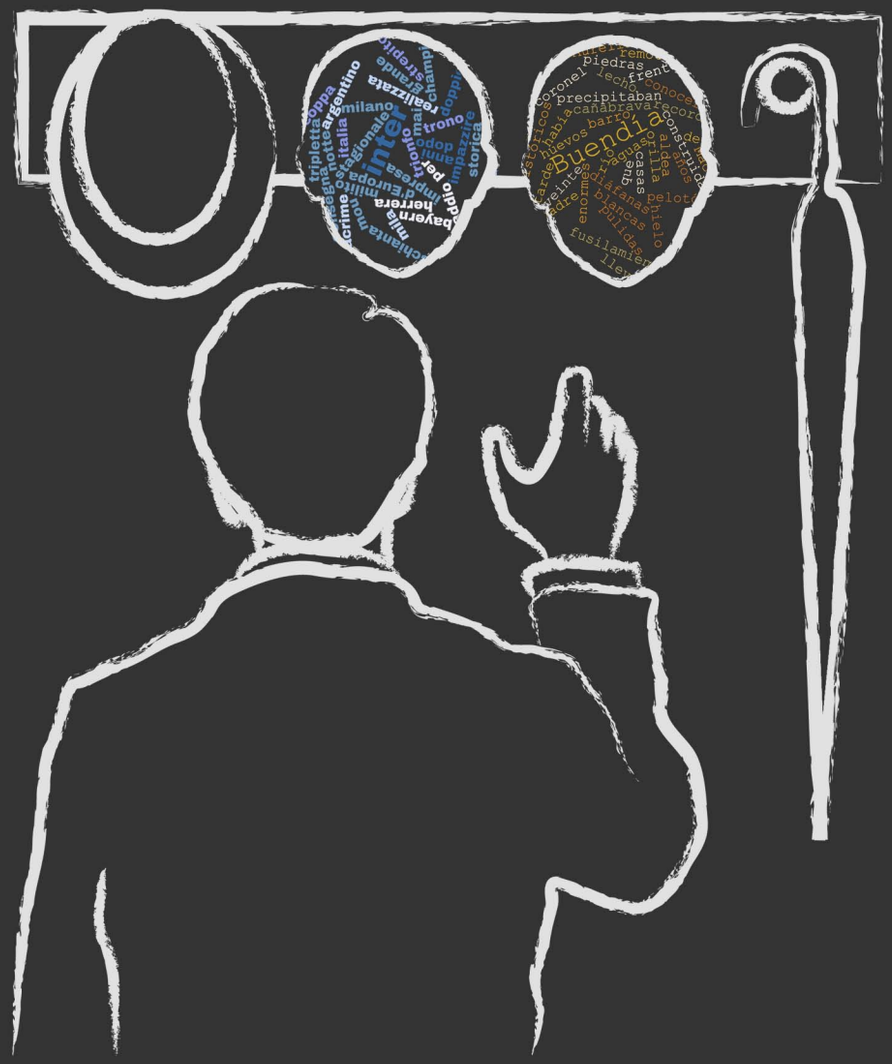
Institute for Logic,
Language and Computation

Universiteit van Amsterdam

marco del tredici

marco del tredici

Linguistic Variation in Online Communities: A Computational Perspective



Linguistic Variation in Online Communities: A Computational Perspective

Marco Del Tredici

Linguistic Variation in Online Communities: A Computational Perspective

ILLC Dissertation Series DS-2020-11

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under VIDI grant no. 276-89-008, *Asymmetry in Conversation*.

Cover design Dario Del Tredici.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6421-063-7

Linguistic Variation in Online Communities: A Computational Perspective

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 6 november 2020, te 16.00 uur

door

Marco Del Tredici

geboren te Gallarate

Promotiecommisie

Promotor:	Dr. R. Fernández Rovira	Universiteit van Amsterdam
Co-promotor:	Dr. W. Ferreira Aziz	Universiteit van Amsterdam
Overige leden:	Prof. Dr. A. Betti	Universiteit van Amsterdam
	Prof. Dr. K. E. Erk	University of Texas at Austin
	Dr. D. Hovy	Università Bocconi
	Dr. E. Shutova	Universiteit van Amsterdam
	Prof. Dr. K. Sima'an	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

A Luciana,
per gioire insieme di quel che io ho potuto avere,
e che a te non è stato concesso.

Contents

Part Two: Modeling User Information for Text Classification

Acknowledgments

This thesis marks the end of a long journey, that I started in Milan in 2004, when I enrolled in my Bachelor's degree. From the very beginning, there was one thing that fascinated me: words. I was lucky (and tenacious) enough to get the chance to spend many years investigating words, in one way or another. During these years, several people have accompanied me along the way. I am sure that, without some of them, I would not have made it to this point. Now, as I think about these people, words that they have told me pop up in my mind. In most cases, these are not the most important or relevant things they told me, but just some observations, comments, and expressions that, for some reason, remained impressed in my mind. I am pretty sure that the people who said these words no longer even remember about doing so. Now, as I thank these special people, I want to go back to these words that helped me find the way.

Malvina: *"Do you mind if I add you on Facebook after the defense?"*
This is what Malvina told me a few days before the defense of my Master thesis, in 2013. For me, Malvina was an alien at the University of Bologna: young, direct, colorful, passionate about NLP. She showed me the beauty of scientific research, and made me think: I want to do a PhD. After the discussion, we indeed became friends on Facebook, and in life. Grazie, Malvina.

Núria: *"There is this guy, Mikolov, have a look at his work"*
Barcelona, 2014, where everything started. I was afraid of every single equation in the papers, and unable to write a line of code. But Nuria always encouraged me, with her motto: understand the math, and code every small thing you do. After many years, I still keep your advice in mind: gràcies, Núria.

Marco *"Well, there are a lot of interesting topics you could work on, you shouldn't be too worried about it"*
I had just started my PhD and, as for most of the PhD candidates, my main worry was: which topic should I work on? One day, while running, I asked Marco for advice, and he calmly gave me this reply. Marco has the capacity to make complicated things

appear easy, and also on that occasion I thought: well, maybe he is right. Shortly after, I found my topic. Grazie, Marco.

Elena: *"It is a great opportunity: go!"*
The offer for a position in Amsterdam had just arrived, and it was indeed a great opportunity, but with a relevant shortcoming: at that time, I was living in Barcelona, with Elena, my partner. Moving to Amsterdam was a huge step, something that would deeply change our lives. I had a lot of doubts and I was terribly afraid, but Elena always encouraged me to go ahead. Grazie Ele: senza di te non ce l'avrei mai fatta.

Raquel: *"Who am I? I am Marco Del Tredici!"*
I do not remember precisely when Raquel said this, but I remember we where in her office, discussing about a metric I wanted to introduce. I was unsure about it, and I said: "Well, who am I to introduce a metric?". Raquel's answer took me by surprise: I suddenly realized that I indeed could introduce a metric, because I had the experience and the knowledge to do that. In a word, I realized in that moment how far I had gone in my academic career. And, if this happened, the person I have to thank the most is Raquel: she offered me a great position, she gave me the time to grow and learn, she guided me in difficult moments, she gave me the opportunity to visit different research groups around Europe. But, most of all, she taught me the value of doing things properly and being self-confident. I know we always use English for work-related stuff, pero esto te lo digo en español: muchísimas gracias por todo, Raquel.

Diego: *"mmm, do you need help with those suitcases?"*
Muiderpoort Station, September 2016. I had just arrived in Amsterdam with two big suitcases and a bike, and Diego was there just to handle me the keys of my apartment. After seeing how much stuff I had, he offered me his help, clearly without being too enthusiastic about it. We met again, few years later, in the Barcelona Amazon lab, and this time he seemed much happier to see me - maybe because I did not have any luggage with me. Thanks Diego, and thanks also Lluís, Hugo, Roi and the other great people I worked with in Amazon: I am proud to be a colleague of yours now!

Elia: *"Trust me, together we can do a lot of stuff"*
November 2016, Elia and I had just arrived in Amsterdam, and we were making great plans for the future. We probably did less than we wanted to, but I did trust him, and it is mainly thanks to Elia if I survived the first year of my PhD: he guided me in my first experiments, taught me how to implement a neural network, and gave me strength with his unlimited enthusiasm. Grazie!

Gemma: *"You should define more clearly your project plan"*
In 2017, I spent six months as a visiting PhD student in Gemma's group. She repeated me this very sentence for about five months. Honestly, at that time I was desperate. But now I see how valuable it was what I learned in that five months. Thanks Gemma, for what you taught me as a researcher, and for helping me in difficult moments. And

thanks also to the great people I worked with in Barcelona: Laura, Ionut, Kristina, Matthijs and Carina.

Sabine: *"Hey, why don't you come to Stuttgart for a while?"*
It was September 2017, and we were at IWCS, in Montpelier. I accepted Sabine's invite, without even knowing where Stuttgart was. This unexpected experience turned out to be one of the best parts of my PhD: during the time I spent with Sabine, Diego, Gabriella, Hanna, Dominik and Jeremy I learned a lot of crucial things for my career, besides having a lot of fun! Thanks guys.

Sandro: *"Pazzesco!"*
The expression Sandro uses to refer to something very nice. An when you are with Sandro, a lot of very nice things happen. When he landed in Amsterdam, in January 2019, everything changed: the gloomy and sad winter days suddenly became full of things "pazzesche" to see, discuss, experience. I am so glad we met: grazie Sandro, sei davvero un grande amico.

Roberto: *"If you want to go, I will support you"*
I was 25, and I was planning to travel the world and become a photographer. I was discussing the thing with Roberto, my father, and this is what he told me. He did not care about how absurd the plan was, because he trusted me: he knew that I would have put all my effort to pursue my goal, and this, for him, was enough. Then, I decided to do other stuff, but his trust remained the same. Grazie papà, mamma, Andrea, Paola, Sara e Fabri: perché anche se le mie scelte mi hanno portato spesso lontano da voi, nemmeno per un momento ho sentito venir meno il vostro supporto. Vi voglio bene.

Besides the people mentioned above, there are several others who, in different ways, played an important role during the years of my PhD, and helped me reach this achievement. First of all, my co-supervisor Wilker: with your calm and deep knowledge of anything related to NLP, you saved my experiments on more than one occasion. Then, all the colleagues at ILLC, especially those in the Dialogue Modelling Group: thanks Mario, Ece, Janie and Lieke, you are all great people, besides being great researchers. Finally, my friends scattered around Europe:
Biondo, Cisto e Mich, non importa in che città io vada, io so che voi verrete sempre, e che io sempre tornerò.
Fa, Tok e Ste: la saga di Del Trenta è iniziata con voi, e con voi finisce. È stato un onore.
Lorenzo, Alessandro e Claudio: che dite, ora che ho un dottorato, glielo mando un curriculum a M.V.?
Jack: forse le pulizie di casa non sono il tuo forte, ma come avrei potuto sopravvivere ai lugubri inverni olandesi senza di te?
Serena: la tua grande pazienza è superata solo dalla tua incredibile memoria. Grazie per essermi stata accanto.

Gian, Toni, los sucios y todos los demás: esperadme, porque ya sabéis que algún día volveré a casa.

To all of you, and to the ones I forgot: thanks for your words.

Chapter 1

Introduction

1.1 Introduction

This work is based on a very simple observation: The meaning of a word is a **dynamic** object that greatly varies depending on the individuals who use that word and when the word is used. As speakers of a language, we are usually aware of the commonly accepted meaning (or meanings) of words. However, we often decide to use them to refer to something new, that is, with a new meaning. This may seem counterintuitive: Since the main purpose of language is to allow communication among people, how is communication possible, if a person arbitrarily changes the rules of the game using a word with a new meaning? In short, the answer is: because communication is not just an exchange of unidirectional messages, but, rather, a joint activity, in which participants continuously **create** and **share** new meanings.

All this might sound as the abstract thought of a PhD candidate with a background in Linguistics, but it is not. The dynamism of language is something we experience, and take advantage of, in everyday communication. Anyone of us can think about common words, or linguistic expressions, that they use with a different meaning, for example, in the WhatsApp group shared with the closest friends, with their family, or with colleagues. If this still sounds too vague, consider this example: In the WhatsApp group I have with two close friends, we say 'take the ferry', 'I am on a ferry' or just 'ferry' to refer to a state of drunkenness - an expression we coined on the ferry on our way home from the pubs in the center of Amsterdam. While not particularly edifying, this example clearly demonstrates the idea of dynamism of words in groups of people. It is highly unlikely that other speakers use the same expression with the same meaning, and it is even less probable that the expression is (or will ever be) recorded in any dictionary: We created it, in a specific context, for specific communicative reasons, and with no need for any explicit agreement.

Similarly to what is described in the example above, countless new meanings emerge every day in groups of speakers of any kind. The result of this linguistic dynamism is that, at a given point in time, the same word can be used to mean many

different things, all different from the commonly accepted meaning of that word. Most of these new meanings disappear shortly after being created, having fulfilled their communicative purposes. Others might survive for a longer time, but be available only in the small group of individuals who created them – as in the 'ferry' example. In some cases, however, the new meaning can spread, and be adopted also by other individuals in the same community of those who created it. This is how **community-specific languages** are created. As an example, consider the AI community: Nowadays, it is perfectly normal to use the word 'attention' to refer to a component of a neural network architecture. This new meaning was introduced a couple of years ago, rapidly spread, and now anyone in the community uses it. However, it is arguably hard, if not impossible, to understand the meaning of 'attention' as a component of a neural architecture for individuals outside the AI community. Finally, in very few cases, the new meaning goes beyond the boundaries of a specific community, and is accepted by all the speakers of a language, finally making its way into general dictionaries. This is the case of 'awesome', that started to be used as a synonym of 'great' in North American English in the 1980s, a meaning that is now commonly accepted by any English speaker, and that is also recorded in **general dictionaries**.[1] A similar, but more recent example, is 'insane', that has recently started to be used as a synonym of 'amazing'. The new usage is rapidly spreading, also thanks to social media, but it has not yet been recorded in (all) dictionaries.[2] To sum up, then, a **hierarchical structure** of nested groups (or communities) of speakers exists, with the smallest groups at the bottom (e.g., me and my two friends), the global community of speakers of a language at the top, and many intermediate communities in the middle. New meanings are introduced at the bottom of this structure, as a response to concrete communicative needs and based on human imagination. Most of these new meanings disappear, others go up through the hierarchy, and, possibly, are accepted in the global community.

While the dynamism of meaning is easily and unconsciously managed by humans as they use language, it has always been, and it still is, a major challenge in Natural Language Processing (NLP) and Computation Linguistics (CL), i.e., the fields this thesis belongs to. Both NLP and CL are interdisciplinary areas that lie at the crossroad between Linguistics, Computer Science and Statistics: While NLP is more concerned with the task of making computers able to read, understand and generate human language, in CL computational techniques and tools are applied to the study of language. In these fields, the main approach to deal with word meaning variation has been, for a long time, the one known as "**sense enumeration**", usually adopted to solve the Word Sense Disambiguation task: Given a word, an automatic technique is applied to select, from a source of knowledge, the most appropriate **sense** of the word in a linguistic context (Navigli, 2009). Knowledge sources usually take the form of machine-readable lexicons, in which each word is associated with a **finite** list of **dis-**

[1] See, for example, `https://www.oxfordlearnersdictionaries.com/definition/english/awesome?q=awesome`.

[2] See `https://www.oxfordlearnersdictionaries.com/definition/english/insane?q=insane`.

crete senses. The sense enumeration approach was shown to be effective when applied to the analysis of meaning in sources of language such as newspapers and books, i.e., in which words are mostly used with highly **conventionalized** meanings, that can be retrieved from knowledge bases (Navigli, 2009; Yarowsky, 2010). However, the sense enumeration approach presents severe limitations when applied to data derived from interactive communication among individuals, for a straightforward reason: It is not possible to include in a knowledge base all the meanings that are continuously created in human interactions.

In order to deal with the limitations of the sense enumeration approach, in the last years, researchers in NLP and CL opened up to the insights from neighboring disciplines such as **Sociolinguistics**, **Psycholinguistics**, and **Historical Linguistics**. As we will see in Chapter 2, in these fields it is well established the idea that the meaning of a word is a nuanced and dynamic object, that varies depending both on the community of speakers that use it and on the moment in time in which the word is used. This opening to new ideas came as the result of two main facts: the acknowledgment of the limitations of the previous approach and the increased availability of **user-generated data** from Computer Mediated Communication, i.e., communication among human beings that takes place via electronic devices. These new data presented two highly valuable characteristics: **abundance**, due to the large and ever-increasing number of people communicating online, and **metadata**, that is, information about the moment in which language was produced, the individual who produced it, and so on. It was, thus, a natural step, for researchers in NLP and CL, to get inspiration from the traditional studies in the aforementioned fields to exploit these metadata, and to do it by using computational tools, in order to manage the unprecedented size of available data.

The union of *old* theories and *new* data and tools led to the birth of **Computational Sociolinguistics**, the research area that has its roots in Sociolinguistics and Computer Science, and that studies the relation between language and society from a computational perspective. In this dissertation I follow the same approach, as I apply computational models to investigate the sociolinguistic phenomenon of **linguistic variation** in online communities of English speakers, that is, the process that underpins the dynamic usage of words described above. For this reason, I consider Computational Sociolinguistics as the most relevant research area for the present thesis. As we will see, I investigate the relation between language and society in different ways, depending on the specific research question. However, what guided me throughout all my PhD was the will of finding efficient ways to account for the dynamism of language, with two main goals in mind: to advance the understanding of the processes underpinning human communication, and to make machines able to better understand human language. In the next section, I will explain how these two goals have been pursued by addressing specific research questions.

1.2 Research Questions

This thesis is organized into two main parts. While both parts are concerned with the investigation of the relation between language and society by means of computational tools, they present different – though, complementary – approaches.

The first part is dedicated to the **analysis of linguistic variation** in online communities. In this part, my goal is to develop methodologies that enable me to effectively identify and describe the linguistic and extralinguistic processes underlying the target phenomenon. This part is characterized by a theoretically oriented approach, as I build on the main findings of traditional theories in Sociolinguistics, and use the tools developed in Computational Linguistics to assess these findings in online communities. I address two research questions, that are highly related to each other. The first one is:

> **RQ-1**: *How to automatically represent and measure meaning variation in online communities of speakers?*

This question aims at assessing the existence of the linguistic phenomenon under scrutiny, and the possibility to identify and measure it using computational tools. Importantly, RQ-1 considers the **synchronic** dimension of variation, that is, the fact that a word, at a specific point in time, is used with different meanings in different online communities. Provided that such a synchronic variation is observed, it is natural to wonder how it comes about. I then address the following question:

> **RQ-2**: *Which are the linguistic and societal processes that lead to variation?*

This question focuses on linguistic **change**, i.e., the **diachronic** dimension of variation, as it deals with the processes whereby new community-specific linguistic practices emerge over time.[3] RQ-2 is a broad question, that touches upon different aspects of meaning change. For this reason, I tackle it from different angles. In the first place, I consider the **linguistic** aspects of meaning change in online communities, analyzing the main linguistic phenomena that underlie it, and the possibility to capture them using computational tools. Subsequently, I focus on the **social** aspects of linguistic change: In this case, the goal is to uncover the role played by different kinds of users in the introduction and diffusion of linguistic innovations in online communities.

In the second part of the dissertation, I leverage the relation between language and society for practical purposes. I build on this intuition: Since the linguistic and extralinguistic practices adopted by the users depend on their social standing, it is possible to leverage the social information about these users to better understand the texts they produce or share. In this part, hence, I take a more task-oriented approach, and focus on how to **improve the performance of NLP models** for text classification

[3] In the rest of the thesis, I will consistently use 'variation' to refer to the synchronic differences existing among communities of speakers, and 'change', or 'shift', to refer to the diachronic modification of word meaning.

by making them able to encode social information about users. I define two research questions, each one focusing on a different kind of user information. The first question is:

> **RQ-3**: *How to identify the relevant information coming from user connections in the social graph and leverage it to improve text classification?*

This question focuses on the information coming from the **social graph** in which users are embedded, and, in particular, on how to extract, from the graph, only the information that is relevant given a specific communicative context.

The second question exploits the relation existing between the way individuals use language and their social and cognitive factors, such as ideas, beliefs and opinions. While the first question is concerned with text classification in general, the second one focuses on a specific problem, namely, fake news, which have recently become a phenomenon of paramount relevance in our society. Since users spreading fake news have been shown to share several of the factors mentioned above, I formulate the following research question:

> **RQ-4**: *How to leverage the linguistic production of users to capture their tendency to spread fake news and, accordingly, to perform fake news detection?*

In this case, hence, the social standing of the users is defined based on their social and cognitive factors, and my aim is to implicitly capture these factors by focusing on the **linguistic production** of the users.

1.3 Contributions

As should be clear from the previous section, this dissertation includes two main *souls*, reflected in the two parts the dissertation consists of. These two souls have equal importance in defining the general contribution of the thesis, that I summarize as follows: On the one hand, my thesis presents a detailed investigation of some relevant sociolinguistic phenomena related to meaning variation and change in online communities, that I identify and describe by means of the computational tools developed in NLP and CL. On the other hand, it introduces methodologies to improve the performance of such tools by making them able to encode information about the social standing of the users in online setups. These two general contributions are deeply intertwined: While the former shows that computational tools can help the theoretical understanding of sociolinguistic phenomena, the latter, by taking the reverse perspective, demonstrates that it is possible to build on sociolinguistic findings to improve the performance of computational tools.

Part One of the thesis provides the theoretically related aspects of the twofold contribution described above. More specifically, in this part, I make the following contributions. First, I show that, in line with the theoretical framework underpinning my

investigation, **online communities show language variation**, whereby the same common word, at a specific moment in time, is used with different meanings in different communities. I obtain this finding by introducing a framework based on computational and statistical tools that allows to identify and quantify variation in a set of online communities, while controlling for the discussed topic. Secondly, I uncover the **diachronic process** that leads to variation in online communities. To this end, I create a longitudinal dataset annotated for **community-specific meaning shift**. I then analyze the dataset, providing a **qualitative analysis** of the linguistic phenomena underpinning meaning change in the community. Also, I use the dataset to **test the performance** of a standard NLP model for meaning change detection. I provide a detailed analysis of the performance of the model, highlighting its main problems, and proposing solutions to them. As a final contribution of Part One, I describe the **social dynamics** related to the introduction and spread of linguistic innovations. I build on **traditional sociolinguistic theories**, and analyze the spread of several thousands of innovations in a large set of online communities. The results of my analysis show that the adopted theoretical frameworks can be used to properly characterize the role of different kinds of users in the process of introduction and spread of linguistic innovations in online setups. Finally, I show that it is possible to leverage information about the users who adopt an innovation to predict if the innovation will successfully spread in a community.

In Part Two, I show that it is possible to improve the performance of several NLP models for text classification by making them able to encode, together with the representation of the target text, the **representation of the users** who produce or share the text. To this end, I introduce two kinds of representation. The first one is based on the connections of a user in the **social graph**. I build on the idea that individuals usually belong to different communities, and that the membership to each of these communities has different relevance depending on the communicative situation. Therefore, I introduce a model that, given a user and their connections in the social graph, **dynamically explores** the connections of the user, finds those that are more relevant for the task at hand, and computes the representations of the user accordingly. In a set of experiments involving the classification of user-generated texts, I show the superiority of dynamic representations compared to representations created by uniformly aggregating information from the connections in the graph. The second kind of representation is based on the **linguistic production** of the users, which I use to perform fake news detection. In my experiments, I show that language-based user representations are a good proxy for the social and cognitive factors shared by people who are more prone to spread fake news, and are beneficial for the target task. Furthermore, I present a deep investigation of the linguistic features that characterize the language used by fake news spreaders, showing their consistency and robustness across domains. Finally, I leverage the relation between the linguistic production of a user and their connections in the social graph to investigate the **Echo Chamber effect**, that is, the condition whereby the ideas expressed by a user are reinforced by their social connections. I introduce a methodology to measure this effect, and analyze its characteristics in light of the linguistic theories underpinning this dissertation.

1.4 Overview

The rest of the dissertation is organized as follows:

Background

- **Chapter 2:** I introduce the main concepts and ideas underpinning this dissertation. The chapter includes three sections: In the first section I define the theoretical framework of the thesis, presenting the theories in the fields of Linguistic and Sociolinguistics that underpin the experiments presented in the following chapters. In the second section, I present the main studies on word meaning done in CL and NLP. I initially define the data used in these studies, then I describe the relevant *descriptive* and *predictive* studies on meaning change, i.e., the studies related to Part One and Part Two of this thesis, respectively. Finally, in the third section, I focus on the computational part of my work, and introduce the neural architectures used in my experiments.

Part One - Analysis of Linguistic Variation in Online Communities

- **Chapter 3:** In this chapter I address RQ-1. To this end, I present a framework for detecting and quantifying semantic variation of common words in online communities of speakers. I apply the framework to several communities, showing that variation is indeed at play in these communities.

  The content of this chapter is based on the following publication:
  Marco Del Tredici and Raquel Fernández. 2017. Semantic Variation in Online Communities of Practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

- **Chapter 4:** I turn my attention to RQ-2, focusing on the linguistic aspects of meaning change in online communities. I present a dataset annotated for this kind of meaning change, provide an analysis of the linguistic phenomena related to it, and then use the dataset to assess the performance of a standard model for meaning shift detection.

  The content of this chapter is based on the following publication:
  Marco Del Tredici, Raquel Fernández and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

- **Chapter 5:** I still address RQ-2, but now I focus on the social dynamics related to emergence of linguistic innovations. In particular, I build on traditional sociolinguistic theories, and investigate the role of different kinds of users in the introduction and diffusion of lexical innovations in online communities.

The content of this chapter is based on the following publication:
Marco Del Tredici and Raquel Fernández. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

Part Two - Modeling User Information for Text Classification

- **Chapter 6:** I focus on RQ-3, and propose a model that dynamically explores the connections of a user, and creates user representations accordingly. I apply the model to several text-classification tasks, and compare its performance against concurring models.

  The content of this chapter is based on the following publication:
  Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- **Chapter 7:** I finally address RQ-4. To this end, I introduce a model for fake news detection which leverages user representations based only on the language they produce. Also, I leverage the relation between language use and connections in the social graph to investigate the Echo Chamber effect.

  At the time of writing, the content of this chapter has not been published.

Conclusions

- **Chapter 8:** I present a summary of the dissertation and the conclusions.

Chapter 2

Background

In this chapter, I provide an overview of the main concepts, theories, and ideas needed to understand the rest of the dissertation. The chapter is split into three sections. The first section introduces the theoretical frameworks underlying this dissertation. In the second section, I present the work in NLP and CL that investigates the interplay between language use and social aspects in online communication. In the third section, I focus on the models developed in the field of Machine Learning that I used to carry out the experiments presented in the next chapters.

2.1 Key Theoretical Approaches

Said in a nutshell, this thesis is concerned with how new word meanings are created, spread and used in communities of speakers. The goal of this section is to define each of the elements that are included in this short definition: what I mean by meaning (2.1.1), how and why the creation of new meanings takes place in human interaction (2.1.2), how I define a community (2.1.3) and how I investigate the spread of new meanings in communities (2.1.4). To define these elements, I build on several theoretical frameworks, mainly developed in the fields of Linguistics and Sociolinguistics.

2.1.1 Word Meaning

There are countless approaches to define what meaning is (Geeraerts, 2010). In Computational Linguistics and Natural Language Processing, the meaning of words is defined based on the observation of their statistical patterns of usage in a large set of texts produced by humans, usually named **corpus** (Turney and Pantel, 2010). This is the approach that I adopt in this dissertation to define word meaning.[1] Such an approach relies on the **Distributional Hypothesis** (DH), a theoretical framework that has its

[1]To stress the difference with the sense enumeration approach introduced in Chapter 1, in this Section, and in the rest of the thesis, I will consistently talk about 'meaning' and 'meanings', rather 'sense' and 'senses', to indicate the single or multiple possible readings of a lexical item.

roots in different strands of research, namely, in the studies by Zellig Harris in Structural Linguistics (Harris, 1954), and by John Rupert Firth in Corpus Linguistics (Firth, 1957).

The main insight of the DH is that the degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the **linguistic contexts** in which A and B appear (Lenci, 2008). This intuition is usually expressed by the famous motto 'you shall know a word by the company it keeps' (Firth, 1957). The crucial consequence of such an intuition is that, if the meaning depends on the context, then the context can be used as the characterizing feature for measuring the similarity of meaning. For example, consider the words 'ocean', 'sea' and 'bicycle': If we look at a large set of contexts of use of each of them and compare these contexts, we can observe that, for example, both 'ocean' and 'sea' tend to occur with similar words (e.g., 'water', 'waves' and 'ship'), while 'bicycle' does not. We can then conclude that the contexts of use of 'ocean' and 'sea' are more similar than those of 'ocean' and 'bicycle' and 'sea' and 'bicycle'. In turn, based on the DH, we say that the meanings of 'ocean' and 'sea' are more similar than the ones of 'ocean' and 'bicycle' and 'sea' and 'bicycle'.

The previous example provides an intuitive explanation of context similarity. It is possible to define this concept in a more precise way: Given a corpus and its vocabulary V, i.e., the alphabetically sorted set of n unique words occurring in it, a matrix M of size $n \times n$ is created, and words in the vocabulary are the headers of both columns and rows. By looking at the occurrences in the corpus of the i-th word in V, it is possible to fill the cells of row M_i with the number of times the target word occurs close to any word in the vocabulary, where *close* means in a window of words of arbitrary length. Row M_i can then be considered as the **vector representation** of the i-th word in V. By measuring the similarity of two vector representations, it is possible to precisely quantify the similarity of two words. The set of rows in matrix M defines a **semantic space**, that is, a space in which a vector representation is available for each word in V, and it is possible to retrieve a measure of similarity between any pair of words.

Semantic spaces present two relevant features: First, they are **holistic** (De Saussure, 2011), which means that the meaning of a word is defined only in relation to the other words in the lexicon, and cannot be defined in isolation. Second, they are **corpus-dependent**, i.e., the meaning of a word in a semantic space completely depends on the corpus in which its occurrences are observed. This is a very relevant aspect. Consider the word 'virus': If its position in the semantic space is defined by observing its occurrences in a corpus including many medical texts, 'virus' will end up in the semantic space close to words such as 'disease', 'infection', and 'sickness'. Conversely, if a corpus including many texts related to informatics and programming is used, 'virus' will be close in the space to 'malware', 'spyware', and 'software'. Hence, by building two semantic spaces on two corpora derived from different sources, the resulting meaning for the same word can be highly different.

Another important aspect of semantic spaces is that words are comparable not only *within* a semantic space, as in the 'ocean' example above, but also *across* different

semantic spaces. For example, given two communities of speakers C_A and C_B, we can collect the language that is produced in each community, and create two semantic spaces, S_A and S_B, representing the meaning of words in the two communities. If word w is used in both the communities, a vector representation w_A exists in S_A, and, similarly, w_B exists in S_B. By computing the cosine similarity between w_A and w_B we can define how similar is the meaning of w in the two communities. The comparison between w_A and w_B is performed as the one between words in the same semantic space, i.e., based on the cosine similarity of the vectors. The only difference is that, in this case, the two representations belong to the *same* word, but in *different* semantic spaces. The same procedure applies to the analysis of semantic change in time. In this case, a longitudinal corpus, i.e, one covering a long time period, is sliced in several sub-corpora, each including texts produced in a specific time interval. A semantic space is then built for each sub-corpus and the meaning change of a word is measured as the semantic similarity of the word to *itself* in different sub-corpora.

The main insight of this section is that it is possible to create vector representations of words and use them to compute meaning similarity. Also, it is possible to create multiple representations of the same word from any number of corpora, and to compare these representations to see how much the meaning of the word changes in each space. In this thesis, I consider corpora of texts produced by different online communities of speakers, I create a semantic space for each of them, and I compare vector representations in different corpora to investigate lexical variation across communities. Similarly, by using corpora of texts produced at different points in time, I will investigate diachronic lexical change.

Finally, the methodology illustrated above to compute semantic spaces is as intuitive as powerful, and it has underpinned the computational studies on lexical meaning for a long time. Nowadays, however, different methodologies, based on neural models, are used. I will provide the details of these methodologies and models in Section 2.3.

2.1.2 Meaning Creation

> *Consider a caller, as reported in the San Francisco Chronicle (November 24, 1980), who asked an operator at the telephone company's directory assistance about toll charges and was told, "I don't know. You'll have to ask a zero". The caller presumably had several conventional meanings for 'zero' in her mental lexicon, including 'naught', 'freezing temperature', and 'nonentity'. If all she could do was access these meanings and select among them, she would have interpreted zero as 'nonentity'. But she did not. According to the report, she interpreted it as 'person you can reach on a telephone by dialing zero'. Surely, this meaning was not in her lexicon. She created it on the spot.* (Clark and Gerrig, 1983)

The quote above introduces one of the main ideas adopted in this thesis: Understanding a word does not merely mean to retrieve its meaning from a list in the mental

Selection Process	Creation Process
Meanings are conventional	Meanings are not conventional
Meanings are enumeratable by lexicon	Meanings are not enumeratable by lexicon
Meaning coherence is guaranteed	Meaning coherence is created
Word has small number of meanings	Word has infinitely many potential meanings
Intended meaning is selected	Intended meaning is created
Word prompts access to intended meaning	Word prompts recall of relevant information
All needed information is in lexicon	Part of needed information is world knowledge

Table 2.1: Comparison of meaning selection and creation processes.

lexicon, rather, the meaning is defined during the interaction between the individuals involved in the communication. Clark proposed this idea of **meaning creation** at the beginning of the 1980s as an alternative to the dominating **meaning selection** approach, that underpinned several theories about meaning comprehension (Blank and Foss, 1978; Marslen-Wilson and Welsh, 1978; Forster, 1981). These theories presented some differences, especially concerning the access strategies to the existing meanings of a word in the mental lexicon. However, they all had in common two main assumptions, namely, **enumerability** and **selectivity**: When the listener hears a word in an utterance, they access their mental lexicon, that includes all the known words, and, for each word, a finite list of possible meanings (enumerability). The listener selects one of the possible meanings, and uses it to interpret the utterance (selectivity). This process can indeed take place, as the listener processes a meaning commonly accepted for a word: For example, it is plausible to think that the meaning of 'radish' as 'pungent root of the plant of the genus *Raphanus*' in the sentence 'I have two radishes' is in the lexicon of the average speaker (Clark and Gerrig, 1983). However, it is very unlikely that the meaning of 'zero' as 'person you can reach on a telephone by dialing zero', as in the initial example, is in the mental lexicon of the listener before they enter the conversation with the operator. However, the listener is able to understand what the operator is saying, and to successfully conclude the communication. Therefore, what the listener did was to start from their knowledge of the meaning of zero, and then build on it, leveraging their world knowledge. Rather than *selecting* a meaning, then, the listener *created* it (Clark and Gerrig, 1983). In Clark's view, hence, meaning creation and selection are complementary, but characterized by very different assumptions. Table 2.1 reports the main features of the two processes.

Clark calls linguistic phenomena like the zero example above **contextual expressions**, by which he indicates expressions whose meaning does not exist in advance, and it is created on the spot (Clark and Clark, 1979; Clark, 1983). The main characteristic of contextual expressions is that they can take on potentially infinite meanings, depending on the circumstances in which they are used. For example, the same word in the initial example, 'zero', could have been used by a teacher in 'All the zeros must redo their papers' to mean 'persons with a grade of zero on a paper' (Clark and Ger-

rig, 1983). Clark investigates the phenomenon of contextual expressions in a series of works, showing that they are pervasive in human communication, and that they apply to all the linguistic categories, such as verbs (Clark and Clark, 1979), nouns (Clark, 1978), adjectives (Clark, 1983), and proper nouns. The latter category offers a very clear example of meaning creation since, differently from common nouns, for proper nouns there is no list of possible meanings to select from. Despite this, the occurrence of proper nouns in contextual expressions is very frequent:

> *Suppose a friend, taking your photograph, asks you with a glint in her eye, "Please do a Napoleon for the camera". Most people to whom we have offered this scenario report imagining, quickly and without reflection, posing with one hand tucked inside their jacket à la Napoleon. Arriving at this sense is a remarkable feat. The proper name Napoleon, though listed in the mental lexicon, does not have senses of the kind common nouns have. [...] All it contains are designations, or pointers, to individuals such as Napoleon Bonaparte and Napoleon III. [...] In understanding "do a Napoleon", you must represent a designation to M. Bonaparte, but you must also search his biography for a characteristic act fitting your friend's request in this context and create a sense around it. Your interpretation is built entirely around elements from your knowledge of Napoleon's life. These elements are not part of the designation of Napoleon, regardless of which theory of proper names one accepts. You are dealing with elements in your biography of Napoleon, not entries in your mental lexicon. The process is one of sense creation without sense selection.* (Clark and Gerrig, 1983)

A crucial assumption underlying Clark's framework is **cooperation**: The listener must assume that the speaker used a specific expression because they want to be successful in the communication, and they think the listener will be able to understand the new proposed meaning. But how does the speaker know that the listener will be able to understand the new proposed meaning? Because the speaker is relying on the **common ground** shared with the listener. The concept of common ground is another key element in Clark's theory, and is used to denote the set of mutual knowledge, beliefs, and assumptions that are shared between the speaker and the listener. The common ground consists of several layers of knowledge, that are organized hierarchically. In the example of Napoleon, these layers include the identity of the proper nouns and his acts, i.e., the speaker must assume the listener knows who Napoleon is, and have basic information about his biography. Also, the speaker must assume these acts are salient for the current content, and that the listener will be able to select, based on this saliency, which among the possible acts is relevant in the specific context in which they are. Importantly, the common ground is not a static object, that is defined once for all. Rather, it is continuously updated during the interaction, and as the communication progresses, new elements are added or removed from it.

Clark's studies show that the process of meaning creation is extremely frequent in communication. New meanings can be forgotten quickly or maintained. The latter usually happens between intimates, e.g., partners, who may develop their own lexicon, in which many words have a specific meaning, usually related to private matters and personal instances, that is not understandable for those who are not involved in the relation (Hopper et al., 1981). In a few cases, the new meaning goes beyond the common ground shared between two individuals, and makes its way in the **community** the individuals belong to. This happens when other individuals of the same community recognize the meaning as useful, and start using it. When this happens, the new meaning is added to the **communal common ground** (Clark, 1996), that is, the common ground shared by all the individuals belonging to a community. The interaction among these individuals leads to the continuous creation of new meanings, that are added to the communal common ground. It is the accumulation of new, community-specific meanings that gives rise to what are commonly known as specialistic or domain-specific languages, that is, the languages that belong to a specific community, in which many common words are assigned meanings that are not understandable for those who are not part of the community. Also, the communal common ground is a dynamic object: New meanings are continuously added, many of them rapidly disappear, while others stay and become conventional in the community.

The concepts of communal common ground and community-specific meaning are highly relevant in Clark's framework, since, given that no speaker exists outside a community of speakers, the meaning of a word can be defined only within each of these communities:

> *Conventional word meanings hold not for a word simpliciter, but for a word in a particular community. You can't talk about conventional word meaning without saying what community it is conventional in. Word knowledge, properly viewed, divides into what I will call communal lexicons, by which I mean sets of word conventions in individual communities. When I meet Ann, she and I must establish as common ground which communities we both belong to simply in order to know what English words we can use with what meaning. Can I use 'fermata'? Not without establishing that we are both music enthusiasts. Can I use 'rbi'? Not without establishing that we are both baseball fans. [...] Every community has a specialized lexicon.* (Clark, 1996)

Importantly, the communities of speakers defined by Clark are part of a **hierarchical structure**, in which communities at the lower levels are nested in the ones higher in the hierarchy. Consequently, an individual belongs to many communities at the same time:

> *Cultural communities [...] generally form nested sets. San Franciscans, for example, are a subset of Californians, who are a subset of Western*

> *Americans, and so on. [...] When I meet a psychologist named Kay, I infer more and more specialized common ground as I discover she is an experimental psychologist, a cognitive psychologist, a psycholinguist, a psycholinguist working on speech production, a student of Charles Osgood's, and a recent visitor to the Max Planck Institute for Psycholinguistics. [...] We all belong to many communities at once.* (Clark, 1996)

Finally, a highly relevant aspect of Clark's framework, which underpins the study presented in Chapter 3, is the focus on **common words**, that is, words that are used by all the communities, but with different meanings:

> *When we think of jargon, slang, and regionalisms, we tend to focus on the words unique to a communal lexicon. [...] But most common word forms belong to many communal lexicon - though with very different conventional meanings.* (Clark, 1996)

The investigation of meaning variation presented in this thesis heavily relies on Clark's theoretical framework, that I sum up as follows: Different communities of speakers use the same word in different ways not because they choose a different meaning for that word from a list of possible meanings. Rather, the meaning specific to a community is created within that community by the individuals who belong to it, for communication purposes. Among the potentially infinite meanings created by the uncountable communities of speakers, a few manage to go beyond the boundaries of that community, and end up being recorded in dictionaries. Starting from this position, I will first use computational tools to investigate how the meaning of common words varies in different communities of speakers, and how this variation comes about (Part One). Subsequently, I will leverage the properties of communal lexicons to encode and classify texts produced by individuals in online communities (Part Two).

2.1.3 Communities of Practice

In the previous section, I showed how common words take on different meanings in different communities of speakers. But how is a community of speakers defined? The question has been deeply investigated in the field of Sociolinguistics, in particular, in variationist studies, that is, the research area in which linguistic variation is studied in relation to the social characteristics of the speakers using the language. Following Eckert (2012), it is possible to identify three main ways of defining communities, that have been introduced in **three consecutive waves**.

The first wave began in the 1960s with the pioneering studies by Labov (Labov, 1963, 1966), that focused on how the usage of different linguistic forms was related to socio-economic patterns. In the first wave, the focus was on **macro categories** that defined permanent attributes of individuals, such as socio-economic class, sex, age and ethnicity. Variation was investigated based uniquely on the membership to these

macro categories, that determined the adoption of a linguistic form. In this view, no active social agency was considered, that is, the individuals were supposed to make no conscious choice when using language.

The second wave dates back to the 1970s, and is characterized by the adoption of ethnographic methods to establish more direct links between the macro categories introduced in the first wave and the observed linguistic features of the speakers and of the communities they belong to. Milroy's studies on phonological variation in Belfast belong to this wave (Milroy and Milroy, 1985, 1987; Milroy, 1987). While still considering the macro categories defined by Labov, the Milroy's refused the idea of passive adoption of non-standard linguistic forms by the speakers and focused on the **social networks** that the individuals create in their life, correlating network types and characteristics to the adoption of specific linguistic forms. The second wave, then, provides a more grounded and local perspective of the macro categories proposed in the first wave. However, these macro categories are still relevant for determining some static, permanent features of the individual.

The third wave brought about fundamental changes, as it proposed to shift from linguistic variation as a *consequence* of the social identity of individuals to variation as the *means* whereby the social identity is shaped. In this view, communities are created by groups of people who voluntarily gather together on the base of a common endeavor, and that define and share a set of **practices**. Also, an individual is now an active agent who shapes their identity by consciously deciding which communities to join, and by adopting the practices that are specific to those communities. Community membership, hence, is not due anymore to some permanent feature of a person, but to a personal choice to get involved with other people and to share with them some practices. Such a process is well defined by the concept of **homophily** (McPherson et al., 2001), that is, the tendency of people to group together with others they share ideas and beliefs with. McPherson et al. (2001) show that homophily is the crucial element in defining the structure of social communities, and that it causes such communities to be homogeneous with regard to many aspects. Thus, while introduced in a different line of research, the principle of homophily well defines the motivation that moves people to gather in communities and share practices. Among these practices, the most important one is language, that is adopted and collaboratively defined by the members of the community together with other symbolic systems, such as dress, body adornment, ways of moving, and so on (Eckert and McConnell-Ginet, 1992).

Since the shared practices in a community are the most relevant aspect in defining it, these kinds of community are called **communities of practice** (Lave et al., 1991; Wenger, 1998).

> *A community of practice is a collection of people who engage on an ongoing basis in some common endeavor. Communities of practice emerge in response to common interest or position, and play an important role in forming their members' participation in, and orientation to, the world around them.* (Eckert, 2006)

Communities of practice have a flexible nature, which means that they are in constant transformation, and the emergence of new communities goes parallel with the disappearance of existing ones. Any group of people can create a community of practice (e.g., people working together in a factory, regulars in a bar, a family) as long as the group presents three features (Meyerhoff, 2002): (i) **mutual engagement** of the members, that motivates the members to get together and engage in their shared practices. Such engagement can be either harmonious, e.g., the supporters of a team gathering to watch a soccer match, or conflictual, e.g., a group of departments chairs who regularly meet to discuss budget allocation to their departments; (ii) **jointly negotiated enterprise**: the shared engagement and the goals of the community are not fixed, but continuously negotiated by its members; (iii) **shared repertoire**, that is, the existence of a set of linguistic patterns recurrently used in time and negotiated by the members of the community, that eventually conventionalize, and become a constitutive part of the identity of the community. There is a clear similarity between the idea of shared repertoire and the concept of communal common ground presented above: In both cases, the focus is on the linguistic practices that are developed and shared between individuals in a community, and that characterize the community-specific language.

Summing up, a community of practice is based on the reciprocal sense of homophily of a group of individuals, who, based on this feeling, decide to engage in a shared activity. By joining different communities, and sharing the practices of each of them, individuals define their identities. I believe that the concept of community of practice, while developed to model offline communities, perfectly fits online communities too, as it captures the dynamic process whereby users in online setups create and join virtual communities and, what is more important for my research, create and share **linguistic practices**. I therefore adopt this framework, and, in the rest of the thesis, the usage of the word 'community' will always imply the concept of community of practice.

2.1.4 Spread of New Meanings in Communities

Communities of speakers show variation, from a synchronic point of view, in patterns of language use. As seen, variation emerges as the result of the interactions among individuals belonging to the community, and communities are created by individuals who engage in common endeavors. A relevant question, then, is: Who are the individuals who propose the linguistic innovations in a community? And which are the patterns of spread that innovations follow, within communities?

In order to answer these questions, representing a community simply as an aggregation of individuals is not sufficient, and a more structured representation is needed. For this reason, researchers have modeled communities as networks (or graphs) (Milroy and Milroy, 1985; Milroy, 1987; Milroy and Milroy, 1987). In a **social network**, the nodes are individuals, and the connections, or ties, among them are the relationships that exist among the individuals. Having this structured representations, then, it is possible to track the diffusion of linguistic innovations in social networks, and to

study the relation between the observed diffusion patterns and the features of nodes and connections.

In the real world, the nodes in the social network, i.e. the human beings, and their connections are highly complex objects. Furthermore, social networks are huge, boundless webs of ties that reach out through a whole society. In order to deal with this complexity, and to create community representations that are manageable, researchers usually create networks only for circumscribed communities. In these networks, the complexity of nodes and ties is (partially) captured by using specific measures. For nodes, the most commonly used measure is **centrality**. Several types of centrality have been proposed. While these types present relevant differences, they all share the core idea of defining how important a node is in a network based on the number and on the distribution of its connections (Newman, 2010). As for connections, the most common measure is **strength**, a concept introduced by Granovetter (1973), and that in its original formulation was defined as 'a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services that characterize the tie'.

Before analyzing how social networks have been used to investigate the spread of new linguistic practices, it is important to stress the complementarity of the concepts of social network and community of practice: While the main focus of the community of practice is on the common enterprise undertaken by a group of people and on the negotiation of their social identity, network analysis mainly attends to the structural and content properties of the nodes and ties existing among these people. It is only by modeling the structural aspects of a community that it is possible to investigate and quantify the processes going on in it, such as, for example, the spread of new linguistic practices (Meyerhoff, 2002).

In the sociolinguistic literature, two main models have been proposed to account for the spread of linguistic innovations in communities of speakers. The first has been proposed by Labov (1972a), and is based on the concept of centrality. The study analyses the usage of Black English Vernacular (BEV) in three groups of adolescents, showing the relevance of social influence on the adoption of linguistic patterns. The centrality of group members was defined based on open questions made to the members, in which they were explicitly asked to judge the popularity of the other individuals in the group. The most central members of the groups were defined as **leaders**. The study showed that leaders were the individuals showing the most salient use of the BEV features, and that these individuals had a strong influence on their close connections in the social network. Conversely, individuals with low values of centrality showed limited usage features related to BEV. Thus, one of the main claims of the study was: The higher the centrality of an individual, the stronger their adoption of BEV linguistic features and their linguistic influence on the closely related members. The second important finding was that leaders were the main sources of innovations in communities, as the connections existing between leaders of different communities allowed the linguistic innovations to flow from one community to the other.

The second model was proposed in Milroy and Milroy (1985). The Milroy's fo-

cused on the ties connecting individuals in a community, analyzing how their strength influenced the spread of linguistic innovations. In order to operationalize the general definition of strength provided by Granovetter (1973), the strength of the ties was defined based on the following variables: (i) common membership to a high-density and territorially based cluster; (ii) kinship relations; (iii) shared working place; (iv) shared voluntary work. Intuitively, **strong ties** usually corresponded to the relation existing between components of the same family, or close friends, and in the social network they were observed in small clusters (or cliques). Members in these cliques provided strong support to each other, and favored the maintenance of linguistic and, more in general, cultural practices. While cliques of strong ties favor the perpetration of linguistic norms, it also happens that an innovation is received by one of the individuals in the clique, who then spreads it to the other individuals in it. (Milroy, 2002).

In contrast, **weak ties** were those that characterized more marginal relationship, like those of acquaintance. Importantly, individuals with many weak ties were found to be crucial in the spreading of linguistic innovations, as they acted as bridges that allowed the innovations to flow from one clique to the other. As mentioned above, it is only when the innovation is received by one of the strong-tie member in a clique, that they can spread it to the other members in it. Authors stress that while the idea that weak ties play a key role in the transmission of innovation might be counterintuitive at the beginning, it should be considered that, from a purely quantitative point of view, weak ties are more numerous than strong ties in a society. Furthermore, weak ties are created in situations such as, for example, business travels or academic conferences, that is, when it is easier to get in touch with new people and new linguistic variants (Milroy, 2002).

The models introduced by Labov and the Milroy's present similarities and differences. In common, they propose that close-knit groups, i.e., the ones surrounding the leaders and those formed by strong ties, are the main agents of linguistic stability, that allow the spread and conventionalization of linguistic patterns within communities of individuals. What differentiates the two models is the way they account for the spread of linguistic innovations. For Labov, the leaders, besides being the individuals who ensure linguistic stability, are also those that introduce innovations. The crucial aspect, here, is the position of the leaders, who have both access to new linguistic variants, thanks to their connections to other leaders, and a direct, strong influence on the members of their cliques. Conversely, for the Milroy's, peripheral individuals are responsible for change, since they connect cliques of strongly connected users and only marginally experience the pressure to conventionalization coming from the leaders.

In Chapter 5 I will address this dispute, and investigate the spread of linguistic innovations in online communities of speakers, with the goal to make clear which are the dynamics underlying the spread of linguistic innovations in online communities, and which of the two models introduced in this section better explains them.

2.2 Computational Approaches to Linguistic Variation

In Computational Linguistics and Natural Language Processing, the analysis and modeling of human language has been concerned almost exclusively with its linguistic content. Several branches developed in the two fields, each of them investigating language at different levels, such as phonology, phonetics, morphology, syntax, and semantics. For a long time, no attention was devoted to the extralinguistic context in which language is produced. It is only in the past decade that the researchers started to take into consideration such a context, and to investigate crucial factors for language understanding such as the social context to which individuals who produce language belong to, and all the situational and psychological aspects that surround language production (Hovy, 2015). As mentioned in Chapter 1, the adoption of this new approach led to the birth of **Computational Sociolinguistics**, a multifaceted research area defined by Nguyen et al. (2016) as 'the field that integrates aspects of Sociolinguistics and computer science in studying the relation between language and society from a computational perspective', and in which converge both the task-oriented approach proper to NLP and the focus on theoretical aspects characteristic of CL. The present thesis is strongly related to the studies in Computational Sociolinguistics, that are the main focus of this Section.

The main reason why the interest in the extralinguistic aspects concerning human communication suddenly rose in the recent past is the availability of new kinds of data, namely, those deriving from **Computer Mediated Communication** (CMC, Herring, 1996). Computer Mediated Communication is an umbrella definition that indicates all the kinds of communication happening between individuals through electronic devices. It therefore includes a wide range of modalities, such as email, chat, instant messaging and posts on social media. Different modalities became available at different moments in time, and their appearance was related to the developments of new communication technologies. While the term 'Computer Mediated Communication' was mainly used in the early days of the studies in the field, nowadays it is more common to talk about **online communication**. In what follows, I will use the two terms interchangeably.

The data deriving from CMC offered an unprecedented opportunity to researchers, who investigated this new kind of data by leveraging, on the one hand, the knowledge acquired in the field of Sociolinguistics, and, on the other, the computational tools offered by NLP and CL. Thanks to CMC data it was possible to overcome two inherent problems of the traditional sociolinguistic studies. First, the limited size of the data used in these studies, that was heavily bounded by the fact that data had to be manually collected. The ever-increasing amount of online communication has made available huge amounts of data, providing the opportunity to investigate sociolinguistic phenomena on a much larger scale than in the past. Second, the *observer's paradox*, that is, the situation in which the social phenomenon being observed is influenced by the presence of the researcher interested in it (Labov, 1972b). Data in CMC is produced by individuals who are not directly observed by researchers, and can therefore behave

naturally.

I identify two main lines of research in Computational Sociolinguistics. The first line is mostly inspired by the traditional work in Sociolinguistics and the theoretical-oriented approach of CL. It focuses on the analysis of the linguistic behavior of individuals involved in CMC, and how this behavior is related to social patterns. This line of research is mainly **descriptive**, and investigates the traditional research questions in Sociolinguistics in online setups, focusing, e.g., on the relation between sociological categories and language use and the spread of linguistic innovations.

The studies in the second line of research adopt the task-oriented approach typical of NLP and address the general question: How can social information be used in order to improve NLP models for language understanding? In this line, thus, the focus is on the **predictive** power of extralinguistic information. While extralinguistic information has initially been leveraged for traditional NLP tasks, such as sentiment analysis and named entity recognition, in the last couple of years it has become increasingly relevant for new tasks related to urgent issues in online communication, such as fake news detection and abusive language detection.

I first analyze the main features of the data derived from Computer Mediated Communication (Section 2.2.1), focusing on the data produced on social media platforms, i.e., the kind of data I use for the experiments in the next chapters. I then present, in Sections 2.2.2 and 2.2.3, the main studies in the two lines of research in Computational Sociolinguistics outlined above. Both these lines are highly relevant for the present dissertation. In particular, descriptive studies are strongly related to the chapters in Part One, while predictive studies to the chapters in Part Two.

2.2.1 Data from Computer Mediated Communication

Studies on Computer Mediated Communication started in the 1990s, and evolved in parallel with the development of the information and communication technologies that enable human communication. A traditional distinction has been made between synchronous and asynchronous CMC, where the former includes communications among individuals that take place in real-time, such as instant messaging, chats and video/audio online conference, while the latter includes communication in which there is a delay in the interaction, such as emails, discussions on forums, and posts on social media platforms. In its initial stages, studies on CMC mainly focused on asynchronous setups, with a strong focus on emails, the first kind of CMC that reached the general public (Romiszowski and Mason, 1996). In the last years, however, large attention has been dedicated to the study of language in online social media. I focus on these setups, and in particular on Twitter and Reddit, the two online social media platforms that are used as data sources in this thesis.

Twitter[2] is a micro-blogging social media platform created in 2006. On Twitter users can post *tweets*, i.e., short messages, and react to tweets posted by the users they

[2] `https://twitter.com`.

follow, i.e., with whom they are connected. Reactions to the tweets can be of different kinds, such as expressing appreciation (*like*), sharing (*retweet*) and commenting on them. Twitter has gained worldwide popularity, and, as 2019, it had approximately 330 million monthly active users.[3]

Reddit[4] is a social media platform founded in 2005 whose main goal is to foster discussions among people about any kind of topic. The platform hosts more than 1 million forums, called *subreddits*, that cover a large variety of topics, such as news, science, cinema, sport, music, etc. Individuals who register on the platform join the subreddits discussing the topics they are interested in. Once in the subreddit, they can submit content (e.g., links, posts, images), that spark discussions among the users in the same subreddit, who can upvote or downvote the content or comment on it. As for 2019, Reddit had 430 million monthly active users, and 1.2 million of subreddits.[5]

Together with new opportunities, data derived from online communication also brought new challenges to researchers. First, this kind of data presents linguistic features that are hard to define, as they span over the linguistic categories used before its advent. More specifically, a substantial amount of discussion has been dedicated to determining if CMC should be considered a form of written or spoken text (Romiszowski and Mason, 1996). The difficulty comes from the fact that, despite being mostly conveyed through written language (e.g., posts and comments), CMC presents most of the informal aspects that characterize the language used in spoken interactions, for example, the large usage of deixis, elision, non-fluencies, etc. These characteristics of CMC not only pose questions related to its definition, but cause concrete problems to the researchers in NLP and CL. For example, many of the tools developed for the analysis of language (e.g., tokenizers, parsers, models for named entity recognition) and optimized on standard written text did not work, or were only partially working, on CMC-derived data. Initially, researchers tackled this problem by normalizing CMC data, that is, by trying to convert the 'incorrect' pieces of texts, i.e., those showing deviations from the standard forms, to the 'correct' form (e.g., 'cooool' $\rightarrow$ 'cool'). This initial approach, however, has been deemed as inappropriate by several researchers, as by normalizing user-generated texts, valuable information regarding variation across users is lost (Eisenstein, 2013). Thus, researchers adopted the opposite approach, whereby deviation from standard forms is not considered as an error anymore but as a possible source of information. This change of perspective was crucial to kick off studies in Computational Sociolinguistics and it is, of course, the one adopted in this dissertation.

Other challenges related to data derived from online communication are related to social information about the users. First of all, this information is not always available. For example, while for Twitter and Reddit it is possible to retrieve social information about users by querying the APIs of the two social platforms, this is not possible for Facebook. Furthermore, platforms that make data available, usually place tight restric-

[3] `https://www.oberlo.com/blog/twitter-statistics`.

[4] `https://www.reddit.com`.

[5] `https://techcrunch.com/2019/12/04/reddits-monthly-active-user-base-grew-30-to-reach-430m-in-2019/`.

tions on the circulation and usage of the data. A further issue is volatility, that is, the fact that data that are available at a specific point in time, might no be available at a subsequent point. This might be due to actions taken by the users, such as deleting their previous posts or canceling their profile from the platform, or by the administrators of the platforms, who have the right to remove posts or ban users due to violations of the platform's rules. This situation has detrimental effects on research as it hinders the possibility to compare the performance of models developed at different moments in time on (exactly) the same set of data. The volatility problem affects also some of the experiments presented in this dissertation, as shown in Chapter 6.

A second relevant aspect is related to the quantity and quality of the data voluntarily made available by users on social platforms. In general, personal information provided on social platforms is optional, and there is no control over it. Consequently, many of the users do not provide any kind of information, and, for those who do it, it is not possible to assess the veracity of the provided information. Typical examples, in this sense, are information about age, nationality, work, and name, that are potentially very valuable for investigating the relationship between language and social categories, but whose reliability cannot be taken for granted.

Given the problems related to the retrieval and reliability of the social information voluntarily provided by users, several researchers mainly, or exclusively, focused on the social information that does not have to be explicitly provided, and that can be extracted by looking at users' behavior on the social platforms. This kind of information includes, in the first place, the language produced by users in social media. Users incessantly produce text, by means posts, comments, status, etc, making text an incredibly abundant resource of information. Since the variation shown in text is related to the social traits of the users who produce them, text is also a potentially very valuable source of information. I leverage this kind of information in the experiments presented in Chapters 6 and 7. Another relevant kind of information is the one related to the social behavior of users on social media, that includes, for example, the connections they create with other users, the degree of participation to the online communities they join, and the number of posts they write. While arguably less informative than the information about social categories voluntarily disclosed, these kinds of information present two highly desirable features: high **coverage**, since it can be retrieved for any active user, and **reliability**, since it is not susceptible to mystification. As I show in the next sections, both the information voluntarily disclosed by users and the one that can be retrieved by observing their online behavior have been exploited by researchers.

2.2.2 Descriptive Studies

The studies in Computational Sociolinguistics that investigate the relation between language use and social patterns followed, at least to some extent, the same three waves of the traditional sociolinguistic studies presented in Section 2.1.3. It is therefore possible to identify studies that focused on macro categories online users belong to, such as gender and age (first wave), on the structure of the online social network in which users are

embedded (second wave) and on the engagement of users in these communities (third wave).

The majority of the studies in this section present a common methodological approach, whereby the social categories under scrutiny are considered as the target classes to be predicted, and a machine learning algorithm is implemented and tuned to predict such classes based on the language produced by users. For example, given the post written by a user, and the target categories *male* and *female*, a model is fed with the post and asked to predict the gender of the user. The level of accuracy of the prediction is usually considered indicative of the strength of the relation between the linguistic features in the input texts and the target social categories. Moreover, many studies perform an analysis of the model, in order to identify the linguistic features that are relevant for the target class, i.e., for the social categories. In what follows, I mainly focus on the findings related to the social features of the users, while I analyze the relevant computational approaches in Section 2.3.

First Wave Studies

Studies in this group analyze the relation between macro categories such as age, gender, and location and the use of language in online setups. As in the 'original' first wave, also in this case the macro categories are considered as fixed characteristics of the users, whose social agency is not taken into account. These studies are not directly related to the ones in this dissertations, in which users are characterized based on their active engagement in online communities and on the connections in social graphs, without considering macro categories. Nevertheless, first-wave studies are important as they share with the research in this thesis the crucial assumption that extralinguistic and, in particular, social information is relevant to understand language. I therefore review them in what follows, focusing on each macro category independently.

Gender A large number of studies focused on gender, especially in the early days of Computational Sociolinguistics. These studies mostly model gender as a binary class. They usually report high accuracy in the prediction task, that indicates that language does show variation between men and women (Goswami et al., 2009; Mukherjee and Liu, 2010; Otterbacher, 2010; Gianfortoni et al., 2011; Fink et al., 2012; Bergsma and Van Durme, 2013; Markov et al., 2016).

Almost all these studies provide a list of linguistic features, related to both content and style, that are found to be distinctive for the two genders. For example, some works found that male users tend to use more number and technology words, while female users use more words related to family and relationships (Boulis and Ostendorf, 2005; Bergsma and Van Durme, 2013; Bamman et al., 2014b). Others report that texts authored by female users are characterized by a higher rate of use of first-person pronouns and verb 'to be', while texts written by male users by the usage of prepositions and third person (Otterbacher, 2010).

While these studies had the merit to open the way to Computational Sociolinguistics and to identify interesting aspects of the relation between gender and language, I highlight three main issues related to this kind of studies. First, it is hard to get a general picture out of the results reported in the different studies. This is due to the fact that the results are very fragmented, with different studies focusing on different linguistic features in a non-systematic way. Also, in some cases results are contrasting. For example, while Otterbacher (2010) identify the usage of pronouns as a marker of male-generated language, Bamman et al. (2014b) associate this feature to texts generated by female users. This difference in results is especially evident for features related to style. Second, few studies have considered the possible bias due to other social variables. Among these studies is Gianfortoni et al. (2011), that show that when controlling for other variables (e.g., occupation), the features that were thought to be predictive of gender were not predictive anymore. Finally, another concern about research on gender comes from the strictly binary approach taken by almost all the studies. Nguyen et al. (2014) question this approach and focus on the concept of agency, stressing that, while it can be the case that some stereotypical patterns can be identified in the language of male and female users, users may also decide not to adopt these patterns, due to their own wish to define their identity.

Age Another set of studies focused on the relation between age and language. While for gender the choice of a binary classification task appears straightforward, for age it is less obvious how to define the set of target classes. The majority of studies stick on the classification approach, and define classes based on age spans. For example, Rangel et al. (2014) propose the spans 18-24, 25-34, 35-49, 50-64, and 65+, while Al Zamal et al. (2012) use spans 18-23 and 25-30. Defining age boundaries is not trivial, as there is no clear motivation why an option should be better than another, and choices are mostly related to the data being used. This causes a great variation across the studies, that makes it difficult to compare the reported results. To avoid this problem, a few studies followed a more intuitive (and difficult) approach, that is, to model age as a continuous variable, and therefore performing prediction as a regression task (Nguyen et al., 2011, 2013; Schwartz et al., 2013; Sap et al., 2014)

The main findings in this line of research are related to the language used by teenagers, that, in general, presents more variation from standard language, compared to the language used by adults. A large number of features have been found to be distinctive of the language of teenagers, such as lengthening (e.g., 'niiiice') (Rao et al., 2010; Nguyen et al., 2013), usage of Internet acronyms (e.g., 'lol') (Rosenthal and McKeown, 2011), slang (Barbieri, 2008; Rosenthal and McKeown, 2011), swear words (Barbieri, 2008; Nguyen et al., 2011), and emoticons (Rosenthal and McKeown, 2011). These findings confirm those in Sociolinguistics, that found that adolescents are more prone to use non-standard forms, presumably due to group pressure to not conform to societal rues, while this tendency is less evident in adults (Nguyen et al., 2016).

Differently from what is observed for gender, the studies about the relation between

age and language use provide a more unified and coherent picture, especially in relation to the differences between teenagers and adults. However, also in this case, there are issues related to the possibility to generalize the reported findings. For example, while it might be that a linguistic feature is found to be relevant for teenagers at a specific point in time (e.g., the large use of emoticons), it is not the case that the same feature will still be distinctive in a future point in time, given that changes take place also between generations of teenagers (Nguyen et al., 2016) and that what might be relevant is not the chronological age of users, but their 'online age', i.e., for how long a person has been active in online setups.

Location In this line of research it is possible to identify two distinct approaches. The first one follows the general methodology defined above, whereby a machine learning algorithm is trained to predict a linguistic variant, or a dialect, among a set of possible ones, given the input text. Studies of this kind have focused mainly on Arabic dialects (Elfardy and Diab, 2013; Sadat et al., 2014; Shoufan and Alameri, 2015), for which shared tasks were also organized (Malmasi et al., 2016), but the research area is very active also for European dialects (Trieschnigg et al., 2012; Zampieri et al., 2018).

In the second line researchers leverage the information in geo-tagged datasets, that is, datasets in which each text is associated to information about the geographical area in which it was produced. This kind of data become very popular in the last years thanks to the geographic information about users made available by Twitter, and to the large use of social media on mobile devices, with the possibility to add to the posted content information about the current position (Eisenstein et al., 2010; Wing and Baldridge, 2011; Gontrum and Scheffler, 2015). In this case, the goal is to predict the location, given the text. The prediction task can be cast in two ways (Han et al., 2016): (i) as a regression problem, in which the target is to predict the correct values of latitude and longitude given the input tweet (Rahimi et al., 2017; Hovy et al., 2020); (ii) as a multi-class classification problem, in which the geographic space is partitioned in cells by means of a grid, and the goal is to predict the right cell. This approach is arguably easier than the previous one, but, somehow similarly to what I observed above for age spans, it presents the problem of how to partition the space in a meaningful way. For this reason, several researchers proposed alternatives to the usage of grids, in order to automatically define geographic areas that are more adherent to real diffusion of linguistic variants (Han et al., 2012, 2016; Eisenstein, 2015).

Studies on geolocation had a large diffusion and presented high accuracy results. However, it is possible to identify a common shortcoming for all of them, namely, that very often the relevant linguistic features for the prediction (e.g., named entities) are not of great interest for the sociolinguistic analysis. For example, a model might learn to use the bi-gram 'Times Square' as a strong indicator for the city of New York, simply because, predictably, the citizens of that city use it more than others (Nguyen et al., 2016).

Lately, several studies investigated the relation between geographic variables and

variables of other kinds. Ying et al. (2018) perform tweet geolocation together with event extraction, based on the idea that the two tasks are complementary and, potentially, beneficial one for the other. In Miyazaki et al. (2018) geolocation is improved by using knowledge basis derived information about the words used in the tweets, the idea being that having more information about a word can help to better understand where it was produced. Finally, Fornaciari and Hovy (2019) show that by considering information about geolocation it is possible to better predict other users' social variables, such as age and gender.

Socio-economic categories The studies that take into consideration socio-economic information are less than the ones considering age, gender, and location. The reason is the aforementioned difficulty to retrieve socio-economic information on social platforms, that is provided by approximately half of the users in social media, and whose veracity can not be assessed (Culotta et al., 2016). In order to overcome these difficulties, a general approach has been adopted that relies on the combination of geolocated data, mostly coming from Twitter, and data derived from censuses, that provide socio-economic information on sectors of population based on their location. In this approach, then, the location of a person is defined based on geolocated data, and their social status is inferred by inspecting the data in the census for that location. In line with previous approaches, a set of classes is then defined, usually based on the degree of education, income, and occupation. Also in this case, the definition of the classes is arbitrary, and mostly dependent on the available data, that makes it difficult to compare results across studies.

In general, studies report that differences between social classes are mostly related to the discussed content, and that words expressing sentiment have a high discriminative power (Preoţiuc-Pietro et al., 2015a,b; Flekova et al., 2016; Lerman et al., 2016; Volkova and Bachrach, 2016; Abitbol et al., 2018). As observed for gender, however, also in this case results do not always agree. For example, Quercia et al. (2012) report that the higher the sentiment score of a community, the higher the community's socio-economic well-being, while Preoţiuc-Pietro et al. (2015b) conclude that higher-income users express more fear and anger than lower-income users. While these differences might be due to the different data used in the two experiments, and the two conclusions considered equally valid, it is hard to draw general conclusions on the relation between language and social status. Despite the difficulties in collecting data annotated with socio-economic information and in creating a common framework that allows a more systematic comparison across studies, this line of research is very promising, as socio-economic information has shown to be more informative compared to other variables such as location (Eisenstein et al., 2014).

Second Wave Studies

I consider now the studies in Computational Sociolinguistics that are related to the second wave in variationist studies, that is, those that investigated linguistic variation

in relation to the social network in which users are embedded, focusing, in particular, on the spread of linguistic innovations. These studies are directly connected to this thesis and, more specifically, to Chapter 5, in which I also focus on the diffusion of linguistic innovations in online communities.

The study of online social networks has attracted a lot of interest since their appearance, leading to the emergence of a new field of studies named Online Social Network Analysis (OSNS) (Kurka et al., 2015). The studies in the field focused on the structural properties of online social networks, in order to define basic features such as the distribution of connections across users and the existence of cliques of users connected by strong ties, as well as to assess the presence of well known phenomena in offline networks, e.g., the small-world property (Mislove et al., 2007; Kwak et al., 2010).

Large attention has been devoted to the analysis of how the content produced by users spread in online social networks. An overview of the studies in this line can be found in Guille et al. (2013) and Kurka et al. (2015). Here, I focus only on the studies that analyze the spread of linguistic innovations in online social networks based on the sociolinguistic work presented in Section 2.1.4. Most of these studies focused on tie strength. For these studies, a crucial methodological question is how to define tie strength in online setups. This revealed to be a difficult task, due to the fact that it is not possible to replicate in online communities the same fine-grained criteria used in offline communities. Several solutions to this problem have been proposed. Onnela et al. (2007) investigate the relations in mobile communication networks, and propose to define the strength of the tie between two users based on the degree of overlapping of their neighbors, that is: The higher the number of common neighbors, the stronger the tie between two users. In line with the studies by the Milroy's, the authors find that information is retained and quickly spread within cliques of strong ties, while it flows from one clique to the other through weak ties. The same measure of tie strength is adopted by Goel et al. (2016), that focus on Twitter data, and track language changes as they take place. The study confirms that the linguistic influence exerted across densely embedded ties is greater than the influence across other ties, thus confirming the influence of strong ties in their cliques found by Labov and the Milroy's.

Another approach to the computation of tie strength is based on the frequency of the interaction between users, and is proposed by Paolillo (1999). The study investigates tie strength in online communities on Internet Relay Chat (Werry, 1996). The reported results only partially confirm the findings in the traditional studies by Labov and the Milroy's, as they find that members of cliques of users connected by strong ties consistently share some linguistic patterns, and that these patterns are often different from those of other cliques. However, they do not find a clear mapping between position and strength of the ties and the diffusion of linguistic innovations. In particular, no clear evidence is found regarding the role of weak ties as the main vectors of linguistic change. A similar approach is proposed by Bak et al. (2012), that measures tie strength based on the frequency of interactions and on their duration. Authors use Twitter data, and correlate tie strength to self-disclosure, that consists of the personal information shared in communication, and it is computed by using a sentiment lexicon. Authors

find that people disclose more to closer friends, i.e., users connected by strong ties, but also that people show more positive sentiment towards weak relationships rather than to strong relationships. They conjecture that this reflects the social norm adopted with first-time acquaintances on Twitter.

Finally, Ferrara et al. (2012) investigate the relation between tie strength and linguistic diffusion on Facebook. Authors define an undirected and unweighted graph based on the *friendship* relation on Facebook, and define as weak ties the edges that connect nodes belonging to different communities in the network, while intra-community edges are considered strong. In order to identify the communities in the network, authors adopt a methodology that, given all the possible partitions of the social graph in several clusters, selects the partition that maximizes the number of intra-cluster connections, and minimize the number of extra-cluster connections. Each cluster in the partition is considered a community in the graph. Also in this case, the results reported by the authors provide some evidence about the role and importance of weak ties, but do not find a clear mapping between tie strength and linguistic spread.

In the experiment analyzing the spread of linguistic innovations presented in Chapter 3, I adopt the same methodology used in Onnela et al. (2007) and Goel et al. (2016) to define tie strength, for two main reasons: First, because it is the one that led to the most clear results, among the ones I analyzed. Second, because it is the most suitable one for the kind of data I used.

Third Wave Studies

I finally present the computational studies related to the third wave in variationist research. I consider these studies as the most relevant for this thesis, as they investigate aspects of linguistic variation in online communities that are highly related to the ones I explore in the next chapters, such as the social agency of individuals, the emergence of (online) communities of practice, and the acquisition of these practices by new members.

The studies in this line of research were conducted on different kinds of data, such as email threads form the Enron corpus, (Bramsen et al., 2011), social media such as Twitter (Danescu-Niculescu-Mizil et al., 2011) and Reddit (Nguyen and Rosé, 2011), discussion forums on Wikipedia (Danescu-Niculescu-Mizil et al., 2012; Noble and Fernández, 2015; Yoder et al., 2017), online reviews (Hemphill and Otterbacher, 2012), and transcripts of Supreme Court arguments (Danescu-Niculescu-Mizil et al., 2012).

While data greatly differ across studies, the majority of them present a common approach. They initially analyze the language produced by the users, by using either shallow features, such as n-grams, or by leveraging lexicons, such as the Linguistic Inquiry and Word Count dictionary (LIWC, Pennebaker et al., 2001), that allow grouping semantically related n-grams. Then, they investigate the relation between the language produced by the users and the social dynamics in which they are involved. For example, it is observed how the language produced by the users changes as they join an online community and adopt the linguistic practices of that community.

One of the first studies in this line, and highly related to our investigation, is Cassell and Tversky (2005), that present a longitudinal analysis of the patterns of linguistic interaction in online communities of young people coming from different cultural, linguistic, and socio-economic backgrounds. The authors show that people involved in the experiment increasingly constituted themselves as a community of practice, negotiating and converging on specific linguistic practices, and setting together goals and strategies to achieve them. This study, thus, suggests that the mechanisms and processes characterizing offline communities of practice can be identified also online. Similar findings are reported in subsequent studies: Elhadad et al. (2014) focus on an online community created to support individuals affected by breast cancer, and show that a community-specific terminology emerges, especially related to the semantic areas that are relevant for the members, such as medications, symptoms, side effects, and emotions. Sharma and De Choudhury (2018) investigate the linguistic practices of an online community created to support individuals with mental illnesses, and find that a positive correlation exists between convergence in the usage of linguistic patterns and emotional and informational support.

While previous studies look at how, in general, linguistic practices are created and shared in communities, others investigate how such linguistic practices are adopted by new members. Nguyen and Rosé (2011) report that users that join an online community initially show a clear shift in language usage towards the norms of the community, with a stabilization after a period of 8/9 months. Also, they find that users who have been in the community for a long time share a strong use of community-specific jargon and style, with which they express familiarity and their emotional involvement in the community. In a related study, Danescu-Niculescu-Mizil et al. (2013) suggest that users in online communities follow a two-stage lifecycle: A first stage in which, similarly to what suggested by Nguyen and Rosé (2011), users adopt the language of the community, followed by a conservative phase in which users stop adopting new linguistic patterns, that continue to emerge in the community.

The extent to which linguistic practices are adopted by community members is leveraged also to define their degree of engagement in a community: Hamilton et al. (2017b) show that highly engaged users employ language that signals collective identity, and that a higher level of engagement of the users in a community correlates positively with the number of interactions and negatively with fragmentation in the structure of the community. Similarly, Zhang et al. (2017) analyze several hundreds of Reddit communities, and measure the degree of users' engagement in a community based on the adoption of linguistic practices specific to the community. They show that different kinds of communities show different patterns. For example, small and dense communities with strong identities are more likely to retain their users, but, at the same time, these communities also exhibit much larger accommodation gaps between existing users and newcomers. Finally, Tran and Ostendorf (2016) show that the linguistic practices developed in Reddit communities are so clear that they are a better indicator of the community identity than the discussed topic, even for communities that are focused on very specific topics. They also show that there is a positive

correlation between the community reception to a contribution and the style similarity to that community, but not so for topic similarity.

Among the third wave-related studies, many focused on the dynamics governing meaning negotiation and, more in general, how coordination among individuals is influenced by their (perceived) power. The concept of power is a generic one, that has been operationalized based on different sources of information: the pre-existing organizational structure, such as the one in the Enron corpus (Bramsen et al., 2011; Gilbert, 2012), the number of connections in social networks (Danescu-Niculescu-Mizil et al., 2011), the status of admin in Wikipedia (Danescu-Niculescu-Mizil et al., 2012) or the degree of centrality of a user in the social network (Noble and Fernández, 2015).

Among the first studies considering power relations is Bramsen et al. (2011), that use the data from the Enron corpus, and develop a classifier to determine if a message is sent to a message of higher or lower status. Gilbert (2012) extend this work by proposing a method that identifies the linguistic indicators of the power status. Prabhakaran et al. (2012) consider a more dynamic situation, in which power relations change depending on the situation and the goal of the interaction. In this case, authors do not find a clear mapping between situational power relation and linguistic features.

Other studies consider power relations on social media, for example, Danescu-Niculescu-Mizil et al. (2011), that show that linguistic convergence among users on Twitter takes place at several different levels, and that is is often asymmetric, i.e., one of the two participants is more prone to accommodate the other. Authors, however, find no correlation between accommodation and social status. Differently, Danescu-Niculescu-Mizil et al. (2012) find that power relations do influence linguistic convergence in discussions between editors on Wikipedia forums. In particular, they observe that people with low power exhibit greater language coordination than people with high power, and people coordinate more with interlocutors who have higher power than with those who have lower power. They also find that when a person changes their status, their coordination behavior changes, and so does the behavior of people talking to them. Noble and Fernández (2015) extend previous work, by finding a positive correlation between linguistic coordination and power, where power is defined based on the centrality of a user in the social network.

The power relations between male and female users in online setups is investigated by Hemphill and Otterbacher (2012), that report that female users adjust their communication styles to the one of male users more than what is observed in the other way round. Yoder et al. (2017) investigate the relation between social influence and code-switching in collaborative editing, finding that code-switching is positively associated with Wikipedia editor success. Finally, Jones et al. (2014) investigate the relation between people's inherent tendency to accommodate and the power of a user, showing that low levels of power correlate to high tendency of accommodation.

2.2.3 Predictive Studies

After reviewing some of the main studies that investigate the relation between the language produced by the users and their social traits, I now focus on the line of the studies that *leverage* these traits in order to improve the performance of the NLP models dealing with the classification of user-generated texts. This line of research is more recent than the one of descriptive studies. Arguably, this is due to the fact that researchers initially focused on the analysis of the linguistic behavior of individuals in online communities, and, subsequently, started to leverage the acquired knowledge for practical purposes.

At a general level, the reason for leveraging user information is very intuitive: It is easier to understand a piece of language if you know who produced it. Since information about the speakers (or, better, *posters*), has become more and more available (see Section 2.2.1), it was a natural step to start encoding this kind of information in NLP models. A more theoretically grounded motivation to exploit user information is provided by the sociolinguistic theories introduced in Section 2.1 and, in particular, by the notions of homophily and community of practice: Since people group in communities that build and share their own linguistic practices, by knowing the community a user belongs to, it is possible to know which practices they adopt and, as a consequence, it is easier to understand what they say.

While studies in the previous section used a variety of data to investigate the language produced by the users, studies in this section make use almost exclusively of data derived from online social platforms and, in particular, Twitter and Reddit, i.e., the same data used in the experiments in this thesis. This is due to the fact that in the last years, that is, the period in which these studies were implemented, data derived from social media became the *de-facto* standard, due to two main reasons: First, the fact that on social media users are free to talk about any kind of topic, thus expressing more overtly their interests and ideas, compared to other setups in which the content produced by the individuals is bounded by specific goals (e.g., Wikipedia forums). Second, the magnitude of the data available in online social platforms, that have reached an unprecedented level of coverage, with respect to both the synchronic dimension, i.e., the number of users and communities for which it is possible to obtain information, and to the diachronic one, i.e., the possibility to retrieve the past content produced by the users during their life on the social platform.

User information has been leveraged to address a large number of tasks. Unsurprisingly, researchers initially addressed typical NLP tasks such as sentiment analysis (Hovy, 2015; Jiménez-Zafra et al., 2017; Yang and Eisenstein, 2017), sarcasm detection (Khattri et al., 2015; Rajadesingan et al., 2015; Wallace et al., 2016; Ghosh and Veale, 2017; Hazarika et al., 2018; Kolchinski and Potts, 2018; Oprea and Magdy, 2019), information extraction (Yang et al., 2016), and stance detection (Lai et al., 2020). Subsequently, researchers started focusing on new tasks that emerged in response to the increasing pervasiveness of specific issues on social media: fake news detection, i.e, the task of classifying a piece of news as either real or fake (Gupta et al.,

2013; Zubiaga et al., 2016; Long et al., 2017; Kirilin and Strube, 2018; Guess et al., 2019; Reis et al., 2019; Shu et al., 2019a,b); abusive speech detection, that is, the task of detecting offensive and hateful speech, especially against minorities (Mishra et al., 2018, 2019a); political perspective detection, in which the goal is to automatically detect the political preferences of users (Li and Goldwasser, 2019); detection of problems related to mental heal (Amir et al., 2017), depression and anorexia, (Schwartz et al., 2014; Wang et al., 2018) and potential suicidal intentions (Mishra et al., 2019b; Sinha et al., 2019). The relevance of these new research areas has become so evident to the research community, that several workshops and shared task have been dedicated to them (Hanselowski et al., 2018; Zirikly et al., 2019), in order to impulse the studies on these topics and improve the performance of the models.

Almost all the studies that leverage user information adopt a similar approach: They consider a classifier that performs the target task using only textual information derived from social media, and make the model able to encode, together with the textual information, also the information about the user who produced that text. User information is deemed useful if the performance of the model in the setup in which textual and user information are leveraged improves over the performance achieved when only textual information is used. This general approach is implemented also in Chapter 6 and 7 of this thesis.

Given this approach, the key point is how to create user representations that effectively capture the homophily principle, i.e, that mirrors the degree of similarity of different users. Two main approaches have been used to define user representations.

The first one relies on the **language production** of users. In this case, all the previous posts of the user are collected, and the textual information in these posts is encoded in a fixed-size vector representing the user (see Section 2.3 for more details). This approach leverages the fact that the online identity of users is created and conveyed mostly through verbal communication, and relies on the theoretical assumption that the language used by an individual mirrors their main psychological and sociological features (Pennebaker et al., 2003; De Fina, 2012). In this case, then, the degree of similarity of two users is based on how similar is the content they produced.

This approach has led to improvements in several tasks. For example, in the task of sarcasm detection in online discussions, for which it was shown that encoding the previous posting activity of users makes it easier to recognize the linguistic patterns that they employ to express sarcasm (Khattri et al., 2015; Wallace et al., 2016; Hazarika et al., 2018). Similarly, Amir et al. (2017) show that representations based on posting history capture the homophilic relations between users who share mental health issues, and that can therefore be used to detect problems of this kind.

The second approach leverages the **social graph** in which users are embedded, and user representations are created in such a way to represent the proximity of users in the graph: The closer the users, the more similar their vector representations. This approach, hence, implements the principle of homophily based uniquely on social connections, relying on the idea that, similar to what happens in offline setups, also in online ones people create connections with those that they perceive as more similar.

The validity of this approach has been proved in many studies. For example, Yang et al. (2016) show that graph-based representations can help disambiguate ambiguous entities in the task of entity linking (i.e., the task of linking an entity extracted from a text to the right entry in a knowledge base): If a user mentions the entity 'Giants', connected in the knowledge base to both the football and baseball teams, knowing that most of the connections of the user usually talk about baseball teams will help to connect the entity to the right entry. The same principle is applicable to other tasks, such as abusive speech detection, where the propensity of a user to use offensive words is related to the one of their connections in the graph (Mishra et al., 2019a), or detection of political affiliation, as the political affiliation of individuals is, most of the times, the same of the connections they share information with (Li and Goldwasser, 2019). In general, then, experimental results show that graph-based approaches well capture the idea of homophily, and that by leveraging information coming from the connections of a user in the social network leads to improvements in several tasks.

The two approaches presented above rely on different but complementary sources of information, as shown by Sinha et al. (2019), that investigate suicidal detection on Twitter, and show that by using user representation created by leveraging both historical posts and social connections it is possible to get higher results than when only one of the two sources of information is used.

Concluding, the studies presented in this section show that user information helps to better understand textual information on social media, and that it can therefore be leveraged to enhance model performance in a large set of NLP tasks. I build on these findings, and introduce a model that creates user representations based on the information from the social graphs (Chapter 6), and one that leverages only the linguistic production of users (Chapter 7).

2.3 Computational Models

In this section, I present the computational models that I use for the experiments in the next chapters. All the models belong to the family of Artificial Neural Networks, that, in the last decade, have become the standard choice for most of the studies in NLP and CL (Goldberg, 2017). For each model, I provide an overview of its general architecture and highlight the reasons why I adopted it for my experiments. The section is split into two parts. The first part includes the models that I use to process textual inputs, while in the second part I present the architectures I use to model social graphs. This organization does not reflect the inner characteristics of the models; rather, it depends on the specific usage that I made of such models to carry out my research.

2.3.1 Models for Processing Language

First, I introduce the neural architectures that I use to model linguistic inputs.

Word2Vec

Word2Vec (W2V) is a neural model introduced by Mikolov et al. (2013a) to create word representations based on their observed occurrences in texts. The model, thus, is based on the Distributional Hypothesis introduced in Section 2.1. The original model was introduced in two variants, named Skip-Gram and Bag-of-Words. In this Section I describe the former, which is the one that I use in my experiments. The idea underlying W2V is to create word representations based on the **prediction** of the linguistic context in which they occur. For example, given the sentence 'we all love Italian pizza', the model uses the central word 'love' to compute the conditional probability of the words in its linguistic context ('we', 'all', 'Italian', 'pizza'), where such a context is defined as the n words on the right and on the left of the target word (in this case, $n = 2$).

The model creates a representation for each word in dictionary D. Initially, each word is represented by two randomly initialized d-dimensional vectors: For word indexed as i in D, its vector is represented as $v_i \in \mathbb{R}^d$ when it is the central, target word (i.e., 'love'), and $u_i \in \mathbb{R}^d$ when it is a context word (the other words in the sentence). The model exploits the vector representations of the words to compute the conditional probability of observing the context words given the central word. The probability of context word c given the central target word w is computed as:

$$p\left(c|w\right) = \frac{\exp\left(u_c^\top v_w\right)}{\sum_{i \in D} \exp\left(u_i^\top v_w\right)} \tag{2.1}$$

Context words are independently predicted given the central word. Therefore, in order to predict all the words in context C, the model maximizes their joint probability, that is, it optimizes:

$$\arg\max \prod_{c \in C} p\left(c|w\right) \tag{2.2}$$

At training time, the model updates the values in the both v and u word vectors in order to maximize equation 2.2. At the end of the training, the v vector of each word is retrieved from the model. The set of vectors returned by the model can be considered as an instance of the **semantic space** described in Section 2.1. Since the model updates word vectors to predict the context, and similar words occur in similar contexts, the vectors associated to similar words will be updated in a similar way. As a result, similar words will end up having vectors that are close in the semantic space.

Word2Vec can be applied to any corpus, to create a semantic space that represents the meaning of words in that corpus. As detailed in Section 2.1, my goal is to create community-specific semantic spaces, and to compare the meaning of words in different spaces, in order to capture variation, or, when analyzing diachronic corpora, change. Both variation and change are identified based on a simple assumption: given two representations of the same word computed in two communities, or in two time bins, the larger the cosine distance between the representations, the stronger the semantic variation, or change.

Unfortunately, different semantic spaces created with W2V cannot be directly compared, due to the stochastic nature of the model. In order to overcome this limitation, several methodologies have been introduced. The one introduced by Xing et al. (2015) leverages the fact that there is a linear correlation between the vectors in two semantic spaces (Mikolov et al., 2013b), and that it is therefore possible to learn a matrix that performs an orthogonal rotation of the spaces and makes the vectors across them comparable, without affecting the pairwise cosine-similarities within each space. Another approach was introduced by Kim et al. (2014): Given a longitudinal corpus split into T consecutive time bins, word embeddings for bin t_i, instead of being randomly initialized, are initialized using those in t_{i-1}, and then updated using the normal Word2Vec model. As a result, vectors in t_i and t_{i-1} are in the same semantic space and, thus, directly comparable. I use this methodology to detect meaning change in Chapter 4. Finally, Bamman et al. (2014a) introduced a methodology to account for semantic variation across the states of the US. In this methodology, the central word indexed as i in D is represented by the vector $v_i \in \mathbb{R}^d$ (as in the original W2V model) and by another state-specific d-dimensional vector: When the target word occurs in a text produced, e.g., in the state of New York (NY), the prediction of its linguistic context is performed based on its general vector v_i and on the vector representation specific for that state $v_{NY,i}$. By jointly optimizing the general and state-specific vectors of the target word, the resulting representations are in the same semantic space and, therefore, they are directly comparable. As we will see, I use this methodology to measure variation across communities in Chapter 3.

Paragraph Vector

Paragraph Vector (PV, Le and Mikolov, 2014) is an extension of Word2Vec that learns vector representations for **documents** of arbitrary length. Similarly to Word2Vec, a randomly initialized vector is initially assigned to each document for which a representation has to be learned. Then, during training, the parameters of the document vector are adjusted in order to maximize the joint probability of the words occurring in it. Thus, the mathematical definition of PV is similar to the one presented in equation 2.2:

$$\arg\max \prod_{c \in C} p\left(c|d\right) \tag{2.3}$$

the only difference being that words in context C are predicted based on document vector d, that is tuned exactly as the word embeddings in W2V. When trained, the document embedding can be regarded as a compact representation of the main information in the document.

In this thesis, I use PV to represent users on social media. While this might sound counterintuitive, the methodology I adopt relies on the assumption introduced in Section 2.2.3: In online setups, and especially on Twitter, users communicate, express opinions, and, in general, define their identities mainly by means of written language.

It seems reasonable, then, to leverage PV to capture this kind of information. I thus adopt this approach in Chapter 6. Concretely, in order to create user representations, I collect all the available posts produced by a user in a single document, and run PV on this document. The output vector can be seen as the representation of the main (textual) features of the user. Similarly to the observation made about word representations in W2V, also in this case, the proximity in the semantic space of two user representations indicates the semantic similarity of their linguistic production.

Long Short Term Memory

Long Short Term Memory networks (LSTMs, Hochreiter and Schmidhuber, 1997) are neural architectures designed to create order-sensitive representations of inputs structured in **sequences**, such as the words in a sentence. LSTMs belong to the family of Recurrent Neural Networks (RNNs, Elman, 1990). Given the sequence S, for the input at time step t, the RNN model computes the hidden state $h_t \in \mathbb{R}^d$ as a function of the vector $x_t \in \mathbb{R}^d$ representing the input and of the vector $h_{t-1} \in \mathbb{R}^d$ representing the previous hidden state:

$$h_t = \tanh\left(Wx_t + Uh_{t-1} + b\right) \tag{2.4}$$

where $\tanh$ is a non-linear transformation, W and U are matrices of learnable parameters, and b is the bias vector. When the input is a sequence of words, the input x_t is the vector representation of the t-th word in the sentence, and the hidden state $h_t \in \mathbb{R}^d$ can be interpreted as the d-dimensional representations of the sequence of words observed up to time t (Tai et al., 2015).

RNNs suffer from the *vanishing gradient* problem, i.e., the situation in which, during gradient-based optimization, the gradient gradually decays, until it approaches zero (Bengio et al., 1994). This problem is particularly evident when the input sequence is long, and its main consequence is that the information from the initial inputs in the sequence is forgotten. The LSTM architecture addresses this problem by using a *memory cell*, whose goal is to avoid the information appearing early on in the sequence to disappear. Also, the model uses an *input gate*, a *forget gate* and an *output gate* to control for the new information to be added to the memory cell and to the hidden state. In practice, gates are just vectors whose values are in range $[0, 1]$, that control how much of the input information is allowed to pass through them (0 means no information will pass, 1 means the whole information flows through the gate). The LSTM function at

time step t is computed as:

$$
\begin{aligned}
i_t &= \sigma\left(W^i x_t + U^i h_{t-1} + b^i\right), \\
f_t &= \sigma\left(W^f x_t + U^f h_{t-1} + b^f\right), \\
o_t &= \sigma\left(W^o x_t + U^o h_{t-1} + b^o\right), \\
c'_t &= \tanh\left(W^c x_t + U^c h_{t-1} + b^c\right), \\
c_t &= i_t \odot c'_t + f_t \odot c_{t-1}, \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

where x_t is the input at time step t, σ is a sigmoid function that squeezes the output in range $[0, 1]$, $\odot$ is an element-wise multiplication, and c'_t is the candidate for the cell state at time step t. Intuitively, the input gate controls how much of the new information is retained, the forget gate controls the extent to which the previous memory cell is forgotten, and the output gate how much of the information in the cell state to incorporate in the hidden state. The joint action of these gates, hence, allows for a more effective control of the information flow, compared to simpler mechanism characterizing RNNs.

LSTMs have been the standard choice in NLP for sentence representation until the recent advent of new architectures based on attention mechanisms (Vaswani et al., 2017). I use LSTMs in several of the experiments in this thesis to model linguistic inputs (see Chapters 3 and 6) and other kinds of sequential inputs (Chapter 5).

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) were introduced in the field of Computer Vision for the task of image classification (LeCun et al., 1998). The main intuition behind CNNs is to use a set of small matrices of learnable parameters called **filters** (or *kernels*) that *convolute*, i.e., slide, on the input image in order to identify in it the features that are relevant for the classification task. Filters have two main characteristics: (i) they are **space invariant**, which means that they can spot the relevant features in the input independently from their position; (ii) each filter **specializes** on a specific feature of the input. These are desirable characteristics also when modeling textual inputs: For this reason, researchers in NLP adopted CNNs, and showed that they achieve strong performance on text classification tasks (Kalchbrenner et al., 2014; Kim, 2014).

Formally, when applied to textual inputs, CNNs work as follows: Given an input sentence S of length n, the input of the CNN is a matrix $M \in \mathbb{R}^{n \times d}$, where d is the dimensionality of the vectors representing the words in the sentence, and row $M_i \in \mathbb{R}^d$ corresponds to the i-th word in the sentence. A convolution operation involves a filter $f \in \mathbb{R}^{h \times D}$ that slides over M, where h indicates the size of the n-grams the filter focuses on. At each step, the filter is applied to an n-gram of h words, and returns a scalar value named *feature*. Feature c_i is generated by filter f from the n-gram x_i in the sentence as:

$$c_i = \sigma\left(f x_i + b\right) \tag{2.5}$$

where b is the bias term, and σ a non-linearity function. By applying the filter to any n-gram of size h in the sentence a *feature map* $c \in \mathbb{R}^{n-h+1}$ is produced. A max pooling operation $c' = \max\{c\}$ is finally applied to the feature map: The n-gram corresponding to c' is the one that the filter identified as the most relevant for the final prediction. A CNN can have one or more convolutional layers, each with an arbitrary number of filters. The output of a layer with k filters is a vector of size $\mathbb{R}^k$, obtained as the concatenation of the c' values returned by the filters in it. Such a vector can be fed either into another convolutional layer, or into a fully connected layer, in order to perform the final classification.

As mentioned above, LSTMs are arguably the best choice when the goal is to encode a sentence in a single representation. However, due to the aggregation of the inputs in the sequence, it is not trivial to identify how much each word contributes to the classification task. This can be done by using CNNs, that allow to clearly determine how much each n-gram in the input text is relevant for the final decision of the model. I leverage this important property of the model in Chapter 7.

2.3.2 Models for Processing Social Networks

In this section, I focus on two neural models that take as input a graph, and return a representation of the nodes (or vertices) and edges in it. These models were developed to encode any kind of graph, such as molecular structures and transportation networks (Scarselli et al., 2009; Hamilton et al., 2017a). I use these architectures to encode **social graphs**, in which nodes are users in online social networks and edges the connections among them. The goal is to create vector representations of the users that reflect their proximity in the graph and, thus, encode the homophily principle introduced in Section 2.1.3. I first introduce the main concepts and practices related to the creation of the social graphs. I then present two models that take as input the social graph, and return a representation for each user in it.

Social Graph

A social graph is a graph representing the connections among users in a social network. In the social graph $G = (V, E)$, V is the set including the users in an online social network, and E is the set of edges between them. While it is straightforward to identify the users in a social network, it is more difficult to define what should be considered as an edge between two users. This choice mostly depends on the social media platform from which data is collected. For example, when the data is obtained from Facebook, a reasonable choice is to create an edge between two users if the two are connected by a **friendship** relation (Ferrara et al., 2012). In other social platforms, however, the friendship relation is not available, and connections are based on different information. When using Twitter, it is common to leverage the **following** relation among users, that has some similarities with the friendship relation on Facebook. Alternatively, researchers have leveraged the **retweet** and **mention** relations, whereby an

edge is instantiated between two users if one retweets or mentions the other (Yang and Eisenstein, 2017). These kinds of relation are particularly interesting because, while the following relation is static and fixed in time, the act of retweeting or mentioning another user is more dynamic and related to specific communicative contexts. For this reason, I leverage these relations to create the social graphs used in the experiments in Chapters 6 and 7. On Reddit, neither friendship/following relations nor retweet or mentions are available. In this case, researchers proposed to define edges based on the interactions among users, namely, when they directly address or reply to each other in a thread (Hamilton et al., 2017b). I adopt this criterion in order to define the social graphs used in Chapter 5.

Node2Vec

Node2Vec (N2V, Grover and Leskovec, 2016) is a neural model that learns vector representations for nodes in graphs. As the name suggests, the model is an extension of Word2Vec: While the latter uses a word to predict the surrounding ones in the linguistic context, N2V leverages the node in a network to predict the ones it is connected to.

More specifically, given the node $v \in V$ and the set of its neighboring nodes $N(v)$, N2V initially assigns to v a randomly initialized vector of size d, that is then updated in order to maximize the probability to predict the nodes in $N(v)$. In N2V, $N(v)$ includes all the nodes encountered in the graph by taking k random walks of length n starting from v, where n and k are hyperparameters of the model. Once the set $N(v)$ is defined, the model again optimizes the objective function:

$$\arg\max \prod_{n \in N(v)} p\left(n|v\right) \tag{2.6}$$

By randomly sampling the nodes in $N(v)$ and by maximizing their joint probability, the model does not make any kind of distinction between connections, i.e., all the neighbors have equal importance in the updating of the parameters of the vector representing v. As shown in Chapter 6, this can be a limitation when representing users in social networks, because the model does not consider that different neighbors can have different relevance, depending on the situation.

Graph Convolutional Network

Graph Convolutional Networks (GCNs, Kipf and Welling, 2017) are neural architectures belonging to the family of Graph Neural Networks (Scarselli et al., 2009; Hamilton et al., 2017a). The general feature of the models belonging to this class is to define the representation of a node in the graph by aggregating the information coming from its connections. In GCNs, node $v \in V$ is initially represented by vector h_v^k, whose update is computed as:

$$h_v^{k+1} = \sigma\left(\sum_{u \in N(v)} W^k h_u^k + b^k\right) \tag{2.7}$$

where W^k and b^k are the layer-specific parameters of the model, $N(v)$ is the set of neighbors of the target node v, h_v^{k+1} the updated node representation, and σ a non-linearity function. In GCNs, $N(v)$ is defined based on the number of layers in the model. Hence, at layer 1, $N(v)$ includes first-degree connections, i.e., the direct connections of the target node. At layer two, $N(v)$ includes second-degree connections, and so on. Thus, the number of layers defines the depth of the neighborhood. Also, $v \in N(v)$, which means that the target node v is included in the set of the its connections. This ensures that the initial representation of the node h_v^k is considered during its update.

By implementing the aggregation step as a weighted sum of all the neighbors, GCNs let each of the neighbors contribute equally to the update of the target node. Similarly to what happens with N2V, hence, no distinction is made among neighboring nodes. In order to make this distinction possible, an extension of the original model, named **Graph Attention Network** (GATs), was introduced by Velickovic et al. (2018). GATs use a self-attention architecture (Bahdanau et al., 2015; Parikh et al., 2016; Vaswani et al., 2017) that is able to assign different importance to different nodes within the neighborhood. For the target node $v \in V$, an attention coefficient e_{vu} is computed for every neighboring node $u \in N(v)$ as:

$$e_{vu} = a\left(h_v \| h_u\right) \tag{2.8}$$

where h_v and $h_u \in \mathbb{R}^d$ are the vectors representing v and u, $\|$ indicates a concatenation operation and a is a single-layer feed-forward neural network, parametrized by learnable weight matrix $W^a \in \mathbb{R}^{2d}$. The attention coefficients computed for all the neighbors are normalized using a softmax function. In GATs, then, the update of the target node v is computed as:

$$h_v^{k+1} = \sigma\left(\sum_{u \in N(v)} \alpha_{vu}^k \left(W^k h_u^k + b^k\right)\right) \tag{2.9}$$

where the α_{vu} coefficient acts as a weight that defines how much neighbor u should contribute to the update of the vector representing v. The model can use several attention mechanisms (*heads*) in order to stabilize the learning process: in this case, given n attentions mechanisms, n real-valued vector $h_v^{k+1} \in \mathbb{R}^{d'}$ are computed and, subsequently, concatenated, thus obtaining a single embedding $h_v^{k+1} \in \mathbb{R}^{n*d'}$.

I use GATs in the experiment presented in Chapter 6, in order to model the different relevance of the connections of a user in a social graph, depending on the communicative situation.

Part One

Analysis of Linguistic Variation in Online Communities

This first part of the thesis analyses lexical variation in online communities of speakers by means of three related experiments. The first experiment focuses on the cornerstone of this dissertation, as it presents a framework that allows to identify and measure the **meaning variation** of common words in online communities of practice. The second chapter investigates the diachronic processes that lead to the observed variation. In particular, it focuses on the **meaning shift** taking place in short periods of time, investigating the linguistic processes related to it, and assessing the performance of standard NLP tools in detecting such a shift. Finally, the third experiment focuses on the **social dynamics** related to the diachronic spread of linguistic innovations in online communities. In this chapter, I build on sociolinguistic theories to uncover the role played by different kinds of user in the process of introduction and spread of new linguistic practices.

Chapter 3

Meaning Variation in Online Communities of Practice

The content of this chapter is based on the following publication:

Marco Del Tredici and Raquel Fernández. 2017. Semantic Variation in Online Communities of Practice. In *12th International Conference on Computational Semantics (IWCS)*.

The two authors jointly produced the idea for the article. Marco performed the experiments with Raquel's supervision. Marco wrote the article, Raquel provided guidance and a substantial contribution to the writing. The text in this chapter minimally overlaps with the one of the original publication.

3.1 Introduction

The sociolinguistic studies introduced in Chapter 2.1 show that communities of speakers exhibit **lexical semantic variation**, the phenomenon whereby the same word has different meanings in different communities of speakers. In this chapter, I focus on lexical variation in online setups. More precisely, I address the first research question introduced in Section 1.2, and investigate how to automatically identify and represent meaning variation in **online communities of speakers**.

I introduce a framework based on computational and statistical tools to measure variation in the language produced by several online communities of English speakers on Reddit (see Section 2.2.1). All the communities under examination can be considered as examples of **communities of practice**, since they are created by people who have a common goal and that voluntarily engage in common activities. Accordingly, members of these communities are expected to develop specific linguistic practices, different from those developed in other communities. The proposed framework aims at capturing these community-specific practices. In particular, it allows quantifying the meaning variation of **common words** in different communities. Since the meaning assigned to a word can vary depending on the discussed topic, I control for this factor by considering different communities concerned with the same topic.

The results obtained by applying the framework show that, while it is possible to observe topic-related variation, different online communities develop different meanings for the same common words, even if they are discussing the same topic. Such variation is especially evident in small communities. The reported results, hence, provide empirical evidence for the previous findings in Sociolinguistics. A quantitative evaluation of these results is presented, in which it is shown that the semantic variation detected by the framework significantly affects the performance of an independent Language Model.

Finally, I investigate the relation between meaning variation and the social dissemination of words, showing that while community specific meanings are used by niches of users, meanings that are common to related communities are widely disseminated.

3.2 Related Work

The current study is highly related to the descriptive works in Computational Sociolinguistics introduced in Section 2.2.2, with which it has in common the basic goal of describing the interplay between linguistic and social variables in online setups. However, when the idea for this study was produced, we aimed at addressing two aspects of online lexical variation that, until that moment, where under-researched in the field.

First, all the related works in Computational Sociolinguistics shared the same approach, whereby the language produced by the users was leveraged as input to a model, whose goal was to predict the social features of the users, such as gender, power differences, or community permanence. In all these studies, then, the focus was on the

community	years	tokens	members
programming	10	21M	72K
Python	8	18M	41K
learn.prog	7	21M	61K
soccer	8	65M	146K
LiverpoolFC	8	55M	29K
reddevils	6	66M	36K
global	–	50M	445K

Table 3.1: Main statistics of each community dataset: years of activity, tokens in the derived dataset, number of members.

social factors characterizing users, and how language could be leveraged to uncover them. Differently, the main focus of the experiments presented in this chapter is on linguistic variation in its own right, that is, the goal is to show how meaning variants can be efficiently identified and measured across communities of speakers in online setups.

Second, the previous studies had concentrated almost exclusively on community-specific **jargon** and **slang**, that is, on the neologisms, acronyms and abbreviations that are created and shared only by the users of one community – e.g., 'dx' for 'diagnosis' in breast cancer discussion forums (Nguyen and Rosé, 2011) or 'scrim' for 'practice match' in online gaming (Kershaw et al., 2016). None of these studies, however, considered the fact that social interaction among speakers often leads to semantic variation of **common words**, that is, words that *"belong to many communal lexicons, though with very different conventional meanings"* (Clark, 1996). The present study concentrates on this type of words.

After the publication of the work that forms the basis for this chapter, several other studies focused on linguistic variation in online setups. Interestingly, these works share with the current one the focus on common words. For example, Miletic et al. (2020) adopt the methodology presented in Section 3.4.1 to model semantic variation in Quebec English, leveraging a Twitter dataset, while Oba et al. (2019a) and Oba et al. (2019b) investigate interpersonal variation, that is, they analyze how the meaning of the same common words change when used by *different* individuals within the *same* community. These works, hence, focus on variation taking place at a level that is different, but clearly related, to the one I investigate here.

3.3 Data

To investigate semantic variation in online communities, I collect data from subreddits in Reddit. Subreddits are created by people that share a common interest, and that voluntarily decide to group together in order to engage in a common endeavor. For this

reason, subreddits can be considered as online instances of **Communities of Practice** (see Section 2.1.3). I thus expect that in subreddits the same process of meaning creation typical of Communities of Practice takes place, and, accordingly, to observe variation across different subreddits.

Data for six subreddits are collected. Provided that variation is observed in different communities, a possible issue might be that such variation is due to the topic discussed by these communities. In order to control for the effect of topic, I select two groups of communities belonging to the same **domain**, i.e., discussing the same topic. The first group includes three communities concerned with the domain of Information Technology:[1] `programming`, `Python` and `learnprogramming`. While the first community includes users interested in programming in general, the second includes users interested in a specific programming language, while the third users approaching programming. The second group is made up of three communities in the domain of Football: `soccer`, `LiverpoolFC`, and `reddevils`. The first community includes users passionate about soccer (i.e., football), while the others are made up of fans supporting specific football teams, namely Liverpool FC and Manchester United. Both groups include two smaller communities (`Python` and `learnprogramming` in Information Technology; `LiverpoolFC` and `reddevils` in Football) and two larger communities, that, at least to some extent, include the smaller ones. In particular, $\sim$25% of the users in `learnprogramming` and `Python` are also in `programming`, while $\sim$45% of the users in `LiverpoolFC` and `reddevils` are also in `soccer`. The rationale behind this choice is to account, within each domain, for the hierarchical structure defined by Clark (1996), whereby communities of different size are nested in a hierarchical structure, and to investigate lexical variation at different levels of such structure (see Section 2.1.2).

For each community, the contents created by all members during its whole lifespan are crawled. Since the resulting datasets have different sizes in terms of number of tokens, some of them are randomly subsampled (`soccer`, `programming`, and `learnprogramming`) in order to make them comparable in size to the other datasets in the same domain. For this experiment, the dataset obtained for each community is considered synchronically as a whole, thus abstracting away from the diachronic dimension of the data, that I will investigate in the next chapters.

The datasets obtained for each community are used to create community-specific word representations. However, in the introduced framework, community-independent word representations are also needed to measure semantic variation. To obtain these representations, an additional dataset is created by randomly crawling posts and comments exchanged within any of the existing subreddits in Reddit during January 2017. A sample of community-independent linguistic practices is thus obtained. This sample can be considered as a proxy for general language use. I refer to this dataset as the `global` community. The `global` community includes 50 million tokens from hundreds of thousands of different subreddits, contributed by more than 445k different

[1] In the rest of the chapter, I use capitalized names to indicate the domain of a group of communities.

users. Less than 1% of these users are members of the target communities defined above. Table 3.1 summarizes the main statistics for all the communities.[2]

3.4 A Framework for Meaning Variation

In this section, I define the framework that I designed to measure semantic variation in online communities. The framework is based on two components: **community-specific word representations** and **statistical indices** that measure the distance between word representations. I first introduce the model that I use to create community-specific word representations (Section 3.4.1). Subsequently, I describe the two indices that I use to compute semantic variation in single communities and in groups of communities belonging to the same domain (Section 3.4.2).

3.4.1 Word Representations

In order to create community-specific word vectors, I use the model introduced by Bamman et al. (2014a). The model is an extension of the Word2Vec architecture introduced in Section 2.3.1, and in the original study is used to represent meaning variation in the states of the United States of America. In the model, the central word i used to predict the context words is represented by a **general** vector $v_i \in \mathbb{R}^d$, as in the original W2V model. Additionally, given the set of states $S = \{AK, AL, ..., WY\}$, including $|S|$ items, the same word is represented by $|S|$ state-specific d-dimensional vectors, where each vector is meant to represent the target word in a **specific** state. The prediction of the surrounding words is then performed based on the general *and* state-specific vectors: For example, when the target word occurs in a text produced in the state of New York (NY), the prediction of its linguistic context is performed based on its general vector v_i and on the vector representation specific for that state $v_{NY,i}$. This methodology captures the fact that the target word has a general representation v_i shared across all the states, that is updated at any occurrence of the word, and a state-specific representation, e.g., $v_{NY,i}$, that captures the meaning variation related to each state. By jointly optimizing the general and state-specific vectors of the target word, all the state-specific representations are in the same semantic space and, therefore, they are directly comparable. While in the original work the methodology is used to model variation related to geographic information, it can be applied to any other kind of variation. This is what I do, as I substitute the set of state-related texts in the original study with a set of texts generated in the Reddit communities defined in Section 3.3.

Before running the model, the datasets of each community is preprocessed by applying tokenization and removing words that appear less than 100 times. The model is trained using the default parameters of the original implementation.[3]

[2]Here and in the rest of the chapter, I use 'communities' to refer to both the six target communities and the `global` community.

[3]The original implementation is available here: `https://github.com/dbamman/geoSGLM`.

3.4.2 Meaning Variation Indices

Let C denote a set of communities of practice and g the global community. I use subsets such as $D \subset C$ to denote sets of communities related by a certain domain, i.e., Information Technology or Football. The model described above returns word embeddings w_c, w_g for each word w and community c in a domain D, encoding how w is used within that community and in the global community g, respectively.

For any two vectors $v, v' \in \mathbb{R}^k$, let $\text{sim}(v, v')$ denote their cosine similarity. Given two sets of communities $A, B \subseteq C \cup \{g\}$, I use $\text{Sim}^w_{A,B}$ to refer to the following multiset of similarity values for word w:[4]

$$\text{Sim}^w_{A,B} = \{\text{sim}(w_a, w_b) \mid (a, b) \in A \times B \text{ with } a \neq b\} \tag{3.1}$$

Let S and S' be two such multisets of similarity values. To measure the extent to which these values are higher in S than in S', I use the following index, where μ and σ are the mean and the standard deviation, respectively:

$$\mathbb{I}(S, S') = [\mu(S) - \sigma(S)] - [\mu(S') + \sigma(S')] \tag{3.2}$$

This generic index can now be used to construct several specific indices to quantify different types of semantic variation. Note that the proposed indices are designed to capture variation across communities belonging to the *same* domain, without considering variation across domains.

Community variation index (cvi) The first index quantifies the degree to which a given word exhibits variation specific to a community, i.e., that is not shared by other communities in the same domain D. This type of semantic variation is particularly interesting because, when present, it shows that meaning variants can arise in a community independently from the topic discussed. In particular, the community variation index aims at capturing situations in which the meaning of a word w in a community $c \in D$ has drifted away from its general use in g, while in other domain-related communities the meaning remains close to that observed in the global community. The index is defined as:

$$\mathbf{cvi}^w_c(D) = \mathbb{I}(\text{Sim}^w_{D\setminus\{c\},\{g\}}, \text{Sim}^w_{\{c\},\{g\}}) \tag{3.3}$$

where $D\setminus\{c\}$ denotes the set of communities in domain D except for c. For words with positive $\mathbf{cvi}^w_c(D)$ values, the higher the index, the stronger the variation in c relative to other domain-related communities.

[4] In practice, in case $A = B$, I only compute one cosine similarity value for every unordered pair rather than for every ordered pair. This does not affect either the mean or the standard deviation of the multiset.

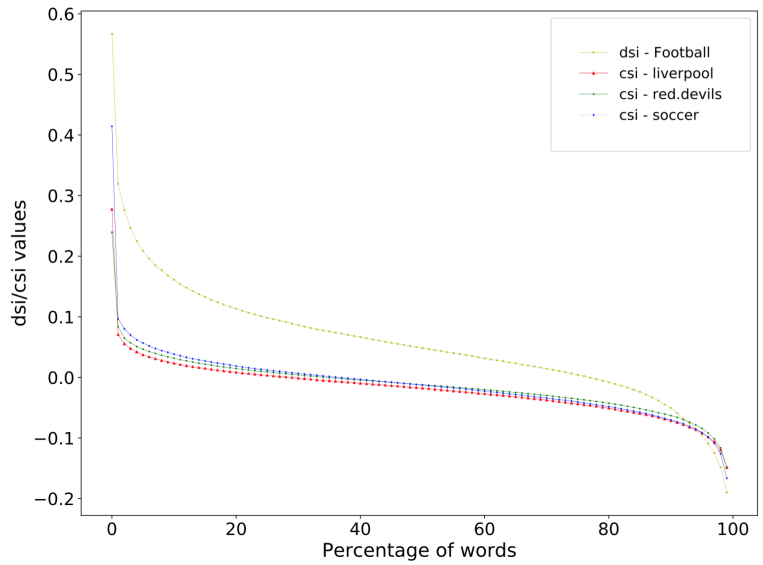


Figure 3.1: The **dvi** and **cvi** values (on the y-axis) for the Football domain and the communities which belong to it. On the x-axis the number of words (in percentage): Since I consider only words in the shared vocabulary, the total amount of words is the same for all the communities and for that domain.

Domain variation index (dvi) The second index aims at spotting words whose meaning is similar in a set of communities belonging to the same domain, and distinct from the one in the global community. Such an index, hence, is used to assess how much variation is due to the fact that related communities discuss the same topic. The domain variation index for a given domain $D \subset C$ and word w is computed as:

$$\mathbf{dvi}^w(D) = \mathbb{I}(\text{Sim}^w_{D,D}, \text{Sim}^w_{D,\{g\}}) \tag{3.4}$$

Again, for words with positive **dvi** values, the higher the index, the more pronounced their semantic variation across a domain with respect to the language use of the global community.

3.5 Observed Variation

The community-specific word embeddings defined in Section 3.4.1 are used to compute the **cvi** and **dvi** values for all the words in all the communities in the Information Technology and Football domain. For each domain, only the words shared among all the communities in the domain are considered.

Figure 3.1 reports the distribution of **dvi** values for the Football domain and the **cvi** values of the communities belonging to the domain. Similar results are found for the Information Technology domain and its communities. The left tails of the distributions for `soccer`, `LiverpoolFC`, and `reddevils` indicate that there is a set of

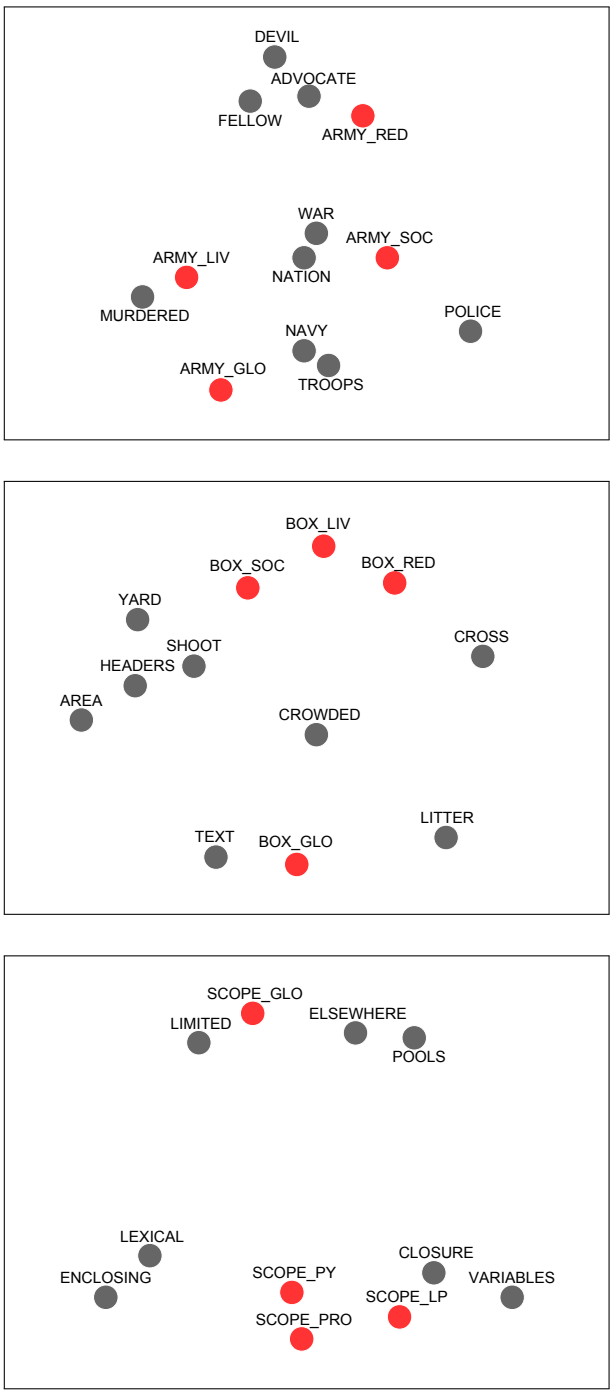


Figure 3.2: Two-dimensional representation in semantic space of meaning variants for words 'army' (high **cvi** in `reddevils`, top), 'box' (high **dvi** in Football, middle), and 'scope' (high **dvi** in Information Technology, bottom). In the plots: GLO = global community, SOC = `soccer`, LIV = `LiverpoolFC`, RED = `reddevils`, PRO = `programming`, PY = `Python`, LP = `learnprogramming`.

words with high **cvi** values in each of these communities. This means that, as hypothesized, in each community there are words that are used with a meaning that is different compared to the one used in the other communities of the same domain. As an example, let us consider the word 'believers', that has a high **cvi** index for the `LiverpoolFC` community. The target word has very similar meanings in the global community and in the other communities in the Football domain (cosine similarity is 0.86 for `global-soccer` and 0.8 for `global-reddevils`), while the meaning is very different in `LiverpoolFC` (similarity `global-LiverpoolFC`=0.56). The word is used with its conventional meaning in `global`, `soccer` and `reddevils`, as shown by the fact that in the semantic spaces of these communities the closest neighbors of 'believers' are, e.g., 'religion', 'beliefs', 'spiritual'. Conversely, in `LiverpoolFC` 'believers' is close in the space to 'doubters', 'we' or 'scousers', that is, to words used by Liverpool FC fans to refer to themselves. Manual inspection of the dataset shows that this usage of 'believers' spread after an interview of the team coach, in which he invited the supporters to *believe* in the team and to 'change from doubters to believers'.[5] This expression remained impressed on supporters' minds, who, as we will see in the next chapter, rapidly adopted it.

Plenty of examples similar to 'believers' can be found in each community. In `reddevils`, for example, the words 'army' and 'theater' are used to denote the supporters and the stadium of Manchester United, respectively, while the words occur with their conventional meanings in `global`, `LiverpoolFC` and `soccer`. In Figure 3.2 (top) the positioning in the semantic space of 'army' is shown.

The left tail of the distribution for Football in Figure 3.1 shows that there are also words that present a common variation in the three communities in the domain. An example of this case of variation is the word 'box', that is highly similar in the three communities in the domain (`soccer-LiverpoolFC`=0.87, `soccer-reddevils`=0.89, `LiverpoolFC-reddevils`=0.89), while it is always very different between these communities and `global` (`global-soccer`=0.38, `global-LiverpoolFC`= 0.36, `global-redreddevils`=0.36). The representation of the word 'box' for `global` conflates its conventional meanings, as it can be inferred by the fact that its closest neighbors are words such as 'text' (related to the usage of 'box' as a text box) and 'cardboard' (when it is used to refer to the physical object). Differently, in all the Football communities the word is used to indicate a specific area of the pitch, namely, the penalty area. For this reason, in the semantic space of each of the football-related communities 'box' is close to words like 'cutback', 'shoot' and 'cross', i.e, terms that refer to actions of the game that typically take place inside the penalty area. This is visually shown in Figure 3.2, middle. The same situation is observed in the Information Technology domain. For example, as shown in Figure 3.2 (bottom), the word 'scope' is used with its conventional meaning in `global`, while in the communities of the domain is used to talk about the property of a variable.

[5] `https://www.thisisanfield.com/2019/04/doubters-to-believers-how-jurgen-klopp-made-and-delivered-his-promise-to-liverpool-fans/`.

As it can be observed in Figure 3.1, all the distributions present a common pattern, whereby few words show a strong semantic variation (left tail of the distributions), while the majority of the words present a small or null variation, corresponding to **dvi / cvi** values included in the range between 0 and 0.2 (central part of the distributions). This pattern reflects the intuitive fact that the majority of words are used with their conventionalized meaning in communities and domains, while some of them show variation.

The right tail of negative values observed for all the distributions has different interpretations for domains and communities. Negative values of **dvi** are assigned to the same words that have high **cvi** values, i.e., words that show a strong variation in just one of the communities part of the general domain. Differently, for each community, negative **cvi** values are assigned to words that undergo strong semantic variation in *another* community of the same domain.

Finally, for both Information Technology and Football, **dvi** values are, on average, larger than **cvi** ones. Arguably, this is due to the fact that **dvi** captures the variation in the domain vocabulary compared directly to the global community, while **cvi** represents the more subtle variation within communities belonging to the same domain.

3.6 Quantitative Evaluation

In the previous section, I leveraged the information captured by community-specific word representations to show that meaning variation is observed in communities of speakers. I now turn to an extrinsic evaluation of the information captured by word representations, that is carried out by leveraging the properties of a Language Model.

3.6.1 Method

As a first step, the dataset of each community is randomly split into training, validation, and test sets (70/15/15). Then, a Language Model (LM) is trained on the train set of each community independently. Importantly, the LM for a community is initialized with the word embeddings previously created for that community (see Section 3.4.1). At test time, given word w in a set of target words, the average perplexity when the word w is used by the LM to predict the upcoming word in the sentence is computed (ppl^{w}_{train}). Subsequently, the original embedding for w is substituted with an **alternative embedding** learned from another community, and compute again the perplexity when predicting the upcoming word (ppl^{w}_{alt}). The change in performance is then measured as the relative perplexity increase:

$$\mathrm{ppl}^{w}_{change} = \frac{\mathrm{ppl}^{w}_{alt} - \mathrm{ppl}^{w}_{train}}{\mathrm{ppl}^{w}_{train}} \tag{3.5}$$

The rationale behind this methodology is the following: If the semantic information provided by the original and alternative vectors is the same, or highly similar, the

change in perplexity will be null or negligible when swapping the two vectors. Conversely, if the two vectors are different, i.e., the word they represent show variation in the two communities, the performance of the model will significantly degrade, and the change in perplexity will be relevant.

I apply this methodology to different sets of words extracted from communities and domains. In particular, for each community, a VARIATION set including the ten words with the highest **cvi** in that community, and a NO.VARIATION set including the ten words with the lowest **cvi** are defined. Similarly, for each domain D, the two sets VARIATION and NO.VARIATION containing the 10 words with the highest and lowest **dvi** values, respectively, are defined. Given these sets of words, I apply the general methodology defined above in different ways.

Recall that for communities variation is defined as the situation in which the meaning of word w in c has drifted away from w's use in the global community, while in the other domain-related communities the meaning is similar to the one in the global community. Accordingly, I leverage the LM trained on `global`, and I hypothesize that using embeddings of VARIATION words in community c as alternative embeddings (e.g., 'believers' in `LiverpoolFC`) will yield significantly higher perplexity than using alternative embeddings from other communities within the same domain. In all cases, I expect that for NO.VARIATION words (i.e., words for which there is no semantic variation according to **cvi** index) the change in perplexity with different embeddings will be negligible.

For domain, variation is defined as the situation in which the meaning of word w is the same in all communities belonging to D, and different from the meaning in the global language. Thus, given the LM of a community, I hypothesize that, for VARIATION words, a significant increase in perplexity will be observed when testing on alternative embeddings belonging to the general community (e.g., when substituting 'box' in `LiverpoolFC` with the representation from `global`), while no significant difference should be observed when testing on alternative embeddings belonging to another domain-related community.

The $\mathrm{ppl}^{w}_{change}$ values are calculated for VARIATION and NO.VARIATION words as described above for all the communities and for the two target domains.

3.6.2 Language Models

The community-specific Language Models are implemented using an existing encoder-decoder LSTM.[6] All the models have 2 layers of size 200. The models are trained for 40 epochs, using Adam optimizer (Kingma and Ba, 2015) for parameter update and dropout for regularization. All the community language models reached an average test perplexity between 45 and 67 on the task of predicting the upcoming word given the preceding word (window size = 1). While it is very likely that the model used in

[6]I use the implementation available at: `https://github.com/pytorch/examples/tree/master/word_language_model`.

	VARIATION		NO.VARIATION	
community	$c \rightarrow g$	$D \backslash \{c\} \rightarrow g$	$c \rightarrow g$	$D \backslash \{c\} \rightarrow g$
`programming`	3.87	0.24	10.05	4.92
`Python`	26.83*	6.73	8.83	0.68
`learn.prog`	56.77*	11.85	13.28	9.57
`soccer`	8.92	11.32	11.93	13.33
`LiverpoolFC`	17.70*	2.45	5.31	3.84
`reddevils`	55.84*	4.98	5.60	2.98
domain	$g \rightarrow c$	$D \backslash \{c\} \rightarrow c$	$g \rightarrow c$	$D \backslash \{c\} \rightarrow c$
Information Tec.	64.9*	6.04	9.02	5.77
Football	40.47*	2.40	-1.04	-0.78

Table 3.2: Perplexity increase medians in each setting. * indicates a significant difference according to a Wilcoxon signed-rank test with $p < 0.05$. At the community level, the following setups are implemented: $c \rightarrow g$: the embedding for word w in community c is fed into the global LM; $D \backslash \{c\} \rightarrow g$: the embedding for w in a community related to c is fed into the global LM. At the domain level: $g \rightarrow c$: the embedding for word w in the global community is fed into the LM of community c; $D \backslash \{c\} \rightarrow c$: the embedding for w in a community related to c is fed into the LM of community c.

this study is outperformed by current models based on attention mechanisms (e.g., the BERT architecture introduced by Devlin et al. (2019)), at the time this experiment was conducted the performance was in line with the state of the art (Zaremba et al., 2014).

3.6.3 Results

Table 3.2 shows an overview of the results. The reported values are the median $\mathrm{ppl}^{w}_{change}$ of each setup. Statistically significant differences between setups are computed using Wilcoxon signed-rank test ($p < 0.05$).

For words that show high variation at the community level (VARIATION), my hypothesis is confirmed for `LiverpoolFC`, `reddevils`, `Python` and `learnprogramming`, as a significant increase in perplexity is always observed when the embeddings from these communities are used with the global LM ($c \rightarrow g$). Conversely, no significant increase in perplexity is observed for `programming` and `soccer`. Also, as expected, no significant difference in perplexity is observed for words with low **cvi** values (NO.VARIATION).

There seems to be a clear pattern in these results: The meanings specific to the communities that are smaller and lower in the hierarchical organization of the communities are not *understood* by the LM trained on the general language, while the meanings in wider communities, positioned at higher levels, are. These results are in line with the theoretical framework defined by Clark (1996): Linguistic innovations

are created at the bottom of the hierarchy, and, as they spread beyond the specific community in which they originated, they move up in the hierarchy. It is therefore possible that some of the meanings proper to communities such as `soccer` and `programming`, that are higher in the hierarchy, have already been accepted in the general language, while those in the lower communities have not.

One might wonder how it is possible that, in `soccer` and `programming`, the same words have high **cvi** values but do not lead to any change in perplexity when swapped in the LM. This is arguably due to the nature of Word2Vec and LSTM models (see Section 2.3.1): The embeddings created with the former conflate all the usages of a word in a single representations, therefore, even if a community-specific meaning (e.g., the one of 'scope' in `programming`), has already been introduced in the general language, this meaning will be much less frequent than the standard meaning, and hence not be represented in the embedding of the word. Differently, the LSTM, by encoding the specific context in which a word is used, is able to correctly represent also meanings that are rare in the global language.

Regarding domain variation, as predicted, for words with low **dvi** values (i.e., those in the NO.VARIATION set), no significant difference in perplexity is observed when different embeddings are used. In contrast, for words with high **dvi** values (VARIATION) the increase in perplexity is always significantly higher when the original embeddings of a community are substituted with those of the general language ($g \rightarrow c$), while perplexity remains reasonably stable when the alternative embeddings come from another domain-related community ($D \backslash \{c\} \rightarrow c$).

3.7 Social Dissemination

In this last section, I consider the social dimension of meaning variation. In particular, the section investigates the relation between meaning variation and **social dissemination** of words, that is, the proportion of community members using them, that has already been shown to be predictive of changes in word frequency over time (Altmann et al., 2011). The dissemination of a word within a community c is computed as follows:

$$\mathrm{Dis}(w, c) = (U_c^w / U_c) \times (1 - \mathrm{RelFreq}(w, c)) \tag{3.6}$$

where U_c^w is the number of community members who use word w and U_c the total number of members in community c. Since words with very high frequencies (such as function words) will be used across the board, the ratio U_c^w / U_c is weighted by the inverse of w's relative frequency. Dissemination in a domain $\mathrm{Dis}(w, D)$ is calculated equivalently.

Similarly to what was done in the previous section, the dissemination value is computed for words that exhibit a strong semantic variation (VARIATION), and words that do not show variation (NO.VARIATION), both at the community and domain level. An unpaired two-sample t-test is then used to assess if a statistically significant difference

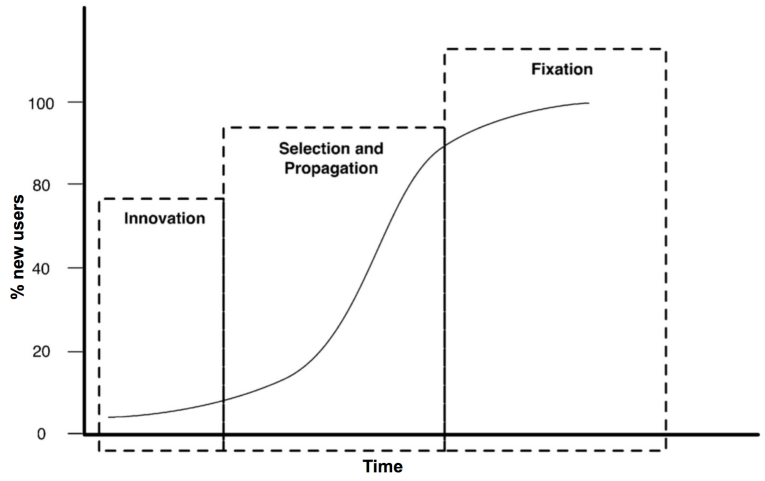


Figure 3.3: The s-shaped curve representing the diffusion of social innovations. The plot is taken from Fagyal et al. (2010).

exists between the two sets of words. In order to obtain more reliable results, VARIATION includes words with **cvi/dvi** value equal or larger than 2 standard deviations above the mean within a community or domain, while NO.VARIATION includes words with index values lower than one standard deviation above the mean.

Results show that VARIATION words in `Python`, `learnprogramming`, `LiverpoolFC` and `reddevils` are *less* disseminated within that community than words that do not exhibit a variation. No significant difference is observed for `soccer` and `programming`, a result that indicates again the difference between communities at different levels of the hierarchical structure defined above. Finally, at the domain level VARIATION words are *more* disseminated than those with low **dvi**.

I see these findings as related to the general process of linguistic innovation and diffusion described in Chambers and Trudgill (1998), and usually represented by the sigmoid function shown in Figure 3.3. Linguistic variants originate among and are initially adopted by a circumscribed number of members. At this stage ('Innovation') few users use the innovation, that is therefore not highly disseminated in the community. My intuition is that the **cvi** index captures innovations that are in this phase. I will explore this intuition in the next chapter, which focuses on the innovation phase in the first part of the curve. Some variants may then rapidly spread within the community ('Selection and Propagation'), until they reach a plateau ('Fixation'), and possibly spread to other domain-related communities. This is the stage that is captured by the **dvi** index. The innovation, at this point, has been largely adopted, and, consequently, has a high dissemination value.

3.8 Conclusion

In this chapter, I investigated linguistic variation in online communities of speakers. I introduced a framework that allows to measure how the meaning of common words varies in different communities of speakers, and in groups of related communities. By applying this framework, I found that while communities concerned with the same topic share some meaning variants, it is also possible to observe community-specific variation. These findings were evaluated using an external Language Model. From this evaluation, it emerged that community-specific linguistic variation is more evident in smaller communities, positioned at a lower level in the hierarchical structure proposed by Clark (1996). Finally, I investigated the correlation between meaning variation and social dissemination of words, finding that words that show high variation in communities and domains are at different stages of spread. While the current study makes several relevant contributions to the general aim of this dissertation, I am aware of several of its limitations, like, for example, the limited number of communities and domains taken into account and the simplicity of the computational architectures used to measure variation. Despite these shortcomings, the present chapter has two main merits: First, it assesses the existence and the relevance of the linguistic phenomenon this whole dissertation is about, i.e., meaning variation in online communities of speakers. Second, it makes some interesting questions emerge, among which the most relevant is: What are the diachronic processes that lead to the meaning variation observed in this study? Such a question forms the basis for the next two chapters.

Chapter 4

The Genesis of Variation: Short-Term Meaning Shift in Online Communities

The content of this chapter is based on the following publication:

Marco Del Tredici, Raquel Fernández and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

The three authors jointly produced the idea for the article. Marco performed the experiments with Gemma's supervision. Marco wrote the article, Gemma and Raquel provided guidance and contributed to the writing. The text in this chapter partially overlaps with the one of the original publication.

4.1 Introduction

The synchronic analysis presented in the previous chapter shows that lexical semantic variation is at play in online communities of speakers, and that such variation can be measured by means of computational tools. In particular, it shows that the same common word can have several community-specific meanings, that are different from the conventional meaning in the general language. As explained in Chapter 2, the synchronic situation presented in the previous chapter is the outcome of a diachronic process of **meaning shift**. In such a process, new meanings are created by individuals during communication and, in some cases, adopted by the other members of the community. Crucially, the adoption of a new meaning variant takes place quickly in communities of speakers, as indicated by the steepness of the central part of the sigmoid function shown when discussing social dissemination in Section 3.7. In this chapter, I focus on this process, and address two questions: (i) which are the linguistic phenomena related to the emergence of new meanings in online communities of speakers? (ii) in NLP and CL, computational models for meaning shift detection have been created to identify shift over long periods of time: are these models suitable also for detecting the quick emergence and adoption of new, community-specific meanings? These two questions, then, focus on the linguistic aspects of RQ-2 defined in Section 1.2.

In order to answer the questions above, I focus on the language produced by one of the communities considered in the previous chapter, namely, `LiverpoolFC`. I split the community dataset in time bins, and collect annotations for the presence of **short-term meaning shift** (**STMS**), that is, the community-specific meaning shift that, as said, takes place in short periods of time. The annotated dataset is then used to answer the two questions above. First, I analyze the main linguistic phenomena related to short-term meaning shift: metonymy, metaphor, and meme. Subsequently, the dataset is used to test the performance of a standard distributional model of semantic change when applied to short-term shift. The results show that the model is able to detect the cases of semantic shift in the data. However, the model strongly overgeneralizes, that is, it labels as semantic shift contextual changes of **referential** nature, in which words change their context of use due to users in the community often talking about particular people and events. In order to remedy this problem, I introduce a measure of contextual variability, that allows to identify words that change context of use not as an effect of meaning shift, but due to specific referential uses.

4.2 Related Work

The study presented in this chapter belongs to the line of studies in CL and NLP concerned with the investigation of how the meaning of words change diachronically. This research topic has received large attention in the last few years, and several models and methodologies have been proposed – see Kutuzov et al. (2018) for an overview. The present study shares with the works in the field the basic assumption whereby

the change in meaning of a word is mirrored by a change in its context of use. This assumption, in turn, stems from the Distributional Hypothesis introduced in Section 2.1.1. Based on this assumption, I implement a model that spots semantic shift based on how much the vector representation of a word changes in different time periods. I thus follow the general approach described in Section 2.3.1, which has been shown to be effective in several studies (Kim et al., 2014; Kulkarni et al., 2015; Del Tredici et al., 2016; Hamilton et al., 2016; Azarbonyad et al., 2017; Phillips et al., 2017; Szymanski, 2017).

There are, however, two main differences between the current study and those that were present in the literature before this was designed and implemented. First, the observed **time span**. Previous studies focused on meaning shift taking place on long periods of time – decades to centuries. While several interesting phenomena take place in such long periods of time, it is only by looking at shorter time spans that it is possible to observe the genesis of meaning shift and the mechanisms that produce it. I focus on this short time spans. Second, the **kind of language** taken into account. While previous studies analyze the standard language of books or newspapers, I focus on the language produced by online communities of speakers. This is a natural choice given that, as said, it is in communities of interacting individuals that new meanings are created, before, eventually, being adopted in the general language.

4.3 Experimental Setup

4.3.1 Data

For the study presented in this chapter, I focus on the data derived from one of the communities introduced in the previous chapter, namely, `LiverpoolFC`, an online community of football fans active on Reddit. This community presents many characteristics that favor the creation and spread of linguistic innovations, such as a topic that reflects a strong external interest and high density of the connections among its users (Hamilton et al., 2017b). The `reddevils` community would have been an equally suitable choice, as it presents the same characteristics as `LiverpoolFC`. What guided me in the final choice was only a personal preference.

I focus on the language produced in the community in the period between 2011 and 2017. In order to enable a clear observation of short-term meaning shift, two non-consecutive time bins are defined: the first one (t_1) contains data from 2011–2013 and the second one (t_2) from 2017. By using the data produced in a time span of two years for the first time bin it is possible to obtain data samples for the two time bins that are approximately of the same size.

Similarly to what was done in Chapter 3, also in this case a large sample of community-independent language (`global`) is collected by crawling data from random subreddits in 2013. As shown in the next section, this data will be used for the initialization of the word embeddings. Table 4.1 shows the size of each sample.

sample	time bin	tokens
global	2013	900M
LiverpoolFC$_{13}$	2011–13	8.5M
LiverpoolFC$_{17}$	2017	11.9M

Table 4.1: The period and the size of each dataset.

4.3.2 Model

I adopt the model proposed by Kim et al. (2014). While other methods introduced in computational studies on diachronic meaning shift might be equally suitable, I expect the results not to be method-specific, because they concern general properties of short-term shift, as we will see in Sections 4.4 and 4.5. Similarly to the model used for the experiments in the previous chapter, also in this case, the model builds on the original Word2Vec architecture, proposing a relatively simple, but effective, methodology. Given a longitudinal corpus split into T consecutive time bins, word embeddings for t_0 are learned using the standard Word2Vec model. The learned vectors are then used to initialize vectors in t_1, that are then further updated. The vectors updated on t_1 are used to initialize those in t_2, and so on. The key idea, thus, is to substitute the randomly initialized embeddings assigned to each word in the dictionary with pre-computed embeddings, and then to updated them using the normal Word2Vec model. As a result, the pre-computed embeddings and the updated ones are in the same semantic space and, consequently, directly comparable.

Based on this methodology, the following steps are implemented. First, randomly initialized word embeddings are created using the large sample global. As a result, word representations that are community-independent are obtained. The rationale behind this first step is to create informative word representations by leveraging the very large amount of data in global. This would not be possible by directly creating word representations on the much smaller LiverpoolFC$_{13}$. The embeddings created on global are then used to initialize those in LiverpoolFC$_{13}$, which are updated on this sample. In this way, embeddings for time t_1 are obtained. In this second step, thus, the community-independent embeddings are adapted to the LiverpoolFC community. Finally, the word embeddings for LiverpoolFC$_{17}$ are initialized with those of t_1, and trained on this sample, resulting in the embeddings representations for words in t_2. It is thanks to this last step that it is possible to observe how much the representation of a word is updated from the t_1 to t_2, i.e., to detect meaning shift.

I consider the words in the vocabulary defined as the intersection of the vocabularies of the three samples (global, LiverpoolFC$_{13}$, LiverpoolFC$_{17}$), including 157k words. For global, only words that occur at least 20 times in the sample are considered, so as to ensure meaningful representations for each word. For the other two samples no frequency threshold is used: Since the embeddings used for the initialization of LiverpoolFC$_{13}$ encode community-independent meanings, if a word doesn't

occur in $\texttt{LiverpoolFC}_{13}$ its representation will simply be as in `global`, which reflects the idea that if a word is not used in a community, then its meaning is not altered within that community. The model is trained with standard skip-gram parameters (Levy et al., 2015): window 5, learning rate 0.01, embedding dimension 200, hierarchical softmax.

4.3.3 Evaluation Dataset

When the present study was conducted, no dataset annotated for short-term meaning shift was available.[1] For this reason, I created and made publicly available a dataset of words from the `LiverpoolFC` subreddit annotated for short-term semantic shift by members of the subreddit. The fact that the annotation was carried out by the members of the community is a crucial aspect: These were the same individuals involved in the creation and spread of new meanings, and, therefore, the ones that more easily could identify and annotate it.

In order to select the target words annotated by the community members, the content words that increase their relative frequency between t_1 and t_2 were initially identified.[2] Such a decision is based on the fact that frequency increase has been shown to positively correlate with meaning change (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015). Although an increase in frequency is not a necessary condition for meaning change, I consider it a reasonable starting point, as a random selection of words would contain very few positive examples. I focused on the words that show an increase in frequency that is larger than 2 standard deviations above the mean (approx. 200 words), and manually identified 34 semantic shift candidates among these words by analyzing their contexts of use in the `LiverpoolFC` data. Two types of confounders were than added: 33 words with a significant frequency increase, but that I did not mark as meaning shift candidates, and 33 words with constant frequency between t_1 and t_2, included as a sanity check. All words have absolute frequency in range [50–500].

In order to perform the annotation, the participants were shown the 100 words, in randomized order, together with randomly chosen contexts of usage from each time period (μ=4.7 contexts per word). For simplicity, the participants were asked to make a binary decision about whether there was a change in meaning. In order to have the participants familiarize with the concept of meaning change, they were initially provided with an intuitive, non-technical definition, and a set of cases that exemplify it. Overall, 26 members of `LiverpoolFC` participated in the survey, and each word received on average 8.8 judgments. The inter-annotator agreement, computed as Krippendorff's alpha, is α= 0.58, a relatively low value but common in semantic tasks (Artstein and Poesio, 2008). Three words were discarded from the initial list after analysis of the annotated data, two due to the homonymy with proper names not detected during the

[1] Subsequently, other datasets annotated for meaning shift over short periods of time were introduced, e.g., the one by Martinc et al. (2020). The corpus includes news about Brexit, it covers the period from 2011 to 2019, and it is split in 5 time bins.

[2] I identified the content words by using the external list of common words available at `https://www.wordfrequency.info/free.asp`.

survey's implementation, and one because the chosen examples clearly mislead the judgments of the redditors. The instructions and examples provided to participants can be found in Appendix A.

While annotators were asked to provide a binary judgment, semantic shift is arguably a graded notion. Therefore, in line with Kutuzov et al. (2018), the annotations were aggregated into a graded **semantic shift index**, defined as the percentage of annotations for a given word indicating a semantic change over the total number of annotations for that word. The index has values ranging from 0 (no shift) to 1 (shift). The shift index is exclusively based on the judgments of the community members, and does not consider the preliminary candidate selection done by me.[3]

4.4 Linguistic Phenomena in STMS

In this section, I address the first question introduced in Section 4.1, and analyze the linguistic phenomena underpinning short term meaning shift. The qualitative analysis introduced in this section is based on the manual inspection of the context of the words with a shift index > 0.5. Note that some of these words are the same that received a high community variation index (cvi) in the previous chapter, in particular, those that started to be used with a community-specific meaning at some point in time included between 2011 and 2017, such as, for example, 'believers'.

Several linguistic phenomena are at play in the data. While it is often hard to draw a clear-cut distinction between these phenomena, I was able to identify three main sources of shift: **metonymy**, **metaphor**, and **meme**.

4.4.1 Metonymy

In metonymic shifts, a highly salient characteristic of an entity is used to refer to it. It is possible to identify several cases of metonymic shift in the dataset. Among these cases is, for example, the word 'highlighter', that in t_2 occurs in sentences like *'we are playing with the highlighter today'*, or *'what's up with the hate for this kit? This is great, ten times better than the highlighter'*. In these examples, members of the community use 'highlighter' to talk about the away kit of the team, that has a color similar to that of a highlighter pen. Another example is the usage of the word 'lean', often occurring sentences like *'I hope a lean comes soon!'*, *'Somebody with speed... make a signing... Cuz I need a lean'*. In these sentences, 'lean' is used to talk about the possible hiring of new players. Such a usage is due to new hires typically *leaning* on a Liverpool symbol when posing for a photo right after signing for the club.

Particularly illustrative is the case of 'F5', whose contexts of use are shown in Table 4.2. While 'F5' is initially used with its common usage of shortcut for refreshing a page (1), it then starts to denote the act of refreshing a web page in order to get the

[3]The annotated dataset is available at: `https://github.com/marcodel13/Short-term-meaning-shift`.

	Example	Date
(1)	*Damn, after losing the F5 key on my keyboard [...]*	16 Jun
(2)	*[he is] so close, F5 tapping is so intense right now*	18 Jun
(3)	*Don't think about it too much, man. Just F5*	1 Jul
(4)	*Literally 4am I slept and just woke up and thought it was f5 time*	3 Jul
(5)	*this was a happy f5*	3 Jul
(6)	*what is an F5?*	3 Jul
(7)	*I'm leaving the f5 squad for now*	5 Jul
(8)	*I made this during the f5 madness*	6 Sep

Table 4.2: Examples of use of 'F5' with time stamps, that illustrate the speed of the meaning shift process. All the examples are from $\texttt{LiverpoolFC}_{17}$.

latest news about the possible transfer of a new player to Liverpool FC (2). This use catches on among the members of the community, and some of them use it to express their tension while waiting for good news (3-4) or their relief when the good news arrive (5). Interestingly, in example (3) the semantic change is accompanied by a change in the part of speech, with 'F5' becoming a denominal verb. However, many members are not aware of the new meaning of the word, and ask for clarification (6). This is in line with the finding presented in the previous chapter about the negative correlation with social dissemination of new linguistic variants in the first stages of diffusion. When the transfer is almost done, someone leaves the *'F5 squad'* (7), and after a while, another member recalls the period in which the word was used (8). This example gives a clear idea of what I mean when I talk about short-term meaning shift, as only few *weeks* passed between examples (1) and (8), and few *days* between examples (1) and (7).

4.4.2 Metaphor

Another relevant source of shift is the metaphorical usage of words, that leads to a broadening or narrowing of the original meaning of a word through analogy. An example of this kind of shift is the word 'snake', that members use in sentences like *'Hope millie gives the snake a f***g nightmare'* or *'Klavan smesh little snake'* to refer to a player of Liverpool FC who was playing for the team in 2013 and moved to a rival team in 2015. The supporters felt betrayed by the player, and started referring to him as the 'snake'. Another example is 'believers'. As said in the previous chapter, the meaning of the word underwent a shift process based on analogy, whereby the conventional meaning of believer as someone who believes in god came to denote those who believe in the team, i.e., the supporters.

While the examples above regard the metaphorical usage of single words, it is possible to observe also cases of *extended* metaphors (Werth, 1994), that is, cases in which the metaphor is conveyed by the whole sentence. Annotators spot these metaphors, and

label as cases of semantic change the words occurring in them. For example, *'shovel'* and *'coal'*, that in t_2 occur in sentences such as *'welcome aboard, here is your shovel'* or *'you boys know how to shovel coal'*. In this case the team is seen as a train that is running through the season, and every supporter is asked to figuratively contribute by shoving coal into the train boiler. A similar situation is observed for the word 'pharaoh', that in t_1 occurs in sentences like *'the pharaoh and the court magician'* and *'our dear own egyptian pharaoh, let's hope he becomes a god by the end'*. In this case, 'pharaoh' refers to an Egyptian player who joined the team, and that is compared by the supporters to a Pharaoh.

4.4.3 Meme

Finally, memes are another prominent source of meaning shift. While the concept of meme is a wide one, and refers to any kind of cultural object or idea that spreads from one person to another, here 'meme' is used to refer to fixed linguistic expressions used by the users of the community in several contexts, usually in an ironic way. As an example, while Liverpool FC was about to sign a new player named Van Dijk, community members started to play with the homography of the first part of the surname with the common noun 'van', its plural form 'vans', and the shoes brand 'Vans': '*Rumour has it Van Djik was wearing these vans in the van*' or '*How many vans could Van Dijk vear if Van Dijk could vear vans*'. Jokes of this kind are positively received in the community ('*Hahah I love it. Anything with vans is instant karma!*') and quickly become frequent in it. A similar usage is observed for the word 'dilly' (*'Dilly ding dilly dong, we're in the Champions League, man!'*), that spread very quickly after being used by a coach in an interview,[4] and 'darkness' that, based on the usage of the word in a famous song, is used by the users in contexts like *'Hello darkness Kevin Friend'* or *'Hello again Darkness I saw you just last week'* to express disappointment after the negative performance of a single player (*Kevin*) or of the whole team.

4.5 Automatic Detection of STMS

I now address the second question introduced in Section 4.1, and test the performance of the model for meaning shift detection introduced in Section 4.3.2 on the dataset annotated for short-term meaning shift.

Initially, vector representations for the words in the dictionary are created as described in Section 4.3.2. The the cosine distance between the representations in t_1 and t_2 for the words in the annotated dataset is then computed. Finally, the correlation between the shift index and the cosine distance is computed. A positive correlation is observed (Pearson's r= 0.49, $p < 0.001$, see Figure 4.1), showing that the model is able, to a certain extent, to capture short-term semantic shift. It is possible, however,

[4] See `https://www.skysports.com/football/news/11095/11553862/claudio-ranieris-best-quotes-from-sausageman-to-dilly-ding-dilly-dong`.

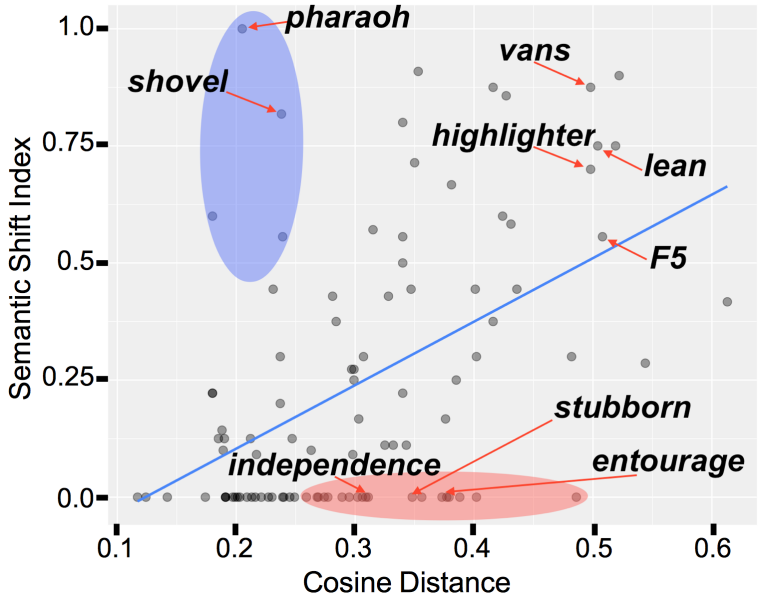


Figure 4.1: Semantic shift index vs. cosine distance in the evaluation dataset (Pearson's $r = 0.49$, $p < 0.001$). Red horizontal ellipsis: false positives; blue vertical ellipsis: false negatives.

to observe some systematic errors made by the model, that I analyze in the rest of this section.

4.5.1 False Negatives

A small, but consistent group is that of words that undergo semantic shift but are not captured by the model. In particular, this group includes words that have shift index>0.5 and cosine distance<0.25. In Figure 4.1, these are the words included in the blue vertical ellipsis. Words in this group are all cases of metaphorical shift. More precisely, they are cases of extended metaphor, in which the metaphor is developed throughout the whole text. The missed identification of the shift is due to the fact that the model used in the experiment, like all the other models based on Word2Vec (see Section 2.3.1), is only able to look at the local context of a word, that, in the cases of extended metaphor, does not change compared to the context of the literal uses. The model, hence, sees the *normal* linguistic context, and the representation of the word is not updated.

4.5.2 False Positives

A larger group of problematic cases is that of words that do *not* undergo a semantic shift but show relatively high cosine distance values between t_1 and t_2. This group includes words with shift index=0 and cosine distance>0.25, indicated by the red horizontal

ellipsis in Figure 4.1. Manual inspection reveals that most of these 'errors' are due to a **referential effect**, that is, the words are used almost exclusively to refer to a specific person or event in t_2, and so the context of use is different from the contexts in t_1. For instance, *'stubborn'* in t_1 occurs in different contexts, always with its conventional meaning, while in t_2, despite being used again with its conventional meaning, is almost always used to talk about a coach who was not there in 2013 but only in 2017. A similar situation is observed for the word *'entourage'*, used in t_2 to talk almost exclusively about the entourage of a specific player of the team during the days of the players market. Same for *'independence'*, that in t_2 occurs almost exclusively in association to 'Catalonia', to refer to the political events taking place in that region. In all these cases, the meaning of the word stays the same, despite the change in context. However, in line with the Distributional Hypothesis, the model spots the context change, and identifies such a change as a meaning shift. The problem, then, does not seem to depend on the model, but, rather, on the fact that not all the changes in context of use indicate a change in meaning. Such a problem is not reported in studies on meaning shift taking place over the long term. My intuition is that this is due to the fact that in the case of long-term shift embeddings are created from a much larger sample of language, including a more varied set of occurrences of the same word. This situation helps to eliminate, or at least mitigate, referential effects like the ones mentioned above, that, arguably, take place on relatively short periods of time.

4.5.3 Modeling Contextual Variability

The analysis presented in the previous section shows that a standard semantic model developed to sport shift over long periods of time is able to detect also short-term meaning shift, but it overgeneralize, that is, it identifies as semantic shift any change in context of use, while some of these changes are due to reference. Now, I introduce a measure that accounts for the difference between semantic shift and referential phenomena.

The measure is based on the observation that in referential cases the contexts of use is *narrower* than with actual semantic shift, since, as shown above, words are used to refer to a specific person or event. Hence, my hypothesis is that a measure of **contextual variability** should help spot false positives. To test this hypothesis, I first define contextual variability as follows. For a target word, a vector is created for each of its contexts in t_2 by averaging the embeddings of the words occurring in it, and variability is defined as the average pairwise cosine distance between context vectors. In this case, the context of the target word includes the 5 words on both its sides.[5] While much simpler, this method shares with current models of contextualized word representations such as EMLo (Peters et al., 2018) and BERT (Devlin et al., 2019) – which had not been developed at the time this research was carried out – the basic goal of capturing the information regarding the different contexts of use of a word.

[5]By experimenting with different window sizes I obtained similar results.

Contextual variability is computed for all the words in the dataset, and, also in this case, the correlation with the semantic shift index is computed, resulting in a positive correlation (Pearson's $r = 0.55$, $p < 0.001$). This result indicates that referential cases tend to occur in a restricted set of contexts, while semantic shifts are characterized by a wider set of new contexts. Figure 4.2 shows this effect visually. The words included in the red ellipsis are the ones included in the set of false positives in Figure 4.1. As it can be observed, all these words have low values of contextual variability, which confirms my hypothesis that referential cases tend to occur in a more restricted set of linguistic contexts.

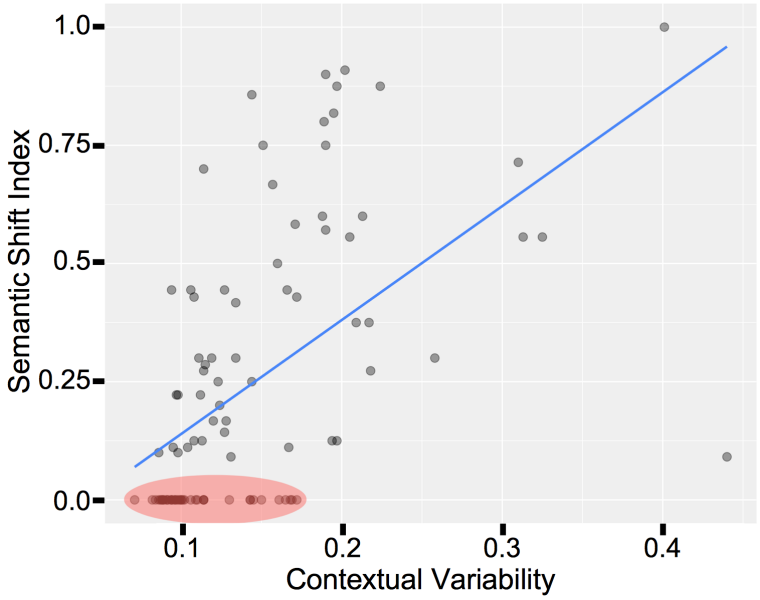


Figure 4.2: Semantic shift index vs. contextual variability. Red horizontal ellipsis: referential cases which are assigned high cosine distance values by the model (false positives).

4.6 Conclusion

This chapter focused on the diachronic aspects of meaning variation in online communities of speakers, making three main contributions: First, it provides a dataset annotated for short-term meaning shift, a kind of resource that, when introduced, was not available in the community. Second, it presents an analysis of the linguistic phenomena related to short term meaning shift. Third, it assesses the performance of a standard model for meaning shift detection on the newly introduced dataset, showing what are the main difficulties encountered by the model, and proposing a measure to remedy such difficulties. Besides the contributions listed above, the main merit of the current study was to bring to the attention of the NLP community short-term meaning shift,

a problem that, despite the large number of studies focusing on diachronic meaning shift, had not been addressed before. Subsequent works investigated questions raised in this study. For example, Boleda (2020) takes a more theoretical-oriented approach, and, building on the findings regarding the kinds of errors made by the model, reflects on the relation between context change and meaning shift. Another related study is Martinc et al. (2020), that, building on the recent advances in the field, use BERT to compute the word representations employed to detect short-term meaning shift. The study shows that, as expected, these new representations better encode information about short-term meaning shift. Martinc et al. (2020) build on Giulianelli et al. (2020), a work I contributed to, and that, while not focusing on short-term meaning shift, addresses and extends many of the problems and ideas originally proposed in this chapter, using contextualized embeddings.

Finally, a limitation of the present study is the size of the evaluation dataset, which is relatively small due to the great effort needed to collect and annotate the data. Recently, some works tried to overcome this limitation by proposing comprehensive evaluation frameworks for semantic change detection, that enable a systematic comparison of different models, and that consider both long and short term meaning shift (Schlechtweg et al., 2019; Shoemark et al., 2019).

Chapter 5

The Role of Community Members in the Introduction and Spread of Linguistic Innovations

The content of this chapter is based on the following publication:

Marco Del Tredici and Raquel Fernández. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

The two authors jointly produced the idea for the article. Marco performed the experiments with Raquel's supervision. Marco wrote the article, Raquel provided guidance and contributed to the writing. The text in this chapter partially overlaps with the one of the original publication.

5.1 Introduction

I continue, in this chapter, the investigation of the diachronic dimension of linguistic variation in online communities of speakers. While the previous chapter focused on the linguistic aspects of diachronic shift, here I focus on its **social** dimension, with the goal to uncover the role played by different types of users in the introduction and spread of new linguistic practices in online communities. In particular, I address two research questions, whose goal is to explore the social aspects of linguistic innovations at the core of RQ-2 defined in Section 1.2. First, traditional sociolinguistic theories account for the role played by different types of members in the spread of linguistic innovations in *offline* communities: Can these theories account for the behavior of users in *online* communities? Second, is it possible to leverage the information about the users who adopt a linguistic innovation, in order to predict whether such innovation will successfully spread in the community?

To address these questions, I build on the sociolinguistic theories introduced in Section 2.1.4. In particular, I adopt the framework introduced by Milroy and Milroy (1985), that maintains that **weak**-tie users are the ones who favor the **introduction** of innovations in clusters of strongly connected users, while the **strong** ties in these clusters favor the **spread** of these innovations. Other accounts of linguistic spread have been proposed. Among these, the one by Labov (1972a), that is based on the concept of **centrality**, and that, differently from the Milroy's, identifies highly central users as the source of innovation.

I assess the claims above in a longitudinal study including 20 online Reddit communities and around 10 million users, that I characterize based on the strength of their ties and on their centrality. I define a set of linguistic innovations introduced in the communities under scrutiny, and investigate the relation between the spreading trajectory of each innovation and the characteristics of the users who adopted them. While in the previous chapter I analyzed the emergence of new meanings, in this chapter I focus on new linguistic *forms*, which are easier to spot and track in online communities.

The results of my experiments show that the theoretical frameworks by the Milroy's and Labov provide a complementary account of innovators, as they are found to be central members of their community, connected to many other users with relatively low tie-strength. Also, as suggested by the Milroy's, my study shows that strong-tie users effectively contribute to the dissemination of a new term. These findings are very consistent across all the communities under investigation. In addition, I show that, by solely using information on users' tie strength as a predictor variable, it is possible to anticipate whether an innovation will successfully spread within a community.

5.2 Related Work

The study in this chapter is tightly related to the ones presented in Section 2.2.2, with which it shares the basic idea of modeling a community as a graph representing the in-

teractions among users, and analyzing language variation in relation to the characteristics of the nodes and connections in the graph. As we saw, the majority of the presented studies leverage the social graph to analyze the extralinguistic factors that drive diffusion, such as geographical (Eisenstein, 2015) and demographic variables (Eisenstein et al., 2014), and the seniority of the users who adopt them (Rotabi et al., 2017).

Other works take a different approach, and, similarly to what I do in the current study, characterize the nodes based on their position and relations in graph. Among these studies is Rotabi and Kleinberg (2016), who investigate the diffusion *trending topics* on Twitter, i.e., on the analysis of patterns of words that experience a frequency burst at a given point in time. Similarly to my investigation, they dedicate particular attention to the users who first adopt these topics. Perhaps the work that is most directly connected to the one in this chapter is Goel et al. (2016), who investigate the amount of interaction required to adopt an innovation, and analyze the types of social network connections that are more influential in the spread of linguistic innovations. While clear similarities exist between these works and mine, none of them address the same research questions defined above.

Several subsequent studies followed on the same research line addressed in this chapter. Sharma and Dodsworth (2020) take an approach that is very similar to the one adopt here, as the authors focus on the relation between traditional theories and processes taking place in online communities. In particular, they analyze the similarities and differences between the social graphs investigated in traditional studies, that are anchored in temporally specific and ideologically mediated cultural norms, and the online social networks. Other works focus on the linguistic properties of innovations. For example, Stewart and Eisenstein (2018) propose an interesting analysis of the constraints posed by the linguistic system in which innovations take place, finding that dissemination across many linguistic contexts is a crucial factor for spread. Also Ryskina et al. (2020) focus on the linguistic properties of innovations, introducing two indices, *semantic sparsity* and *frequency growth rates of semantic neighbors*, and showing that both are predictive of innovation emergence. Both these studies are complementary to the one proposed in this chapter, which focuses on the social aspects of spread.

5.3 Methodology

In this section, I describe the dataset used in the experiments (5.3.1), the methodology used to define the social network of a community and the social role of its members (5.3.2), and the procedure to identify linguistic innovations and characterize their diffusion (5.3.3).

5.3.1 Data

I use again data from Reddit, leveraging the great abundance of active communities (subreddits) hosted on the platform. In order to overcome the limitations of the experiments in Chapters 3 and 4 related to the size of the dataset, 20 different communities are considered. This large set of communities allows me to conduct a large-scale analysis, and to draw conclusions that generalize across communities. The communities taken into account show substantial variability in terms of subject, matter, and size. Table 5.1 offers an overview of the main statistics of each community. Despite their rich heterogeneity, however, they all have characteristics similar to those of the `LiverpoolFC` community, that, as shown in previous chapters, allow a proper investigation of the processes related to linguistic variation. In particular, all the communities present features that are indicative of highly active and interconnected communities, where the spreading process leading to innovation diffusion finds a favorable environment (Hamilton et al., 2017b; Guille et al., 2013), namely:

- a topic that reflects a strong external interest, such as sport teams, videogames or TV series;
- small-to-medium size, that in contrast to very large subreddits such as `news` (15M users) or `funny` (18M), are less dispersive and favor tighter connections;
- high density, i.e., a high ratio of existing connections over the number of potential connections.[1]

The entire content of each subreddit from its first post to the end of 2016 is downloaded, and the data from each subreddit segmented into consecutive time bins corresponding to one month.[2] Time bins with less than 200 active users – which are particularly common during the first few months of a subreddit lifespan – are discarded. Also, posts whose author is unknown are ignored.[3]

The next section explains how this longitudinal data is leveraged to extract information about the social role of community members, as well as to detect linguistic innovations and characterize their diffusion.

5.3.2 Social Networks

Creating the social graph I create an undirected and unweighted graph representing a community's social network for each time bin (i.e., month) during the community lifespan. Following the general approach described in Section 2.3.2, in the graphs nodes are users and edges encode the interactions among users. I follow Hamilton

[1] Hamilton et al. (2017b) report density values in the range $[0.001 - 0.016]$ in their set of subreddits.

[2] I also experimented with smaller bins of one week, obtaining similar results.

[3] When users delete their account, the posts, comments, and messages submitted prior to the deletion are still visible to others, but information about the user is not available (see Reddit Privacy Policy at `https://www.redditinc.com/policies/privacy-policy`).

subreddit	**years**	**tokens**	**users**	**density**	**innovations**
Android	7	158	1.03M	0.006	730
apple	8	89	580K	0.006	584
baseball	6	101	576K	0.014	520
beer	7	29	291K	0.008	360
boardgames	6	88	313K	0.004	380
cars	6	101	544K	0.014	605
FinalFantasy	4	22	137K	0.009	218
Guitar	7	71	387K	0.009	496
harrypotter	5	39	287K	0.005	227
hockey	7	191	847K	0.012	602
LiverpoolFC	5	40	173K	0.018	314
Patriots	5	26	151K	0.009	231
pcgaming	5	52	350K	0.003	360
photography	8	81	353K	0.006	485
pokemon	6	107	1.02M	0.006	695
poker	6	28	104K	0.012	258
reddevils	4	49	186K	0.008	329
running	6	56	279K	0.008	367
StarWars	6	56	542K	0.008	381
subaru	5	21	187K	0.005	340

Table 5.1: Statistics of the subreddits in our dataset, including: years of activity considered until end of 2016; total # of tokens (in millions); total # of active users (including users who may have left the community); average ratio of network ties over all possible ties computed over all the time bins (density); total # of linguistic innovations analyzed.

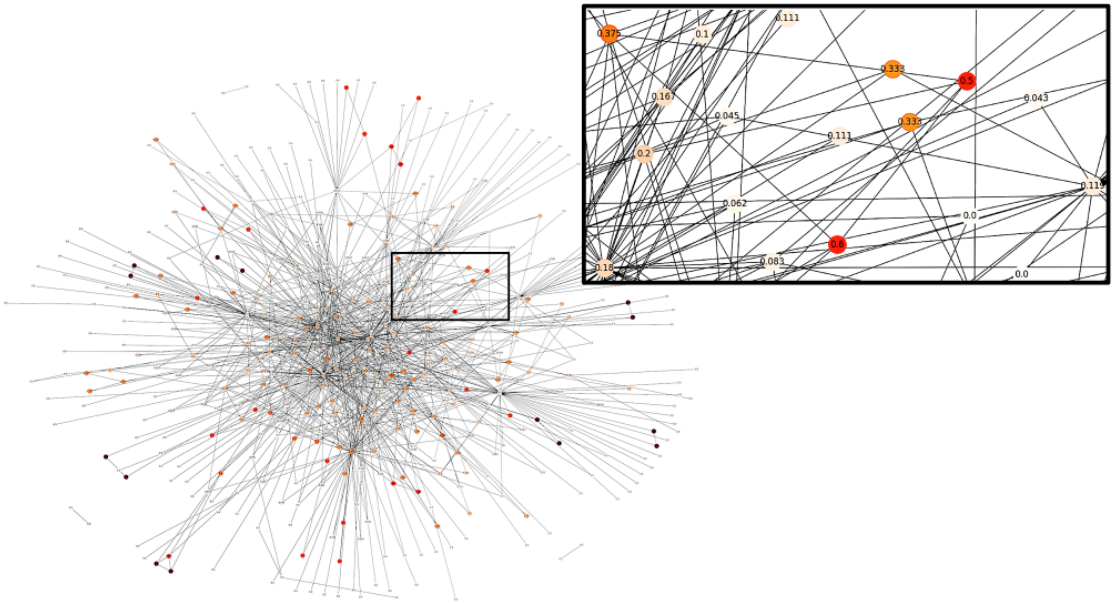


Figure 5.1: The social graph of the `beer` community in time bin t. The color of the nodes reflects the tie-strength of the users: The darker the color, the stronger the ties.

et al. (2017b) and instantiate an edge between two users if they comment within the same thread in close proximity, namely, if they are separated by at most two posts.

Computing tie-strength As detailed in Section 2.2.2, the cornerstone of the theoretical framework introduced in Milroy and Milroy (1985) is the concept of tie-strength. In line with previous studies (Zhao et al., 2010; Weng et al., 2015; Goel et al., 2016), I adopt the measure introduced by Onnela et al. (2007) to define the strength of the tie connecting two users in the social graph. This measure determines the strength of the tie between two individuals i and j based on **local** information, that is, on the overlap O_{ij} of their adjacent neighborhoods:

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \tag{5.1}$$

where n_{ij} is the number of adjacent nodes between i and j, and k_i, k_j their respective degree, i.e., the number of edges incident to each of them. Possible values for O_{ij} are in the range [0-1], where 0 indicates no common neighbors (weakest possible connection between i and j) and 1 exactly the same adjacent neighbors (strongest possible connection between i and j).

Characterizing users Equation 5.1 defines the strength of a tie between two users. I now explain how this strength value is leveraged to characterize users given their connections. According to Milroy and Milroy (1985), weak-tie individuals have *only* weak

connections, since they are not part of close-knit clusters. Conversely, strong-tie individuals have strong connections with other users, but may *also* have weak connections if they are linked to weak-tie users. To capture this, the tie strength of each individual i is defined as the **highest value** of their incident edges. That is, for all individuals j directly connected to i:

$$\textit{tie-strength}(i) := \max\left(O_{ij}\right) \tag{5.2}$$

By taking the maximum, I aim at capturing the idea expressed in the original theory by the Milroy's: A community member who has weak connections only will have low tie-strength and be considered a **weak-tie user**, while a member with either strong connections only or with both strong and weak connections will have high tie-strength and be considered a **strong-tie user**. Figure 5.1 shows an example of a social graph, where each user (node) is assigned a value based on the methodology just described.

Besides computing tie-strength as defined in Equation 5.2 for all users in the social graphs, I also compute their degree and betweenness **centrality** values. As detailed in Section 2.1.4, these are **global** indices of the importance of a node with respect to all other nodes in a graph. In particular, degree centrality is defined as the number of connections (i.e., incoming edges) of a node, while betweenness centrality is based on the calculation of the shortest paths existing between any two nodes in the graph: The higher the number of times a node is included in the shortest path between two other nodes, the higher its betweenness centrality.

5.3.3 Linguistic Innovations

Differently from the experiments in the previous chapters, the current experiment focuses on new linguistic *forms*, rather than new *meanings*. This choice is due to two main facts: First, tracking new forms is easier than new meaning, and thus allows me to concentrate on the sociological aspects of linguistic spread, that are the main focus of this study. Second, the annotation of new meanings on the large dataset used for this study would be too costly. I thus consider terms belonging to **Internet slang**, a general term commonly used to refer to a range of linguistic phenomena such as abbreviations (e.g., 'cu' for 'see you'), acronyms ('IIRC' for 'if I remember/recall correctly') and phonetic spellings ('dat' for 'that'). These forms are very abundant and continuously introduced in online communication, thus offering a very large set of datapoints for the current investigation.

I initially tried to automatically extract the list of target terms from the data, based on the statistical analysis of the occurrences of the terms. However, I encountered several difficulties related to the detection of typos, misspelling, nicknames, etc. I thus decided to leverage the terms included in the dictionary available at `NoSlang.com`,[4] a comprehensive record of Internet slang that is constantly updated. After removing

[4] `https://www.noslang.com/dictionary/`

terms including non-alphabetic characters, I obtain a list of approximately 6K terms.[5] Finally, for each subreddit, I only consider terms that:

- are used at least 10 times in the subreddit;
- are not present during the first 3 months of the community's existence;
- are introduced within the initial quarter of the community lifespan.

By adopting these criteria, I restrict my analysis to newly introduced Internet slang terms, i.e., that are not present from the very beginning of the community's activity, but are not introduced too late, so as to be able to observe their trajectory for a substantial period of time.

The number of innovations considered across all subreddits is 7962, while the number of unique innovations amounts to 1456. Most of the terms (around 76%) occur in more than one community, although no innovation is present in all the subreddits in the dataset. Around 24% of innovations tracked occur in just one community. Some of these are clearly topic-related, e.g., 'pkemon' in `pokemon`, while others are more general purpose abbreviations that, in principle, could appear in any community, such as 'txs' ('thanks', in `Android`) or 'omgz' ('oh my god/gosh', in `subaru`). Regarding frequency, 72% of terms occur at least 50 times on average.

The goal of the present experiment is to investigate the relation between the spread of a term, and the social features of the users who use the term. I operationalize this idea by implementing the vector representations introduced in the next two paragraphs.

Dissemination trajectory As highlighted in studies presented in Sections 2.1.2 and 2.1.3, innovations are continuously introduced in communities of speakers. However, once introduced, different innovations may have different fates: They can spread widely within the community, be used by just a small sub-group, or fail to make an impact and disappear altogether. In line with the experimental setup defined in Section 3.7, I define the spread of a term as its **dissemination**, which is computed as the proportion of community members who use it at a given moment in time. Recall that although dissemination is often correlated with frequency, in principle a term can have high relative frequency but low dissemination (and vice versa), and that it is only dissemination that gives a measure of the spread of a term within a community. Thus, for each innovation I calculate its dissemination in each time bin, and define the dissemination trajectory as the vector of its monthly dissemination values since the innovation was introduced.

[5]To assess whether ambiguity may be an issue for the set of target terms, I checked whether they also appear in a standard English dictionary, PyDictionary. For example, the slang term 'bra' for 'brother' also has the standard meaning of 'brasserie'. However, given that less than 2% of terms in the dataset could potentially be ambiguous, I decided to not treat them in any special way.

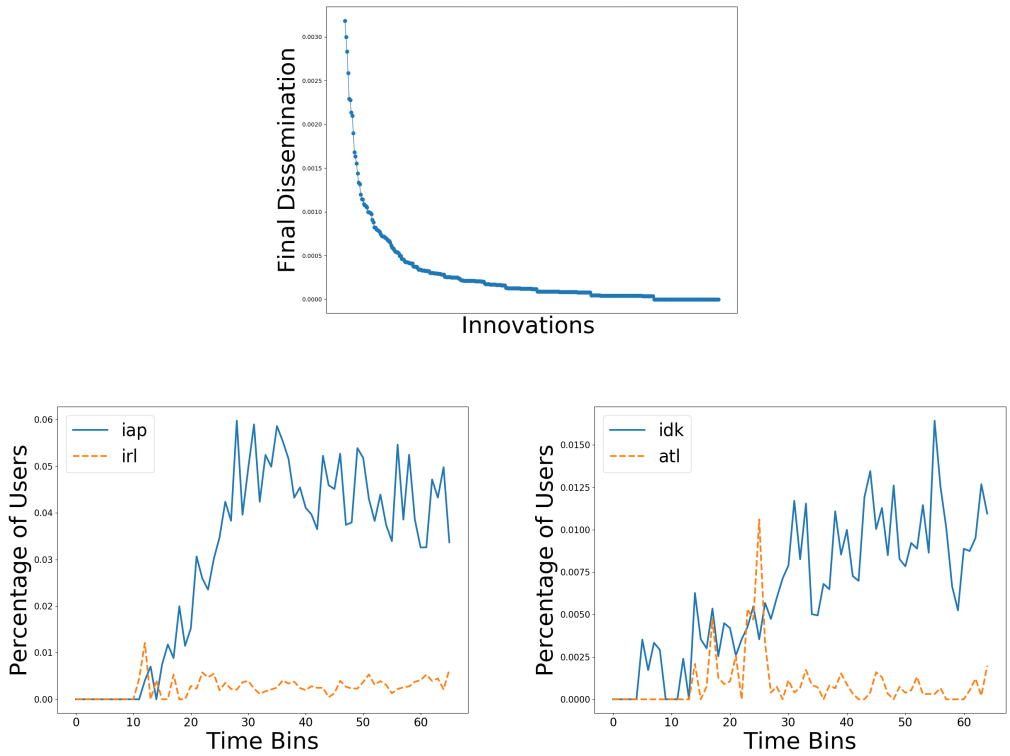


Figure 5.2: Top: Distribution of the final dissemination values in `beer`. Bottom: Examples of dissemination trajectories of successful (blue solid line) and unsuccessful (orange dashed line) innovations in `Android` (left) and `Patriots` (right).

Tie-strength trajectory Similarly, the tie-strength trajectory of each innovation is computed as a vector whose features correspond to the maximum tie-strength value among the users that used it in the corresponding month. Considering only the maximum value provides a simple way to test whether any individual with high tie-strength has used the term in a given month. Note that the dissemination and the tie-strength vectors always have the same magnitude, namely, the number of time bins considered for a community.

5.4 Empirical Observations

In this section, I present an analysis of the common patterns regarding the spread of linguistic innovations (5.4.1) and the structure of the social networks (5.4.2) in the set of communities under scrutiny.

5.4.1 Linguistic Innovations

I initially focus on the level of dissemination reached by the target innovations at the end of the period covered by my data. The final level of dissemination of a word is defined as the average dissemination value in the last six months.[6] The distribution of these final dissemination values is highly skewed for all the subreddits, as can be observed in Figure 5.2 (top) for the `beer` community. The same pattern is observed for all the communities: While a few innovations disseminate successfully, i.e., are adopted by a relatively high number of community members, most of them do not spread, and either disappear or barely appear in the last period. This finding is in line with the theoretical standpoint defined in Section 2.1.2, whereby only a minor part of the uncountable innovations introduced during human interactions manage to be durably adopted by communities of speakers.

In Figure 5.2 it is possible to observe examples of **successful** and **unsuccessful** innovations in two different communities, `Android` (bottom left) and `Patriots` (bottom right). Successful innovations such as 'iap' ('In App Purchases') and 'idk' ('I don't know') show a stable increase in dissemination after their introduction, which can reach a plateau at some point. Note that, in the case of 'iap', the spread resembles the S-shaped curve typical of the linguistic diffusion processes observed in Section 3.7, while for 'idk' the process is likely to still be ongoing at the end of the period covered by my analysis. Unsuccessful innovations, in contrast, can either have an almost flat dissemination trajectory, as in the case of 'irl' ('In Real Life'), indicating that the term has never experienced a spread in the community, or present a peak at some point, followed by a sudden decrease with no stable recovery, as for 'atl' ('Atlanta'). In this case, the term, after being adopted, is quickly abandoned by community members.

Based on these observations, I formally define the classes of **successful** and **unsuccessful** innovations based on the **dissemination slope** of a term, i.e., the difference between its average dissemination value in the first six months and in the last six months in the dissemination trajectory vector. The unsuccessful class includes innovations with slope index ≤ 0, i.e., those with trajectories similar to the 'irl' and 'atl' examples in Figure 5.2. In order to discard innovations with very low positive slope (i.e., those that do not disappear, but are only sporadically used) the successful class only includes terms whose slope index is above the average value of the community.[7] I will make use of these two classes in the prediction experiment in Section 5.6.

5.4.2 Social Networks

Next, I analyze the distribution of users' tie-strength values in the social graphs of the communities in the dataset. A clear pattern can be observed for all subreddits: The large majority of users have **low tie-strength**, with around 39% having values $\leq$ 0.05 and almost 50% having values ≤ 0.1; while, around 15 to 20% of users have

[6] By considering six months, instead of just the very last one, more robust measurements are obtained.

[7] Average slope index is positive for all subreddits.

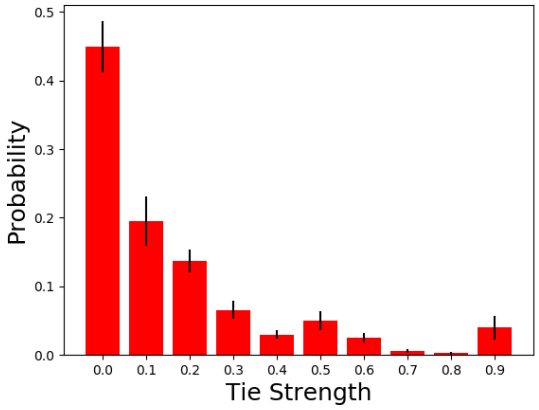


Figure 5.3: Average tie-strength distribution for all subreddits, with standard deviation.

strong tie-strength, with values ≥ 0.5. Figure 5.3 shows the average tie-strength value distribution computed over all the monthly graphs of all subreddits in the dataset, with probabilities calculated for bins of size 0.1 for illustration purposes.

This distribution is the same reported by Milroy (2002). Moreover, it mirrors the typical power-law distribution observed for centrality measures in online communities (Mihalcea and Radev, 2011). The topological properties captured by the tie-strength measure defined by Equation 5.2, however, are different from those captured by centrality, as already hinted at in Section 5.3.2. This is clearly shown by the fact that the two centrality measures considered (degree and betweenness) correlate strongly with each other (Spearman's r=0.89), while there is only a moderate correlation with tie-strength: r=0.63 with degree, and r=0.61 with betweenness.[8] These results confirm the difference between the adopted tie-strength measure and centrality: While centrality values are **global** indices of the role of a node with respect to the entire graph (Newman, 2010), tie-strength captures the **local** topological information around a node. In the online social communities under scrutiny, individuals at the core of the social network, who interact with many other individuals and have high posting activity, receive high centrality values. In contrast, high tie-strength values are the signature of users who belong to small cliques, but who do not act as hubs for the entire network.

Provided that the adopted measure is a suitable choice to represent the reference theoretical framework, it remains to be seen whether the social processes taking place in large online social communities and identified using such a measure lend support to the theory's main claims.

[8] All correlation coefficients reported are averages across time bins and subreddits and are all significant with $p < 0.05$.

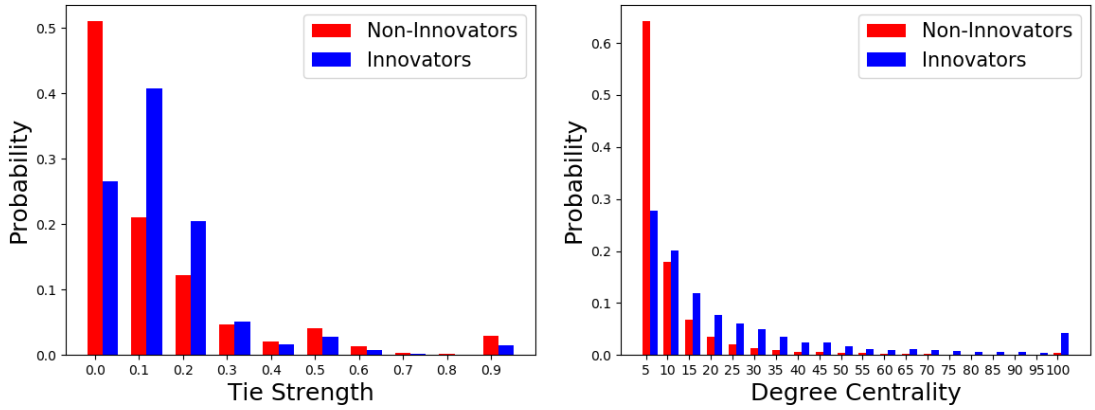


Figure 5.4: Comparison of the probability mass distribution of non-innovators' and innovators' tie-strength (left) and degree centrality (right).

5.5 Assessing Sociolinguistic Claims

In this section, I present the features that characterize innovators (5.5.1) and analyze the role of strong-tie users in the dissemination process (5.5.2).

5.5.1 Innovators

I assess here some of the claims put forward in the linguistic theories introduced in Section 2.1.4, namely, the claim in Milroy and Milroy (1985) stating that innovators are weak-ties individuals, and the claim by Labov (1972a) whereby they are highly central individuals. Following a straightforward intuition, I consider **innovators** the members who introduce a new term, that is, those who use it for the very first time in a community. There are 6.5K innovators across all communities, i.e., 0.9% of the users.

Initially, I verify whether innovators are **weak-tie** users. To this end, I compare the distribution of the tie-strength values of innovators and non-innovators across communities, and find a significative difference between the two groups (unpaired Welch's t-test $p < 0.0001$).[9] Figure 5.4 (left) shows how the tie-strength of the two groups differ. The plot shows that innovators are weak-tie users, but do not have the *weakest* possible ties (i.e., $p < 0.1$): rather, the tie-strength values of innovators cluster in the range [0.1–0.3]. This trend is observed consistently across all the communities. Also, innovators tend to *not* be strong-tie users, as shown by the fact that blue bars are lower than red ones for tie-strength ≥ 0.4.

I then focus on **centrality**, the key element for linguistic spread in Labov (1972a). Similarly to what I have done for tie-strength values, I compare the centrality val-

[9] As a sanity check, I randomly define the sets of innovators and non-innovators (keeping the same size of the original sets), and compute the statistical difference between the two sets. I repeat the process ten times, and in no run I observe a significant difference (average $p = 0.4$).

ues of innovators and non-innovators, for the two centrality measures considered, i.e., degree and betweenness. For both a significant difference is observed.[10] Figure 5.4 (right) shows the distribution of the values for degree centrality of innovators and non-innovators – bins of size 5 are chosen for illustration purposes. As it can be observed, innovators (blue bars) are *less* numerous among users with low degree centrality, i.e., with value in range [1-5] (first column), while as the values of degree centrality increase, the proportion of innovators is larger. This is particularly evident in the last column, which includes users with 100 or more connections.

Thus, a very robust pattern emerges across all subreddits, showing that innovators do have a particular profile in terms of their social standing: First, they have weak ties, which indicates that they do not belong to tightly connected cliques. Second, they occupy a central position in the network (high betweenness centrality), as hubs with many connections (high degree centrality). These results, hence, lend support to both the claims by the Milroy's and by Labov, since they confirm that innovations are spread by weak ties not belonging to close-knit clusters, who occupy a core position in the network.

5.5.2 Strong-Tie Users and Innovation Spread

I now turn the attention to the role of strong-tie users, who, in Milroy and Milroy (1985), are the individuals who spread the innovations after their introduction. I consider strong-tie users those with a tie-strength value ≥ 0.5. To analyze the role of strong-tie users in the innovation diffusion process, I proceed as follows. First, I identify any time t_i in the tie-strength trajectory vector when the innovation is used by some strong-tie member for k consecutive months. Then, I check its average dissemination in the period up to time t_i and compare it to the average dissemination in the six months following t_{i+k-1}.

The analysis shows that when $k = 1$ (i.e., when an innovation has been used by a strong-tie member only in one month) the probability that the dissemination increases in the next six months is around 50% for all subreddits — a value similar to the likelihood of dissemination increase after the usage by a weak-tie user. However, as k increases, and thus the adoption by strong-tie users becomes more stable, a future increase in dissemination becomes progressively more likely, for all the subreddits. Importantly, the same effect is not observed for weak-tie users, for whom, independently from the number of months, the probability of a future increase in dissemination is always approximately the same. Figure 5.5 shows examples of how the probability of dissemination changes after the adoption by either strong- or weak-tie users. A table with the results for all the communities can be found in Appendix B.

These results, thus, are consistent with the claims by Milroy and Milroy (1985), and show that innovation diffusion is connected to **sustained** adoption by strong-tie

[10] I repeat the sanity check described above also for the two centrality measures. Again, in no case a significant difference is observed between the samples. For degree centrality the average p over the ten runs is 0.44, for betweenness centrality the average p is 0.7.

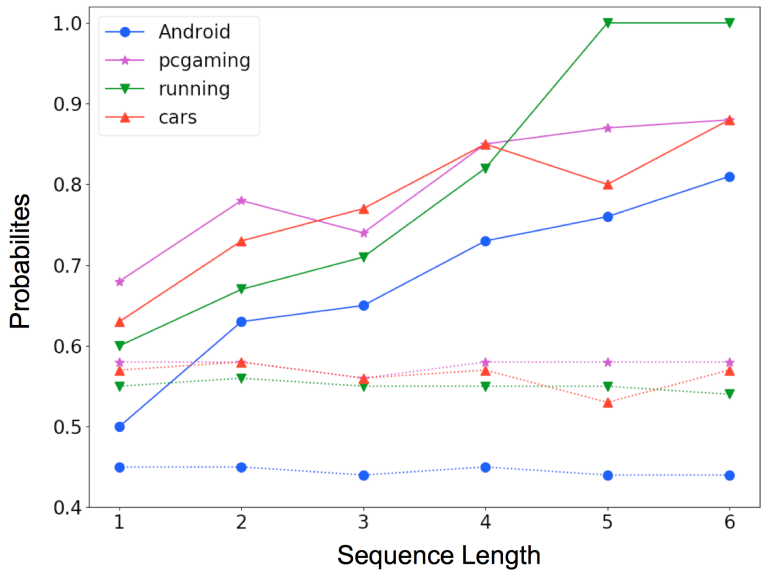


Figure 5.5: Probability of dissemination increase after a term is adopted by a strong-tie user (solid line) or by a weak-tie user (dotted line) for k consecutive months, computed for k in range $[1-6]$.

community members.

5.6 Predicting Innovation Success

Most innovations do not succeed in becoming community norms, but some do. Here I assess whether information about the tie-strength of members who use an innovation in the first months after its introduction can predict whether it will be successful in the future. This provides further theoretical insight into the importance of tie-strength for innovation diffusion, and has practical significance by contributing to identifying new terms that NLP systems should be able to process. My aim here is not to maximize prediction accuracy – which is likely to require taking into account several factors beyond users' tie-strength – but rather to explore whether the statistical effects regarding tie-strength presented in the previous sections are strong enough to have some predictive power.

Prediction is approached as a binary classification task, making use of the distinction between **successful** and **unsuccessful** innovations defined in Section 5.4. A subvector of length k is extracted from the tie-strength trajectory vector of innovations and used for the prediction. For instance, with $k = 3$, the tie-strength information from the first three months of usage of a term is used to predict if it will be successful or not. Subvectors of increasing magnitudes are used. For the final classification, a Random Forest classifier with default parameters is used.[11] The model is trained performing

[11] `http://scikit-learn.org/stable/modules/generated/sklearn.`

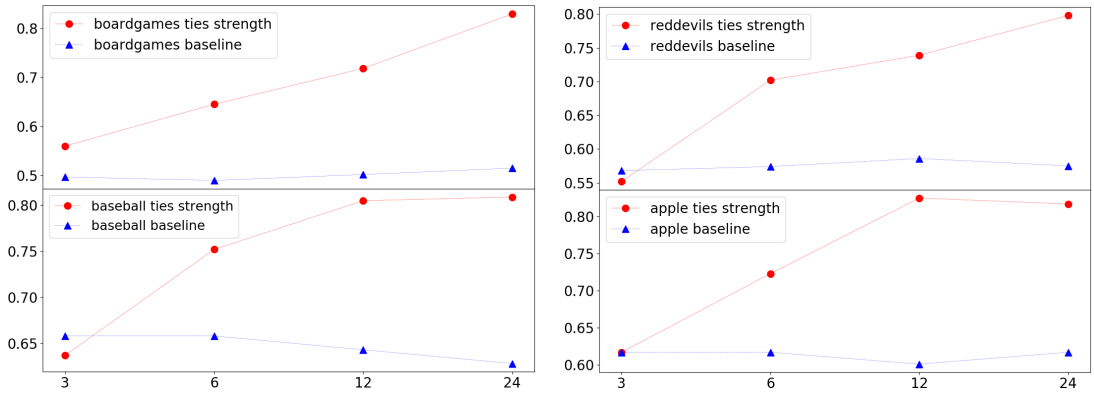


Figure 5.6: F-score (y-axis) for the successful class obtained with the tie-strength values of the first k months (x-axis) after the introduction of a term.

100-run cross-validation, in which 90% of the data is used for training and 10% for testing. The results of the model are compared against a weighted baseline, in which the two labels (successful/unsuccessful) are randomly assigned taking into account their frequency in the training set. The classes are fairly balanced across subreddits, with an average proportion of 55% successful and 45% unsuccessful.

When leveraging tie-strength information from only the first 3 months of usage, the F1 score of the model is significantly higher than the one of the baseline for 12 out the 20 subreddits. However, the overall performance remains rather low, with an average F1-score for the successful class of 0.62 vs. 0.58 for the baseline. Given that new terms are introduced by users with relatively low tie-strength (as shown in Section 5.5.1), arguably in the initial few months before a novel term is picked up by a strong-tie user, there is little difference between successful and unsuccessful innovations. With tie-strength information from the first 6 months of usage, the models is able to make predictions with results significantly above baseline for 18 out 20 subreddits, with an average F1-score of 0.68. Not surprisingly, performance increases substantially when information for a longer period (first/second year of usage) is exploited, reaching an average F1-score of 0.76, significantly above the baseline for all communities. Figure 5.6 graphically illustrates the results for a few communities, that are representative of the general trend observed. Detailed results, including precision, recall, and F1-score for each subreddit, are reported in Appendix B.

5.7 Conclusion

This chapter has provided a large-scale analysis of the interplay between the introduction and spread of new terms and users' social standing in large online social communities. Building on sociolinguistic theories – in particular, the version of *The Strength of Weak Ties* theory proposed by Milroy and Milroy (1985) – I proposed a simple

`ensemble.RandomForestClassifier.html`.

measure to quantify tie-strength of the users in an online social network in Milroy's sense, and I used it in combination with common centrality measures to uncover the characteristics of innovators and to assess the role of strong-tie users in the dissemination process. Regarding innovators, the results show that they are central community members, connected to many other users with relatively low tie-strength. These results provide support to both the theoretical frameworks by the Milroy's and by Labov introduced in Section 2.1.4. As for strong-tie users, the study indicates that in online social networks they are a small proportion of community members, organized in small cliques, and that the stable adoption of a linguistic innovation by this kind of user is related to the success of such an innovation. Also this finding is in line with the theoretical standpoint introduced by the Milroy's. Importantly, the reported patterns are highly consistent across the 20 online communities under scrutiny.

Finally, I showed that the information about tie-strength has some predictive power: By looking at the tie-strength of the users who adopt a new term in the first months after its introduction, it is possible to predict if the term will successfully spread in the community or not. While this experiment has limited scope, and it is performed using a basic classifier, it exemplifies the kind of predictive approach that characterizes the experiments I will present in the next chapters, and, thus, it ideally marks the transition to the second part of this dissertation.

Part Two

Modeling User Information for Text Classification

In Part One I described some relevant linguistic and societal processes underpinning linguistic variation in online communities. In the second part of this dissertation, I focus on how to represent individuals in these communities, with the goal to better understand the linguistic practices that they adopt and, more in general, to characterize them. Concretely, I investigate how to compute user representations that effectively capture **homophily** relations among users, and how to exploit these representations to improve the performance of different NLP models concerned with text classification tasks. Homophily relations are captured by considering two kinds of **user information**. In Chapter 6, I consider the **social connections** of the users, in order to perform user-generated text classification. In particular, building on the idea that a user is part of several communities at the same time, I implement a model that is able to identify which connections are relevant in a specific communicative situation, and to create user representations accordingly. In Chapter 7, I focus on the **linguistic production** of the users, with the goal to perform fake news detection. Based on the relation existing between social traits and language use, and between such traits and the propensity to spread fake news, I introduce a model that creates user representations based uniquely on the language they produce, and leverages them for the task at hand.

Chapter 6

Dynamic Representations for Social Media Users in NLP

The content of this chapter is based on the following publication:

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marco and Raquel jointly produced the idea for the article. Marco performed the experiments: Sabine provided supervision in the first stages of the experiments, Diego and Raquel in the following ones. Marco wrote the article, Raquel provided guidance and contributed to the writing, Sabine and Diego provided feedback. While not among the authors, I would like to thank Mario Giulianelli and Jeremy Barnes for their (very) valuable contribution in the earlier stages of this study. The text in this chapter partially overlaps with the one of the original publication.

6.1 Introduction

The idea that user information can help language understanding has gained a lot of traction in NLP in the last years, especially after the large diffusion of data derived from online communications (see Section 2.2.1). As detailed in Section 2.2.3, such an idea is usually operationalized by enabling a model for text classification to encode, together with the representation of the text, a representation of the user who produced that text. A common approach to model user information relies on the **social graph** in which users are embedded. In this approach, a representation is created, for each user, that encodes information about the connections of the user in the social graph. Such a representation, hence, aims at capturing the concept of **homophily** introduced in Section 2.1.3, whereby people create connections with those they share ideas, beliefs, and linguistic practices with. In this chapter, I also focus on creating user representations based on the connections in the social graph. In particular, I build on one of the main insights introduced by Eckert and McConnell-Ginet (1992) when describing communities of practice, namely, that individuals usually belong to **many communities**, and that they adopt, each time, the practices proper to each of these communities (see Section 2.1.3). Different community memberships, then, have different relevance depending on the situation. For example, consider one of the members of the `Liverpool` community introduced in Part One, and imagine that this individual is also part of a community of supporters of the UK Labour Party. While the membership to these two communities is equally important to characterize the person in general terms, the former is much more relevant when it comes to understanding the meaning of a tweet written to comment on the performance of the supported team, while the latter is important for understanding a tweet posted by this person regarding Boris Johnson. Here, I focus on this idea of **multiple membership** to address RQ-3 introduced in Section 1.2: How to identify the relevant information coming from user connections in the social graph and leverage it to improve text classification?

To answer this question, I propose a model that creates **dynamic user representations**. The model dynamically explores the social relations of an individual, learns which of these relations are more relevant for the task at hand, and computes the vector representation of the target individual accordingly. This is then combined with linguistic information to perform text classification. The model is tested on three different classification tasks involving **user-generated** texts, and its performance compared against models that create static vector representations, i.e., that uniformly aggregate information from the connections in the social graph. The results show that, when social information is relevant for the task, my model significantly outperforms competing alternatives. I also provide an extended error analysis, which shows why dynamic representations better encode homophily relations compared to static ones.

6.2 Related Work

As mentioned above, several works introduced in the last years propose to leverage social connections to compute user representations. Among these works are, for example, Yang et al. (2016), who leverage this kind of representation for the task of entity extraction; Mishra et al. (2018) and Mishra et al. (2019a), two studies I contributed to, that use graph-based representations for the task of abusive language detection. All these studies implement the methodology described in Section 2.3.2, which relies on creating a social graph where users are nodes and the edges among them are defined based on their interactions on the social platform. Computational techniques such as Node2Vec and Graph Convolutional Networks are then applied to the graph to learn low-dimensional embeddings for each user, that are finally used to perform the classification of the target text.

While reporting positive results, and thus showing the importance of social information, all the models present a common shortcoming, namely, they create user representations by **uniformly** aggregating the information coming from their connections in the social graph. This approach does not consider the crucial fact that individuals typically belong to **several communities**, and that when they interact with different communities, they adopt different practices. Thus, in order to create representations that better represent homophily relations, the information from the neighbors should be **weighted**, depending on the specific communicative situation. The study in this chapter implements this idea, leveraging the most recent neural architectures for graph encoding to model a crucial aspect of sociolinguistic theories.

6.3 Model

This section introduces the model that I designed to create dynamic user representations, and to leverage them to classify user-generated texts. The model operates on annotated corpora made up of triples (t, a, y), where t is some user-generated text, a is its author, and y is a label classifying t. The task is to predict y given (t, a). Note that label y is assigned to t only, and no label exists for a. Since my goal is to observe how model performance varies depending on user representations, the model can operate with different kinds of representations, as we will see in Section 6.4.1.

The model consists of two modules, one encoding the **linguistic information** in t and the other one modeling **user information** related to a. The general architecture is shown in Figure 6.1. The output of the linguistic and social modules are vectors $l \in \mathbb{R}^d$ and $s \in \mathbb{R}^{d'}$, respectively. I adopt a standard late fusion approach in which these two vectors are concatenated and passed through a two-layer classifier, consisting of a layer $W_1 \in \mathbb{R}^{d+d' \times c}$, where c is a model parameter, and a layer $W_2 \in \mathbb{R}^{c \times o}$, where o is the number of output classes. The final prediction is computed as follows, where σ is a ReLU function (Nair and Hinton, 2010):

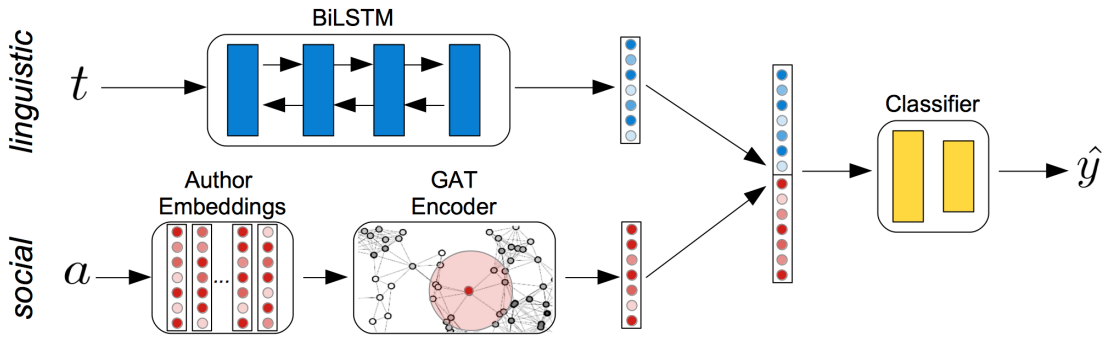


Figure 6.1: General model architecture. The linguistic module returns a vector representation of the input tweet t. The social module takes as input the pre-computed representation of the author a and updates it using the GAT encoder. The output embeddings of the two modules are concatenated and fed into a classifier.

$$\hat{y} = softmax(W_2\,(\sigma\,(W_1\,(l\|s)))) \tag{6.1}$$

The **linguistic module** is implemented using an LSTM, that, at the moment when the current experiments were performed, was the standard choice to represent sentences in NLP. In particular, a bidirectional LSTM (BiLSTM) (Graves, 2012) is used, and its final states are concatenated in order to obtain the representation of the input text.

The goal of the **social module** is to return author representations that encode homophily relations among users, i.e., that assign similar vectors to users who are socially related. To create these representations, the module leverages the social graph $G = (V, E)$, where V is the set of nodes representing individuals and E the set of edges representing relations among them. The module takes as input a pre-computed user vector, performs a dynamic exploration of its neighbors in the social graph, and updates the user representation given the relevance of its connections for a target task. Node2Vec (N2V) is used to pre-compute initial user representations. Recall from Section 2.3.2 that in N2V neighbors are randomly selected, and hence the model makes no distinction among them. Relevant connections are then identified using Graph Attention Network (GAT), which leverages a self-attention mechanism to assign **different relevance to different neighboring nodes** depending on the task. The user representation returned by the model is finally concatenated with the output of the linguistic module and fed into the classifier.

6.4 Experimental Setup

This section describes the details of the experimental setups in which the performance of the model introduced above is assessed. I initially introduce the alternative models

against which the performance of my model is compared (6.4.1), including hyperparameter search (6.4.2). I then introduce the target tasks and the related datasets (6.4.3), and the metrics optimized for each task (6.4.4). Finally, I present the main features of the social graphs used for the experiments (6.4.5).

6.4.1 Alternative Models

The performance of the model introduced above is compared against several competing models.

Frequency Baseline In this baseline, labels are randomly sampled according to their frequency in the dataset.

LING Only the linguistic module (LING) is used: In this way, is it possible to assess the performance of the model when no social information is provided.

LING+random A setting similar to Kolchinski and Potts (2018) is implemented. In this setting, each user is assigned a random embedding, that is updated during training. The implementation in the current study differs from the Kolchinski and Potts's model in two aspects: They use GRUs (Cho et al., 2014) rather than LSTMs for the linguistic module, and their author vectors have size 15, while the vectors used in my experiments have size 200 to allow a fair comparison with the other models (see below).

LING+PV As explained in Section 2.3.1, and as we will see in the next chapter, a very relevant source of information to model users in online setups is the language they produce. For this reason, a representation for each author is computed by running Paragraph Vector (PV) on the concatenation of all their previous tweets. The model, thus, makes no use of the social graph, and author representations are based uniquely on past linguistic usage.

LING+N2V While none of the previous models make use of the social graph, the current one represents authors by means of the embeddings created with N2V. In contrast to my GAT-based model, N2V computes user embeddings without making any distinction among neighbors, and without updating them with respect to the task at hand.[1]

6.4.2 Hyperparameter Search

For all the models and for each dataset, grid hyperparameter search is performed on the validation set using early stopping. For batch size, I explore values 4, 8, 16, 32,

[1] I also experimented with updating author embeddings during training, but did not observe any difference in the results.

64; for dropout, values 0.0, 0.1, 0.2, 0.3, ..., 0.9; and for L2 regularization, values 0, 1^{e-05}, 1^{e-04}. For all the settings, I use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001, β_1 = 0.9 and β_2 = 0.999.

I run PV with these hyperparameters: 30 epochs, minimum count of 5, vector size of 200. For N2V I use the default hyperparameters, except for vector size (200) and epochs (20). For the GAT encoder, I experiment with values 10, 15, 20, 25, 30, 50 for the size of the hidden layer; for the number of heads, I explore values 1, 2, 3, 4. The number of hops is kept equal to 1 and the alpha value for the Leaky ReLU of the attention heads equal to 0.2 across all the settings.[2]

Since my focus is on social information, I keep the hyperparameters of the linguistic module and the classifier fixed across all the settings. Namely, the BiLSTM has depth of 1, the hidden layer has 50 units, and uses 200-d GloVe embeddings pretrained on Twitter (Pennington et al., 2014). For the classifier, I set the dimensionality of the non-linear layer to 50.

6.4.3 Tasks and Datasets

All the models are tested on three Twitter datasets annotated for different tasks. For all datasets, the text is tokenized and lowercased, and any URL, hashtag, and mention is replaced with a placeholder.

Sentiment Analysis Given a tweet, the task is to classify its polarity (i.e., if it is positive, negative of neutral). I use the dataset in Task-4 of SemEval-2017 (Rosenthal et al., 2017), that includes 62k tweets labeled as `POSITIVE` (35.6% of labels), `NEGATIVE` (18.8%) and `NEUTRAL` (45.6%). Tweets in the train set were collected between 2013 and 2015, while those in the test set in 2017. Due to the volatility issue mentioned in Section 2.2.1, information for old tweets is difficult to recover. To have a more balanced distribution, hence, the dataset is shuffled and then split it into train (80%), validation (10%) and test (10%).

Stance Detection For this task, given a tweet and a topic, the goal is to determine whether the tweet expresses a stance that is in favor or against of the given topic, or whether neither inference is likely. Thus, while stance detection is related to sentiment analysis, a crucial difference is that in the latter the focus is only on the polarity of the text, while the former is more complex, as it aims at capturing the opinion expressed by the text toward a specific entity. I use the dataset released for Task-6 (Subtask A) of SemEval-2016 (Mohammad et al., 2016), that includes 4k tweets labeled as `FAVOR` (25.5% of labels), `AGAINST` (50.6%) and `NEUTRAL` (23.9%), with respect to five

[2]I implement PV using the Gensim library: `https://radimrehurek.com/gensim/models/doc2vec.html`. For N2V, I use the implementation at: `https://github.com/aditya-grover/node2vec`. For GAT, the implementation at: `https://github.com/Diego999/pyGAT`.

topics: 'Atheism', 'Climate change is a real concern', 'Feminist movement', 'Hillary Clinton', 'Legalization of abortion'. The dataset is split into train and test. 10% of tweets in the train split are randomly extracted and used for validation.

Hate Speech Detection This is a binary a task, in which the model has to classify a text as hateful, when it denigrates a person or a group based on social features (e.g., ethnicity), or normal, i.e., non-hateful. I employ the dataset introduced by Founta et al. (2018), from which I keep only tweets labeled as `NORMAL` (93.4% of labels) and `HATEFUL` (6.6%) for a total of 44k tweets.[3] The dataset is randomly split into train (80%), validation (10%) and test (10%).

6.4.4 Optimization Metrics

Models are tuned using different evaluation measures, according to the task at hand. The rationale behind using different metrics is to use, whenever possible, established metrics per task.

For Sentiment Analysis average recall is used. This is the same measure used for Task 4 of SemEval-2017 (Rosenthal et al., 2017), computed as:

$$AvgRec = \frac{1}{3}(R^P + R^N + R^U) \tag{6.2}$$

Where R^P, R^N and R^U refer to recall of the `POSITIVE`, the `NEGATIVE`, and the `NEUTRAL` class, respectively. The measure has been shown to have several desirable properties, among which robustness to class imbalance (Sebastiani, 2015).

For Stance Detection, the average of the F-score of `FAVOR` and `AGAINST` classes is used:

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \tag{6.3}$$

The measure, used for Task-6 (Subtask A) of SemEval-2016, is designed to optimize the performance of the model in cases when an opinion toward the target entity is expressed, while it ignores the neutral class (Mohammad et al., 2016).

Finally, for Hate Speech Detection, a more recent task, it was not possible to identify an established metric. For this reason, I use the F-score for the target class `HATEFUL`, the minority class accounting for 6.6% of the datapoints.

6.4.5 Social Graph Construction

In order to create the social graph, I follow the approach described in Section 2.3.2. I initially retrieve, for each tweet, the ID of its author using the Twitter API and scrape

[3]The other labels in the dataset are `SPAM` and `ABUSIVE`.

	Sentiment	Stance	Hate
# tweets	62,530	4,063	44,141
% with author	71.4%	71.7%	77.1%
# nodes	50k	6.9k	25k
# edges	4.1m	258k	1.3m
density	0.003	0.010	0.004
# components	1	1	1
homophily	38%	60%	68%

Table 6.1: Statistics for each dataset: number of tweets; percentage of tweets for which it was possible to retrieve information about the author; number of nodes; number of edges; density; number of connected components; and amount of homophily as percentage of connected authors whose tweets share the same label.

their timeline, i.e. their past tweets.[4] An independent social graph $G = (V, E)$ is then created for each dataset. V is the set of users authoring the tweets in the dataset, while for E an unweighted and undirected edge is instantiated between two users if one retweets the other. Information about retweets is available in users' timeline. In order to make the graph more densely connected, V includes external users not present in the dataset who have been retweeted by authors in the dataset at least 100 times. When information about the author of a tweet is not available, they are assigned an embedding computed as the centroid of the existing author representations. In the datasets used for this study, authors with more than one tweet are rare (6.6% on average).

Table 6.1 summarizes the main statistics of the datasets and their respective graphs. The three social graphs have different number of nodes: The network of the Sentiment Analysis dataset is the largest (∼62k nodes) while the Stance network is the smallest (∼4k nodes). The number of edges and the density of the network (i.e., the ratio of existing connections over the number of potential connections) vary according to graph size, while the number of connected components is 1 for all the graphs: This means that there are no disconnected sub-graphs in the social networks.

The most relevant aspect for which differences can be observed across the three graphs is the amount of **homophily**, that I define as the percentage of edges that connect users whose tweets have the same label. This value is similar for the Stance and Hate Speech social graphs, and much higher in these graphs than in the Sentiment Analysis one, in which such a value is similar to a random distribution.[5] This indicates that, in the datasets, users expressing similar opinions about a topic (Stance) or using offensive language (Hate Speech) are more connected than those expressing the same sentiment in their tweets (Sentiment).

[4] I access the API using the Python package Tweepy: `http://docs.tweepy.org/en/v3.5.0/`. The API returns a maximum of 3.2k tweets per user.

[5] Given that there are three possible labels for the task, by randomly assigning labels to the neighbors of a node, the level of homophily would be 33.3% on average.

Model	Sentiment	Stance	Hate
Frequency	0.332	0.397	0.057
LING	0.676	0.569	0.624
LING+random	0.657	0.571	0.600
LING+PV	0.671	0.601*	0.667*
LING+N2V	0.672	0.629$^{*\diamond}$	0.656*
LING+GAT	0.666	0.640$^{*\diamond\dagger}$	0.674$^{*\diamond\dagger}$

Table 6.2: Results for all the models on the three datasets. Marked with * are the results that significantly improve over LING and LING+random ($p < 0.05$, also for the following results); $\diamond$ indicates a significant improvement over LING+PV; † a significant improvement over LING+N2V.

6.5 Results

The performance of the models is evaluated using the same metrics used for the optimization process described in Section 6.4.4. Table 6.2 reports the results, computed as the average of ten runs with random parameter initialization. The unpaired Welch's t test is used to check for statistically significant differences between models. The standard deviation for all models is in the range [0.003-0.02]. The full table including the results for each dataset and for each class is in Appendix C.

Tasks The results show that **user information** helps improve the performance on Stance and Hate Speech detection, while it has no effect on Sentiment Analysis. This result contrasts with the one reported by Yang and Eisenstein (2017), who use a previous version of the Sentiment dataset (Rosenthal et al., 2015). However, such a result is not surprising, given the analysis made in the previous section regarding the amount of **homophily** in the three social graphs, which shows that, in my version of the data, sentiment is not as related to the social standing of individuals as stance and hatefulness are. My intuition, also, is that this is not a specific feature of the used dataset, but of the task in general, as the polarity of a text is a very general feature, that is arguably loosely connected to the social standing of the individual who produced it.

Models I now turn to the analysis of the performance of each model. LING+random never improves over LING: I believe this is due to the fact that most of the authors in the dataset used for this study have just one tweet, which hinders the possibility to learn at training time the representations of the users used at test time. Differently, both PV and N2V user representations lead to an improvement over LING. N2V vectors are especially effective for the Stance detection task, where LING+N2V outperforms LING+PV, while for Hate Speech the performance of the two models is comparable (the difference between LING+PV and LING+N2V is not statistically significant due to the high variance of the LING+PV results). Finally, the GAT-based model out-

(1) *@user: Yurtle the Turtle needs to be slapped with a f***ing chair..many times!*	HATEFUL
(2) *You stay the same through the ages... Your love never changes... Your love never fails*	AGAINST atheism
(3) *Why are Tumblr feminists so territorial? Pro-lifers can't voice their opinions without being attacked*	AGAINST abortion
(4) *@user No, just pointed out how idiotic your statement was*	HATEFUL

Table 6.3: Examples from Stance (2, 3) and Hate Speech (1, 4) datasets, and their label.

performs any other model on both Stance and Hate Speech detection. These results confirm my initial hypothesis that a social attention mechanism that is able to assign different relevance to different neighbors allows for a more dynamic encoding of homophily relations in author embeddings and, in turn, leads to better results on the prediction tasks.[6]

6.6 Analysis

I now analyze in more detail the strengths and weaknesses of the models leveraging user representations, for the tasks where such information proved useful.

6.6.1 Paragraph Vector

Figure 6.2 (left) shows the user representations created with PV for the Hate Speech dataset.[7] The plot shows that users form sub-communities, with authors of hateful tweets (orange dots) mainly clustering at the top of the plot. The similarity between these individuals derives from their consistent use of strongly offensive words towards others over their posting history. This suggests that representing speakers in terms of their past linguistic usage can capture certain factors characterizing the users, in particular, those sociological and psychological factors related to general ideas and beliefs that, as I will show in the next chapter, are mirrored by linguistic production. For example, tweet (1) in Table 6.3 is incorrectly labeled as NORMAL by the LING model (note that *'f***ing'* is often used with positive emphasis and is thus not a reliable clue for hate speech). By leveraging the PV author representation (which, given previous

[6] In preliminary experiments, Graph Convolutional Networks with no attention showed no improvements over the N2V baseline. The result is to be expected since, similarly to N2V, GCN computes the representation of the target node without making any distinction among its neighbors (see Section 2.3.2).

[7] Plots are created using the Tensorflow Projector available at: https://projector.tensorflow.org.

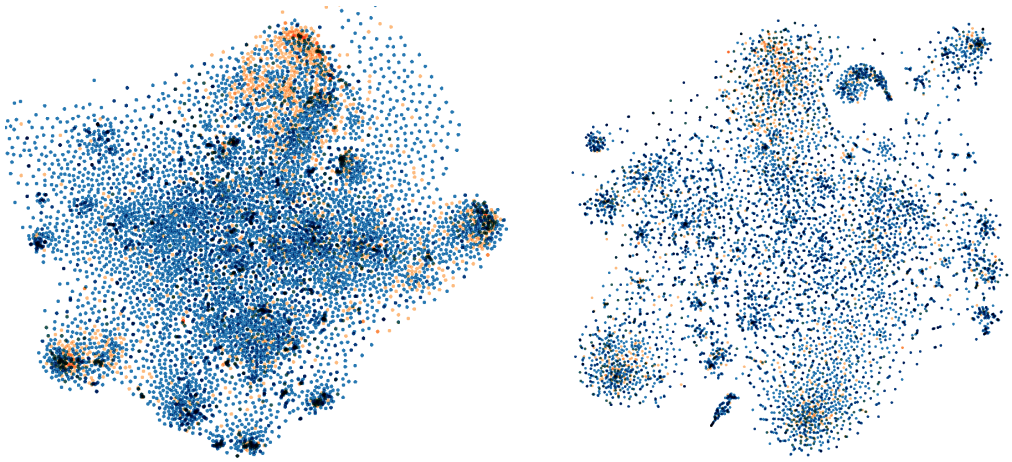


Figure 6.2: PV (left) and N2V (right) user representations for the Hate Speech dataset. In both plots, orange dots are authors of HATEFUL tweets, blue dots of NORMAL tweets.

posting behavior, is highly similar to authors of hateful tweets) the LING+PV model yields the right prediction in this case.

For Stance detection, which arguably is a less lexicaly determined task (Mohammad et al., 2016), PV user representations are less effective. This is illustrated in Figure 6.3 (left), where no clear clusters are visible. Still, PV vectors capture some meaningful relations, such as a small, close-knit cluster of users against atheism (see zoom in the figure), who tweet mostly about Islam.

6.6.2 Node2Vec

User representations created by exploiting the social network of individuals are more robust across datasets. For Hate Speech, the user representations computed with N2V are very similar to those computed with PV – see Figure 6.2 (right). However, for the Stance dataset, N2V user representations are more informative. This is readily apparent when comparing the plots in Figure 6.3: Users who were scattered when represented with PV now form community-related clusters, which leads to better predictions. For example, tweet (2) in Table 6.3 is authored by a user who is socially connected to other users who tweet against atheism (the orange cluster in the right-hand side plot of Figure 6.3). The LING+N2V model is able to exploit this information and make the right prediction, while the tweet is incorrectly classified by LING and LING+PV, which do not take into account the author's social standing.

N2V, however, is not effective for users connected to **multiple communities**, because, as explained in Section 2.3.2, the model will conflate this information into a fixed vector located between clusters in the social space. For instance, the author of tweet (1) is connected to both users who post hateful tweets and users whose posts are not hateful. In the N2V user space, the ten closest neighbors of this author are

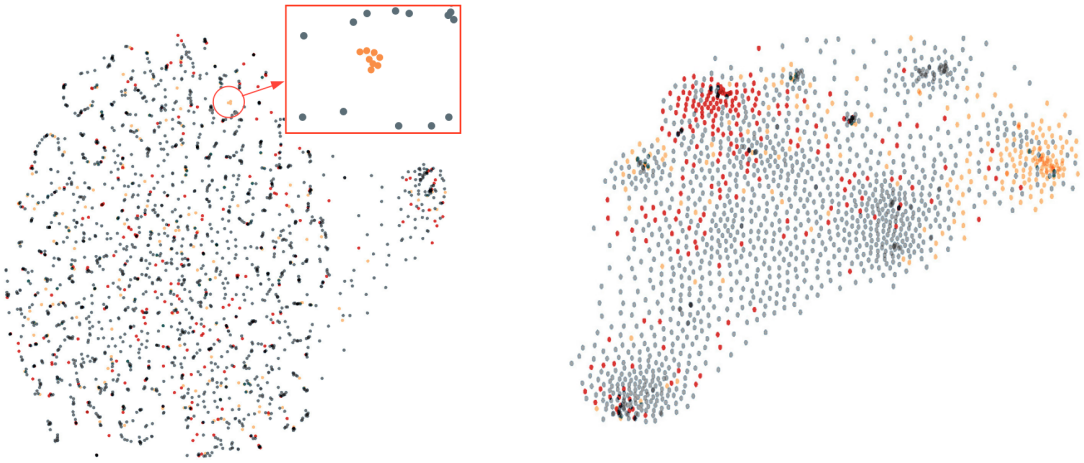


Figure 6.3: PV (left) and N2V (right) user representations for the Stance dataset. In both plots, orange dots are authors of tweets AGAINST atheism, red dots authors in FAVOR of 'climate change is a real concern'. All other users are represented as grey dots.

equally divided between these two groups. In this case, the social network information captured by N2V is not informative enough and, as a result, the tweet ends up being wrongly labeled as NORMAL by the LING+N2V model, i.e., there is no improvement over LING.

6.6.3 Graph Attention Network

As hypothesized, the GAT model is able to address the shortcoming of N2V described above. When creating a representation for the author of tweet (1), the GAT encoder identifies the connection to one of the authors of a hateful tweet as the most relevant for the task at hand, and assigns it the highest value. The user vector is updated accordingly, which results in the LING+GAT model correctly predicting the HATEFUL label.

This dynamic exploration of the social connections has the capacity to highlight homophily relations that are **less prominent** in the social network of a user, but **more relevant** in a given context. This is illustrated by how the models deal with tweet (3) in Table 6.3, which expresses a negative stance towards legalization of abortion, and is incorrectly classified by LING and LING+PV. The social graph contributes rich information about its author, who is connected to many users (46 overall). Most of them are authors who tweet in favor of feminism. The N2V model effectively captures this information, as the representation of the target user is close to these authors in the vector space. Consequently, by simply focusing on the majority of the neighborhood, the LING+N2V model misclassifies the tweet, i.e., it infers FAVOR for tweet (3) on the legalization of abortion from a social environment that mostly expresses stances in favor of feminism. However, the information contributed by the majority of neighbors is not

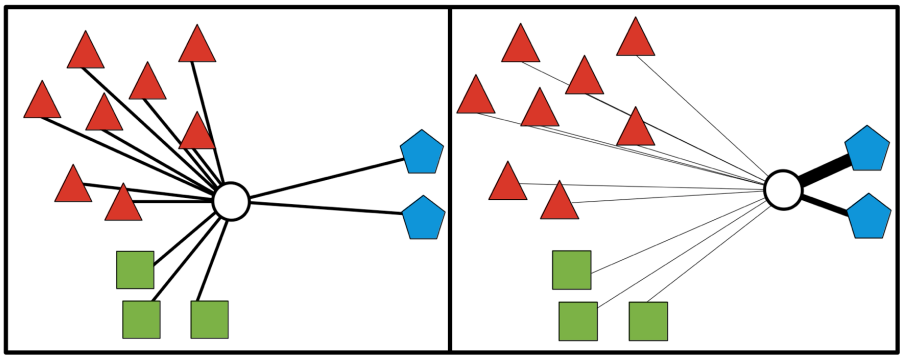


Figure 6.4: Left: Author of tweet (3) in Table 6.3 (white node) has many connections with users tweeting in favor of feminism (red triangles), fewer with authors tweeting in favor of Clinton (green squares) and against legalization of abortion (blue pentagons) (for simplicity, only some connections are shown). Recall that no label is available for nodes (users), and that their color is based on the label of the tweet they posted. Proximity in the space reflects vector similarity. Right: the GAT encoder assigns higher values to connections with the relevant neighbors (0.54 and 0.14; all other connections have values $\leq$ 0.02; thickness of the edges is proportional to their values) and updates the target author vector to make it proximal to them in social space.

the most relevant in this case. In contrast, the GAT encoder identifies the connections with two users who tweet against legalization of abortion as the most relevant ones, and updates the author representation in such a way as to increase the similarity with them, which leads the LING+GAT model to make the right prediction – see Figure 6.4 for an illustration of this dynamic process.

Interestingly, GAT is able to recognize when the initial N2V representation already encodes the necessary information. For example, the LING+N2V model correctly classifies tweet (4) in Table 6.3, as the N2V vector of its author is close in social space to that of other users who post hateful tweets (7 out of 10 closest neighbors). In this case, the LING+GAT model assigns the highest value to the self-loop connection, thus avoiding to modify a representation that is already well tuned for the task.

The error analysis reveals that there are two main factors that affect the performance of the GAT model. One is the size of the neighborhood: As the size increases, the normalized attention values tend to be very small and equally distributed, which makes the model incapable of identifying relevant connections. The second is related to the fact that a substantial number of users ($\sim$800 for Stance and $\sim$2.4k for Hate Speech) are not connected to the relevant sub-community. This means that in the case of Stance, for example, a user is not connected to any other individual expressing the same stance towards a certain topic. While external nodes in the graph (see Section 6.4.5) help to alleviate the problem by allowing the information to propagate through the graph, this lack of connections is detrimental to GAT.

6.7 Conclusion

In this chapter, I focused on user representations based on the connections in the **social graph**, investigating their usefulness in downstream NLP tasks. My goal was to account for one of the main aspects described in the theoretical framework introduced by Eckert and McConnell-Ginet (1992), namely, the fact that individuals belong to **several communities**, and that they adopt, each time, the practices of the community they are interacting with. To this end, I introduced a model that **dynamically** explores the connections of the users, identifies the ones that are more relevant for the task at hand, and computes user representations accordingly. The model, hence, captures the fact that not all the social connections of an individual are equally relevant in different communicative situations. I applied the model to three tasks involving the classification of user-generated texts, and showed that, when social information is proved useful, the dynamic representations computed by my model better encode **homophily relations** compared to the static representations obtained with other models, that uniformly aggregate the information from the connections. Finally, I performed an extended analysis of the performance of all the models that effectively encode author information, highlighting the strengths and weaknesses of each model.

As mentioned above, and explained in detail in Section 2.2.3, social connections are one of the two main sources of information that can be derived from the behavior of users in online social media. The second information source is the linguistic production of the users, which I investigate in the next chapter.

Chapter 7

Language-Based User Representations for Fake News Detection

At the time of writing, the content of this chapter has not been published.

7.1 Introduction

After showing how to represent users on social media based on their connections in the social graph, I now turn my attention to their **linguistic production**, which I leverage to perform fake news detection. As explained in Section 2.2.3, the language produced by users is a highly valuable source to model **homophily** relations. In the previous chapter, I showed that this is particularly true when such relations rely on shared social and cognitive factors related to entrenched beliefs and ideas, as in the case of users sharing homophobic ideas. The relation between these factors and language use has been documented also in previous studies, both in Sociolinguistics and Psycholinguistics (Pennebaker et al., 2003; De Fina, 2012), and, more recently, in NLP (Plank and Hovy, 2015; Preoţiuc-Pietro et al., 2017). Furthermore, other studies show that some people are more prone than others to spread fake news, and that these people usually present similar social factors (Pennycook et al., 2015; Pennycook and Rand, 2017). I build on these findings to address RQ-4 introduced in Section 1.2: How to leverage the linguistic production of a user to capture their tendency to spread fake news and, accordingly, to perform fake news detection?

I implement a model for fake news detection that, similarly to the model presented in Chapter 6, jointly models news and user-generated texts. Since I am particularly interested in understanding which are the linguistic features that characterize the language produced by fake news spreaders, I use Convolutional Neural Networks (CNNs), which, as explained in Section 2.3.1, allow me to extract the informative linguistic features of the input texts. I leverage two kinds of **user-generated texts**, namely timelines and self-descriptions, as I expect these two textual resources to provide different kinds of information about the users. I perform a set of experiments and show that the performance of the model improves when language-based user representations are used, and that it achieves surprisingly high results when leveraging *only* user representations.

I then present an extended analysis of the language of fake news spreaders, showing that it has distinctive features related to both **content** and **style**, and that these features are largely independent from the domain of the dataset, and consistent across datasets. The analysis also shows that the two kinds of user-generated language considered provide partially overlapping information, but with some relevant differences.

Finally, I consider the relation between the language produced by the users and their connections in the social graph. In particular, I investigate the **Echo Chamber effect**, i.e, the situation in which the ideas expressed by a user are reinforced by their social connections (Jamieson and Cappella, 2008). I operationalize this idea by introducing a methodology in which the linguistic production of a user is leveraged to define the content they produce, and the Echo Chamber effect is computed as a function of the similarity between the content of connected users and their distance in the social graph. By applying this methodology, I show the existence of the Echo Chamber effect in my data. I also provide an analysis of the characteristics of the effect, showing the relation between some of them and the sociolinguistic theories underpinning this thesis.

7.2 Related Work

Several studies exploited user information for fake news detection, leveraging different kinds of features. Gupta et al. (2013) and Zubiaga et al. (2016) leverage simple features such as longevity on Twitter and following/friends relations; however, the reported results show that these features have limited predictive power. Others use more informative features, such as users' political party affiliation, job and credibility scores (Long et al., 2017; Kirilin and Strube, 2018; Reis et al., 2019), manually annotated lists of news providers (Guess et al., 2019; Shu et al., 2019a), or a mix of these features (Shu et al., 2019b). All these studies report significant improvements on the task, showing that user information is indeed beneficial for it. However, they all present a common shortcoming, namely, they create user representations based on features that are either hard to retrieve or have to be manually defined, thus hindering the possibility to scale the methodology to large sets of unseen users. The key idea of the experiments introduced in this chapter, hence, is to leverage *only* the **language** produced by the users, a resource that is both highly **informative** and largely **available**, thanks to the ever-increasing amount of online communication discussed in Section 2.2.1. The availability of such data has been a crucial factor also for the study of the Echo Chamber effect, a phenomenon that has recently received large attention in NLP. Most of the studies on the topic implement a similar approach, whereby the Echo Chamber effect is said to exist if users that are connected in the social graph post the same content, where this content is usually a link to a web page from an annotated list (Del Vicario et al., 2016; Garimella et al., 2018; Gillani et al., 2018; Choi et al., 2020). In this chapter I adopt a similar approach, but, crucially, instead of defining content based simply on the sharing of a link, I propose to represent it based on the linguistic production of the users.

7.3 Data

This section describes the datasets used for the current study (7.3.1), and which kind of information is leveraged to represent the users in such datasets (7.3.2).

7.3.1 Datasets

I use two datasets, PolitiFact and GossipCop, available in the data repository FakeNewsNet[1] (Shu et al., 2018). Both datasets consist of a set of news labeled as either fake or real. PolitiFact (PF) includes political news from a single website, `https://www.politifact.com/`, whose labels were assigned by domain experts. News in GossipCop (GC) are about entertainment, and are taken from different sources. The labels of these news were assigned by the creators of the data repository. For each news

[1] `https://github.com/KaiDMML/FakeNewsNet`.

	fake	real	users	description
PolitiFact	362	367	20.7k	79%
GossipCop	2.5k	4.9k	62.5k	82%

Table 7.1: Statistics for each dataset after preprocessing: number of fake and real news; number of users; percentage of users for which a self-description is available. Timeline is available for all users.

in the datasets, its title and body are available,[2] together with the IDs of the tweets that shared the news on Twitter.

Titles and bodies are tokenized, a maximum length of 1k tokens for bodies and 30 tokens for titles is set, and news are defined as the concatenation of their title and body. Words that occur less than 10 times in the dataset are removed, and URLs and integers are replaced with placeholders. The tag <CAP> is added before any all-caps word in order to keep information about style, and the text is lowercased. Finally, only news that are spread by at least one user on Twitter are kept (see below). Each dataset is randomly split into train (80%) validation (10%) and test (10%). Table 7.1 reports the number of fake and real news per dataset after my preprocessing.

7.3.2 Users

The only information about users that I leverage is the language they produce, which is retrieved as follows. First, for each news, the users who posted the tweets spreading the news are identified.[3] Due to the volatility problem mentioned in Section 2.2.1 and already encountered in the previous chapter, for some news it is not possible to find any user. These news are removed from the datasets. Also, in both datasets there are some users who spread many news. One risk, in this case, is that the model may memorize these users, rather than focus on general linguistic features. For this reason only unique users per news are keep, i.e., users who spread only one news in the dataset. Finally, a set of maximum 50 users per news is randomly subsampled, in order to make the data computationally tractable. As a result, for each news a set including 1 to 50 users who retweet it is obtained (on average, 28 users per news for PolitiFact and 9 for GossipCop). For each of these users, their **timeline** (TL), i.e., the concatenation of their previous tweets, and their **description** (DE), i.e., the short text where users describe themselves on their profile, are retrieved. I expect descriptions and timelines to provide different information, the former being a short text written to present oneself, while tweets are written to comment on events, express opinions, etc. Note that descriptions are optional, and have to be intentionally provided by the users. A max-

[2]The body of the news is not in the downloadable dataset files, but it can be obtained using the code provided by the authors.

[3]Similarly to what I did for the experiments in the previous chapter, in order to identify users and retrieve their information, I query the Twitter API using the Python library `tweepy`.

imum length of 1k tokens for timelines and 50 tokens for descriptions is set, and the same preprocessing steps detailed in Section 7.3.1 are applied to both. Additionally, the tag <EMOJI> is added before each emoji. Table 7.1 reports the number of users per dataset, and the percentage for which a description is available.

7.4 Model

The architecture of the model used for the experiments in this chapter resembles the one introduced in Chapter 6, as it processes the information from the text to be classified and the related user information in parallel. More specifically, the model takes as input a news n and the set $U = \{u_1, u_2, ..., u_i\}$ of texts produced by the users who spread n, and classifies the news as either fake or real. Also in this case, the model consists of two modules, one for **news** and one for **user-generated texts**, that can be used in parallel or independently. Both modules are implemented using Convolutional Neural Networks (CNNs). As explained in Section 2.3.1, while CNNs are a less common choice when modeling language, they allow to easily identify the features in the input that are relevant for the final classification. As we will see, this is of the utmost importance for the current study, as in Section 7.7 I will use the features (i.e., n-grams) returned by the model to investigate the characteristics of the language produced by fake news spreaders.

The news module takes as input n and computes vector $\mathbf{n} \in \mathbb{R}^d$, where d is equal to the number of filters of the CNN (see below). The users module takes as input U and initially computes the matrix $\mathbf{U} \in \mathbb{R}^{m,d}$, where m is the number of users in U, and vector $\mathbf{u}_i \in \mathbb{R}^d$ represents user u_i in set U. I assume not all the users to be equally relevant for the final prediction, and I therefore implement a gating system as linear layer $\mathbf{W}_g \in \mathbb{R}^{d \times 1}$, that takes as input $\mathbf{U}$ and returns the vector $s \in \mathbb{R}^m$. A sigmoid function is applied to s, squeezing the values in it in range [0-1], where 0 means that the information from a user-generated text is not relevant, and 1 that it is maximally relevant. The matrix of the weighted representations of the users is thus obtained as $\mathbf{U'} = \mathbf{U} \times s$. User information is finally compressed in a single vector $\mathbf{u} \in \mathbb{R}^d$ computed as $\mathbf{u} = \sum_{i=1}^{m} \mathbf{u}_i \in \mathbf{U'}$. Vectors $\mathbf{n}$ and $\mathbf{u}$ are weighted by a gating system that controls for their contribution to the prediction, concatenated, and fed into a one-layer linear classifier $\mathbf{W} \in \mathbb{R}^{d+d \times 2}$, where 2 is the number of output classes (real and fake), that returns the logits vector $\mathbf{o} \in \mathbb{R}^2$, on which softmax is computed.

7.4.1 Extracting Linguistic Features from CNNs

Recently, model interpretability has gained much traction in NLP, and an increasing number of studies have focused on understanding the inner-workings and the representations created by neural models (Alishahi et al., 2019). My choice to inspect the model in order to extract the linguistic features it leverages for the final prediction is

inspired by this line of work, and, in particular, by the analysis of CNNs for text classification presented by Jacovi et al. (2018). In what follows, I describe how to extract the relevant linguistic features from the model.

As detailed in Section 2.3.1, a CNN consists of one or more convolutional layers, and each layer includes a number of filters, i.e, small matrices of learnable parameters that *activate* (i.e., return an activation value) on the n-grams in the input text that are relevant for the final prediction: The higher the activation value, the more important the n-gram is for the prediction. In order to leverage these properties of the CNN, all the relevant n-grams returned by the filters in the model are initially collected. Then, I assess which n-grams, among the collected ones, are relevant for the fake class, and which for the real class. This is done by considering the **contribution of each filter** to the two target classes, which is defined by the parameters in $\mathbf{W} \in \mathbb{R}^{d+d\times 2}$ (Jacovi et al., 2018). The contribution of filter f to the real and fake classes is determined, respectively, by parameters $\mathbf{W}_{f0}$ and $\mathbf{W}_{f1}$: If the former is positive and the latter negative, I say that f contributes positively to the real class, and, therefore, the n-grams detected by f are relevant for that class. Consequently, for n-gram x returned by the filter f with activation value v, the importance of x for the class real is computed as $R_v = v \times \mathbf{W}_{f0}$ and for the class fake as $F_v = v \times \mathbf{W}_{f1}$.

7.5 Experimental Setup

This section describes the different setups in which my model is tested and the baselines it is compared to (7.5.1), as well as the hyperparameter search for each model used in the experiments (7.5.2).

7.5.1 Setups and Baseline

The goal of this study is to assess the contribution of language-based user representations to the task of fake news detection. Thus, for each dataset, the following setups are implemented:

News This setup is similar to the '**LING**' setup implemented in the previous chapter. Also in this case, the goal is to assess the performance of the model when user information is not available.

TL / DE / TL+DE The model is provided only with user information. User information can be either the timeline (TL), the description (DE), or their concatenation (TL+DE).

N+TL / N+DE / N+TL+DE The model is provided with combined information from both news (N) and user-generated texts, that can again be in the three variants defined above.

	Model	News	User Information			Combined Information		
			TL	DE	TL+DE	N+TL	N+DE	N+TL+DE
PF	SVM	0.839	0.654	0.714	0.673	0.654	0.686	0.682
	CNN	0.865*	0.812	0.706	0.824	0.888$^{*\diamond}$	0.879*	0.882$^{*\diamond}$
GC	SVM	0.629	0.505	0.439	0.514	0.518	0.609	0.525
	CNN	0.641*	0.545	0.463	0.526	0.710$^{*\diamond}$	0.714$^{*\diamond}$	0.719$^{*\diamond}$

Table 7.2: Results on the test set, computed with binary F-score, for all the setups in my experiment. Standard deviation is in range [0.01-0.02] for all CNN setups. For CNN, I mark with * the results that significantly improve over setups in User Information, while $\diamond$ indicates a significant improvement over the News setup.

I implement two baselines. The first one is a Frequency baseline that, similarly to the one used in Chapter 6, randomly samples labels according to their frequency in the dataset. The second baseline is a Support Vector Machine (SVM, Cortes and Vapnik, 1995). SVMs have been shown to achieve results that are comparable to those by neural-based models on text classification tasks (Basile et al., 2018), and I thus expect this model to be a strong baseline.

7.5.2 Hyperparameter Search

For each setup, grid hyperparameter search is performed on the validation set using early stopping with patience value 10. I experiment with values 10, 20 and 40 for the number of filters, and 0.0, 0.2, 0.4 and 0.6 for dropout. In all setups batch size is equal to 8, filters focus on uni-grams, bi-grams and tri-grams, and I use Adam optimizer (Kingma and Ba, 2015) with learning rate of 0.001, β_1 = 0.9 and β_2 = 0.999. All the CNN modules have depth 1, and are initialized with 200-d GloVe embeddings pretrained on Twitter (Pennington et al., 2014).

The SVM baseline is trained on uni-grams, bi-grams and tri-grams.[4] When modeling user information, the user-generated texts of the users spreading the target news are concatenated. I use the `rbf` kernel, and perform grid hyperparameter search on the validation set. I explore values 1, 2, 5, 10, 15 and 30 for the hyperparameter C, and 1^{e-05}, 1^{e-04}, 1^{e-03}, 1^{e-02}, 1.0 for γ.[5]

For both CNN and SVM models, binary F-score is used as optimization metric, and the fake class is indicated as the target class.

7.6 Results

Table 7.2 reports the results of the fake news detection task. In the table, setups in which only information from user-generated texts is used (User Information) and those in which news and user-generated texts are jointly modeled (Combined Information) are grouped. The results of the CNN-based model are computed as the average of 5 runs with different random initialization of the best model on the validation set. For SVM, the single result obtained by the best model on the validation set is reported. The results of the Frequency baseline are not reported, as they are constant across setups.

Both CNN and SVM outperform the Frequency baseline, that obtains an F-score of 0.33 in GossipCop and 0.48 on PolitiFact. CNN outperforms SVM in all the setups, except for one. The largest improvements are in the TL and TL+DE setups for PolitiFact and in all the Combined Information setups. My intuition is that these improvements are due to the weighted sum of the user vectors and to the gating system of the CNN, which allow the model to pick the relevant information when the set of user-generated texts is large and includes long texts,[6] and when news and user-generated texts are jointly modeled.

I now focus on the performance of the CNN in the different setups. First, results in the News setup are significantly higher than those in the User Information setups.[7] This was expected, as classifying a news based on its text is presumably easier than by using only information about users who spread it. Nevertheless, the results in the TL setup are surprisingly high, especially in PolitiFact, which indicates that the language used in timelines is highly informative. The results in the DE setup, both in PolitiFact and GossipCop, are lower than those in TL. The two setups, however, cannot be directly compared, as descriptions are not available for all users (see Section 7.3.2). When the models in the User Information setups are re-run keeping only users with both timeline and description, there is no statistically significant differences between the results in the TL and DE setups. Lastly, no significant improvement is observed when descriptions are added to timelines – i.e., TL+DE and N+TL+DE do not improve over TL and N+TL, respectively. Finally, in all the Combined Information setups the performance of the model significantly improves compared to the News setup – except for N+DE in PolitiFact, for which the improvement is not statistically significant. When user vectors are substituted with random ones in the Combined Information setups, no improvement over the News setup is observed.

Overall these results confirm my initial hypothesis that leveraging user representations based only on the language produced by users is beneficial for the task of fake

[4] I use the `sklearn` implementation available at `https://scikit-learn.org`.

[5] Values for C and γ were defined following the guidelines provided here: `https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html`.

[6] Recall that, on average, there are 28 users per news in PolitiFact and 9 in GossipCop (see Section 7.3.2).

[7] I compute statistically significant differences between sets of results using the unpaired Welch's t test.

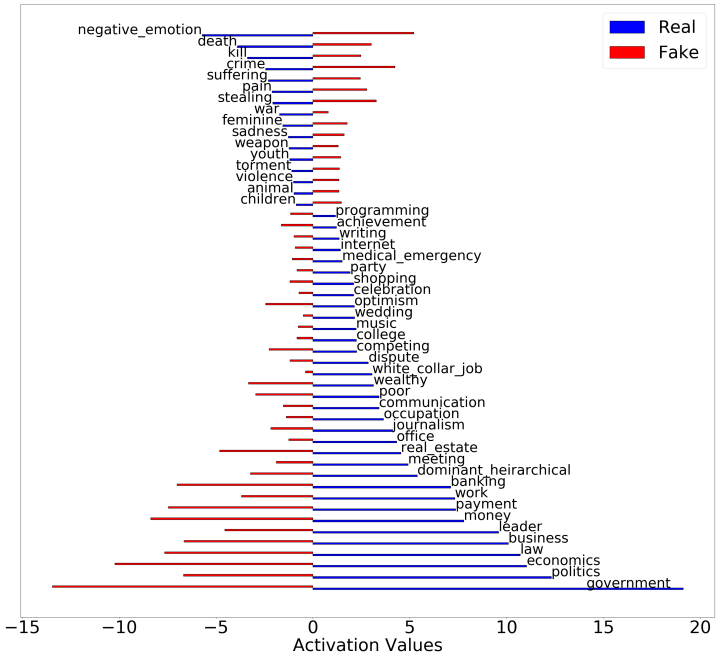


Figure 7.1: Activation values of topics for the News setup in PolitiFact.

news detection. They also raise interesting questions related to what makes user-generated language informative, and which qualitative differences exist, if any, between timelines and descriptions. I address these questions in the next section.

7.7 Linguistic Analysis

As mentioned earlier, one of the main goals of the current study is to uncover the linguistic features that characterize the language of the users who are more prone to spread fake news. Accordingly, in this section, I analyze the language of news and of user-generated texts, addressing two main questions:

Q1: Which features of the language used by **fake news spreaders** are relevant for fake news detection, and how are they different from those of the language used by **real news spreaders**?

Q2: Which linguistic features do **timelines** and **descriptions** share, and how do they differ? Also, which features do these two kinds of user-generated text share with the language of **news**?

To answer these questions, the language used in timelines, descriptions, and news has to be analyzed independently. I therefore consider, for both datasets, the models used at test time in TL, DE and News. For each model, the set of relevant n-grams is extracted. Subsequently, the R_v and F_v values for all the n-grams are computed, and the R_v and F_v of n-grams returned by more than one filter summed (see Section 7.4).

The n-grams are used to analyze both **content** and **style**. Regarding content, I analyze the **topic** of the n-grams, using the Empath lexicon (Fast et al., 2016); **proper names**, detected using the Python library `name-dataset`; and, for user-generated texts, **hashtags**, extracted by using regular expressions. Regarding style, I consider **punctuation marks** and **all-caps**, again identified through regular expressions; **function words**, detected with the LIWC lexicon (Pennebaker et al., 2001); and, for user-generated texts, **emojis**, identified with the Python library `emoji`. I check to which category, if any, each n-gram belongs to (e.g., 'trump' $\rightarrow$ proper names and '#usarmy' $\rightarrow$ hashtags). The category topic includes a list of topics (e.g., Politics and War), and n-grams are assigned to these topics (e.g., 'missile', 'army' $\rightarrow$ War). Similarly, the category function words includes several parts of speech (POS), hence, e.g., 'me', 'you' $\rightarrow$ Pronouns.

The importance of each topic and POS for the two target classes is defined by summing the R_v and F_v values of the n-grams they include. Finally, to consider only the n-grams that are relevant for one of the target classes, I compute the difference between R_v and F_v for each n-gram, compute the mean μ and standard deviation σ of the differences, and keep only n-grams whose difference is larger than $\mu + \sigma$.

Figure 7.1 shows the analysis of the topics for the News setup in PolitiFact. Red bars represent F_v values, blue bars R_v values: The higher the R_v (F_v) value, the more the importance for the real (fake) class. For example, the topics Negative Emotions and Death are important for fake news; Government and Politics for real news. Usually, to a large positive F_v value corresponds a large negative R_v value, and vice versa. I apply this methodology to address the two questions introduced at the beginning of this section. More specifically, I address **Q1** in Section 7.7.1 and **Q2** in Section 7.7.2.

7.7.1 The Language of Fake News Spreaders

Figure 7.2 shows the main categories of the language of fake news spreaders (red circles) and real news spreaders (blue circles) in PolitiFact (top) and GossipCop (bottom). Underlined categories refer to style, the others to content. For simplicity, similar topics are aggregated, e.g., 'positive emotions' includes topics such as Affection, Love and Optimism.

A first observation is that very few categories are shared by the language of fake and real news spreaders (overlap between blue and red circles), and that those in common are mostly related to the domain of the dataset (e.g., law and politics in PolitiFact). The language of fake news spreaders shows many common categories across datasets (overlap between red circles), mostly related to content. In particular, fake news spreaders of both datasets extensively talk about emotions and topics such as friendship, family, animals and religion. Interestingly, many of these topics are not directly related to the domain of either dataset. The most important proper names (e.g., Jesus, Lord, Jehovah, Trump) and hashtags (e.g., #usarmy, #trumptrain, #god, #prolife, #buildthewall) are again the same in the two datasets, and are highly related to the topics above. Some content-related categories that are not shared across datasets (non-overlapping areas in

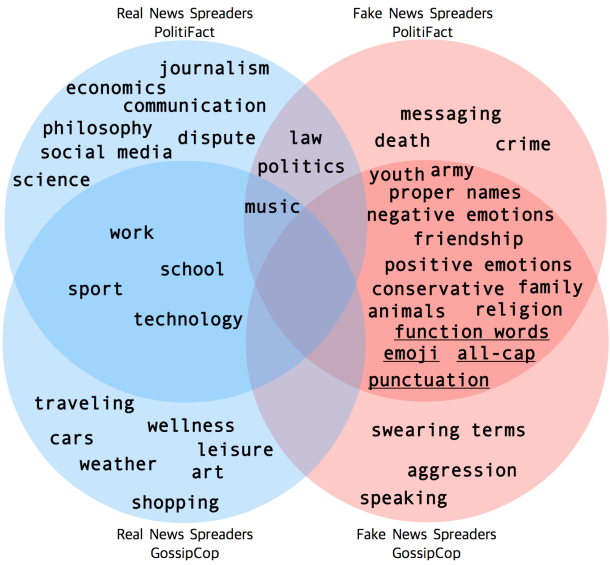


Figure 7.2: The language of real news spreaders (blue circles) and fake news spreaders (red circles) in PolitiFact (top) and GossipCop (bottom).

red circles) are observed, as they are related to the domain of the dataset (as we will see in Section 7.7.2). Cross-dataset consistency is even more evident for style: Fake news spreaders steadily use specific punctuation marks (quotes, hyphen, slash, question and exclamation mark), function words (first person pronouns and prepositions), emojis and words in all-caps.

The language of real news spreaders has different characteristics. Many categories are dataset specific (non-overlapping areas in blue circles), while few of them are shared (overlap between blue circles). Also, dataset specific categories have higher activation values and are related to the domain of the dataset. Finally, no relevant style-related category is found for the language of real news spreaders.

Overall, the analysis shows that the language of fake news spreaders is clearly characterized by a set of linguistic features, related to both style and content. Crucially, these features are largely **domain-independent**, and are consistently identified **across datasets**. This is in stark contrast with what is observed for the language of other users, which is more related to the domain of the dataset. These findings support the hypothesis that people who are more prone to spread fake news are connected by homophily relations based on the sharing of cognitive and sociological factors, and that these factors are mirrored in the features of the language they use.

7.7.2 The Language of Timelines, Descriptions, and News

I now analyze the relation between timelines, descriptions, and news. Figure 7.3 shows the relevant categories of timelines and descriptions for fake (a) and real (b) news spreaders, in both datasets. The plots include the same information displayed in Fig-

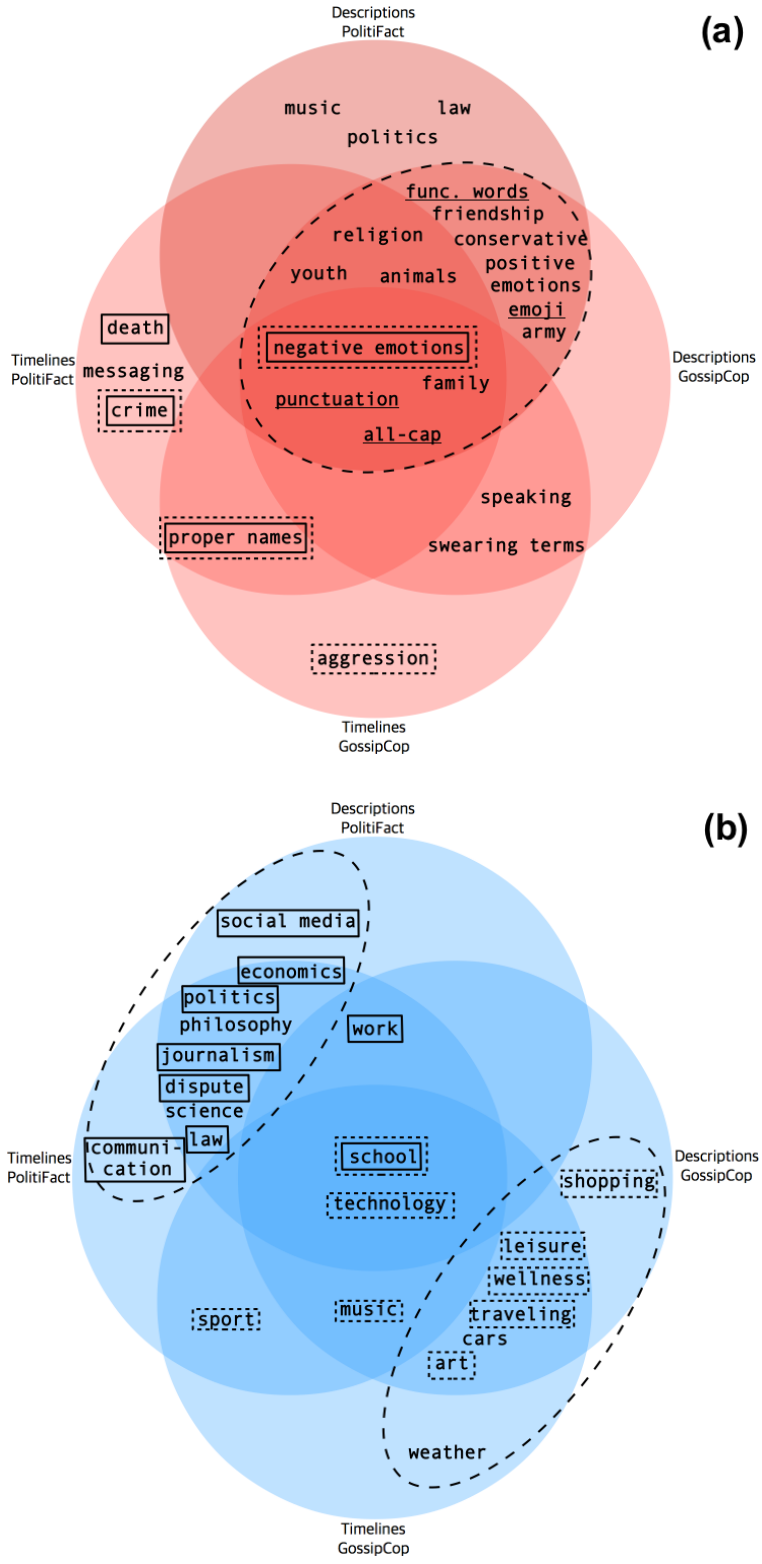


Figure 7.3: Relevant categories for timelines and descriptions of fake news spreaders (a) and real news spreaders (b). Solid-line boxes: categories of fake (a) and real (b) news in PolitiFact. Dashed-line boxes: categories of fake (a) and real (b) news in GossipCop.

ure 7.2, but in greater detail. In the two plots, solid-line boxes indicate the relevant categories for the news shared by fake/real news spreaders in PolitiFact, dashed-line boxes the relevant categories for news shared by fake/real news spreaders in GossipCop.

For fake news spreaders, I highlight the following findings. First, the largest overlap (dotted ellipse) is observed between the **descriptions** *across* the two datasets. Importantly, in this area we find the majority of categories which are not directly related to the domain of the datasets. Second, in both datasets, **timelines** have some categories shared with descriptions (e.g., Negative Emotions and Punctuation), plus other categories related to the semantic field of violence (e.g., Crime and Aggression), together with Proper Names. These timeline-specific categories are also the relevant ones for the fake news in PolitiFact (solid-line boxes) and in GossipCop (dashed-line boxes). The relevance of similar categories across datasets is due to the fact that in both of them fake news are often built by mentioning a famous person (mainly Trump in PolitiFact, a celebrity in GossipCop) in relation to some negative event – a usual scheme in **sensational news** (Davis and McLeod, 2003). In summary, all user-generated texts share some linguistic categories (central area of the plot), but it is in descriptions that we find the largest number of dataset-independent categories, related to both content and style, that characterize the language of fake news spreaders. Conversely, timelines share more categories with the news spread by the users. These findings are in line with my expectations about the **different nature of descriptions and timelines**, as the former include more personal aspects of a user, while the latter are more related to the domain of the news they spread. Furthermore, the limited similarity between the language of fake news spreaders and of the news they spread provides further evidence to the hypothesis that the language of fake news spreaders is largely shaped by **sociological** and **cognitive factors**, and mostly independent from the domain.

For real news spreaders, there is a large overlap of content-related categories between timelines and descriptions *within* a given dataset (dotted ellipses), while no style-related category is relevant for either kind of text. Differently from fake news spreaders, then, for real news spreaders descriptions and timelines do not present clear differences. Also, in both datasets, the relevant categories of real news strongly reflect the topics discussed in user-generated texts (see solid-line boxes for PolitiFact, and dashed-line boxes for GossipCop). Thus, it is possible to conclude that a set of **domain-related topics** exists in each dataset, and that these topics are the relevant linguistic categories in **timelines**, **description**, and in **news**. In contrast, these texts do not share any characteristic related to style.

7.8 Echo Chamber Effect

In this last section, I turn my attention to the relation between the **linguistic production** of the users and their connections in the **social graph**, thus jointly considering the two main kinds of information leveraged throughout Part Two. In particular, these

	nodes	edges	density
PolitiFact	32k	1.6m	0.0031
GossipCop	109k	4.9m	0.0008

Table 7.3: Graphs' statistics.

two kinds of user related information are used to investigate the Echo Chamber effect (ECE). I adopt the operational definition of the ECE proposed by Garimella et al. (2018), and say that the effect exists when users in a social graph mostly receive content from their connections that is similar to the content they produce. I thus introduce a methodology to define the content produced by a user based on their language use, and to compute the ECE as a function of the content similarity of connected users and their distance in the social graph.

7.8.1 Graph

As a first step, the social graph including the users and the connections among them is defined. I follow the approach described in Section 2.3.2, and already implemented in the previous chapter. For each dataset, hence, a graph $G = (V, E)$ is created, where V is the set of users in the dataset, and E is the set of edges between them. An unweighted and undirected edge is instantiated between two users if one retweets the other. The information about retweets is retrieved in users' timeline (see Section 7.3.2). In order to make the social graph more connected, users who are not in the dataset but have been retweeted at least 20 times by users in the dataset are added. Table 7.3 reports the main statistics for each graph.

7.8.2 User Representations

To represent users based on their linguistic production, I adopt an approach similar to the one of Section 7.7. First, the set of relevant n-grams and their activation values of each user is retrieved.[8] Since the ECE is related to the content posted by users, only the topic of the n-grams is considered, while their style is ignored.[9] For each user, the topics in their set of n-grams are analyzed using again the Empath lexicon (Fast et al., 2016), and a topic vector $t \in \mathbb{R}^d$ is defined, where d is the number of topics in the Empath lexicon, and t_i is the activation value of the i-th topic. I consider again the two kinds of user-generated language analyzed in the previous sections, and create two topic vectors per user, one based on the timeline (*TL-topic*) and one on the description

[8] In this case I ignore the class the n-gram is relevant for (i.e., the R_v and F_v values), and only consider value v (see Section 7.4).

[9] Proper names and hashtags are not considered because the dimensionality of the resulting user vectors would be intractable.

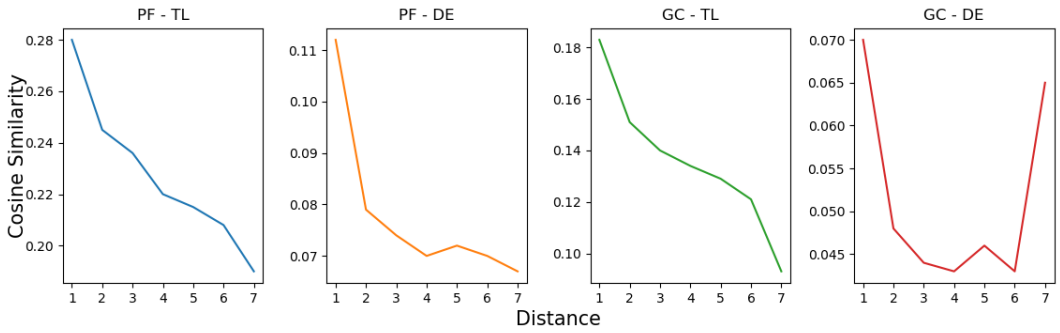


Figure 7.4: The similarity values obtained for the two datasets with the *TL-topic* vectors (TL) and with the *DE-topic* vectors (DE).

(*DE-topic*), using the best models at test time in the TL and DE setups introduced in Section 7.5.1.

7.8.3 Computing the Echo Chamber Effect

After defining the social graph and the user representations, I introduce a method to leverage them in order to compute the ECE. In particular, I conjecture that the ECE exists for a user if the cosine similarity between their topic vector and the one of their connections *decreases* as the distance (i.e., the number of hops away in the graph) *increases*. To check the effect for all users in the graph, I compute, for each distance value, the average cosine similarity of the users at that distance. There is no clear indication in the literature about the reach of the ECE, and it is therefore not possible to define a priori a maximum distance. I thus consider distance values for which there are at least 100 connections, which results in a maximum distance of 7 for all social graphs.

The relation between cosine similarity and distance is then assessed. As shown in Figure 7.4, it is possible to observe a **monotonic decrease in similarity** as the distance increases for all setups (Spearman $\rho \leq -0.9$, $p < 0.005$), except for GC-DE. In this setup, the decrease in similarity is much less pronounced and, consequently, the descending curve is more subject to fluctuations – see the increase after distance 6. Also, the Welch's t test between sets of values at consecutive distances (i.e., 1 and 2, 2 and 3, and so on) is computed, showing that a **significative negative difference** exists in all the setups up to distance 4 (Welch's t test $p < 0.005$). Overall, then, these results indicate that the ECE, as defined in my methodology, is present in the data used for the experiments in this chapter, even if with different strength depending on the setup. Also, the effect presents some relevant characteristics, that I describe in the rest of the section.

First, no difference is observed between fake and real news spreaders regarding the ECE. This indicates that the effect is common to all users in the datasets, and not related to the cognitive and social traits that influence the language production of fake

news spreaders observed in the previous section.

Second, in all the setups, the largest drop in similarity is observed between values at distances 1 and 2 or 2 and 3. I interpret this fact as an indication that the ECE is mostly at play, in my data, for these distance values. This result is consistent with one of the main aspects characterizing the adoption of common practices in social graphs, that is, the **reinforcement action** of **cliques of users** linked by first or second order connections, observed by both Milroy and Milroy (1985) and Labov (1972a). Arguably, the ECE observed in my experiments is related to the existence of such cliques.

I now focus on the difference between timelines and descriptions. Timelines show higher similarity values on average, while the drop in similarity at distance 2/3 is more evident for descriptions. I interpret these findings in the light of what was observed in Section 7.7 and my expectations regarding the different nature of the two kinds of language: Timelines share more **domain-related topics**, which causes them to be more similar to each other in general, while descriptions include more **personal aspects**, presumably shared with close connections belonging to the same clique, which causes the large drop in similarity observed beyond such connections.

Finally, the similarity values for both TL and DE are higher in PolitiFact than in GossipCop, which indicates that the ECE is stronger in the former dataset than in latter. I believe this result is due to the **polarization** of political groups in social networks, a phenomenon already observed in previous studies (Conover et al., 2011; Bakshy et al., 2015). These studies show that online communities made up of users belonging to the same political party are more inclined than other kinds of group to segregate in the social space, and to foster the discussion only within the community, while having few external connections.[10] My intuition is that this phenomenon is at play also in the data used in my experiments, and that it is responsible for the observed difference between PolitiFact and GossipCop.

7.9 Conclusion

In this chapter, I focused on the **linguistic production** of users in online setups, which I exploited to perform fake news detection. I showed that by leveraging user representations based *uniquely* on the language that users produce, it is possible to improve the performance of the model on the target task. This result confirms my initial hypothesis that the linguistic production of users can be leveraged to capture the **homophily** relations among them. Furthermore, I dedicated close attention to the analysis of the language used by different kinds of users, showing that the language of fake news spreaders has specific features related to both content and style, and that these features are, to a large extent, **independent from the domain** in which they are observed and

[10]Conover et al. (2011) find that polarization is particularly evident in social graphs based on retweet connections, i.e., the same kind of graph used in the present study.

consistent across datasets. Also, I jointly considered the two main sources of information leveraged in Part Two of the thesis, namely, linguistic production and social connections, to investigate the **Echo Chamber effect**. I introduced a methodology to compute this effect, whereby I was able to identify it in my data, and to uncover some relevant characteristics of such an effect.

Concluding, the results reported in this chapter offer empirical confirmation of the sociolinguistic studies underpinning it, and contribute new findings to the academic research on the topic. While this is of course a very relevant outcome, I also hope that the ideas and tools introduced in this study might offer concrete help to fight the diffusion of fake news on social media.

Chapter 8

Conclusion

Linguistic variation is a ubiquitous phenomenon in our daily communications. When talking with other individuals, we continuously create and share new linguistic practices, with no effort nor explicit agreement. The creation and spread of these practices allow us not only to communicate more effectively, but also to shape our social standing, as by adopting specific practices we mark our membership to the communities of people such practices are proper to. In turn, the combination of these community memberships is what defines our identity.

In this thesis, I focused on linguistic variation, investigating this phenomenon from different points of view, and with two main goals: describing how lexical variation operates in online communities, and making NLP models able to account for it. In order to reach these goals, at the beginning of this dissertation I defined four research questions, two of which are related to the first goal (RQ-1 and RQ-2), and two to the second goal (RQ-3 and RQ-4):

> **RQ-1**: *How to automatically represent and measure meaning variation in online communities of speakers?*
>
> **RQ-2**: *Which are the linguistic and societal processes that lead to variation?*
>
> **RQ-3**: *How to identify the relevant information coming from user connections in the social graph and leverage it to improve text classification?*
>
> **RQ-4**: *How to leverage the linguistic production of users to capture their tendency to spread fake news and, accordingly, to perform fake news detection?*

In this chapter, I review the main findings of the studies that addressed the research questions above (8.1), and discuss the limitations and possible extensions of such studies (8.2). I then reflect on the ethical aspects of my work (8.3), and conclude with some final remarks (8.4).

8.1 Main Findings

In Part One of this dissertation, I focused on the analysis of linguistic variation in online communities, in order to find an answer to the first two research questions listed above. I initially addressed RQ-1 (Chapter 3), introducing a framework that allows to identify and measure synchronic meaning variation of common words in online communities. I applied this framework to a set of online communities, showing that community-specific variation can indeed be observed, even across communities belonging to the same domain, and that the observed variation is independent from the topic discussed in a community. I then addressed RQ-2, which I investigated in two different studies. In Chapter 4, I presented an analysis of the main linguistic phenomena related to meaning change taking place over short periods of time in online communities. I also tested the performance of a standard NLP model for meaning change detection in detecting this community-related shift, highlighting its weaknesses, and proposing solutions to them. In Chapter 5, I analyzed the social processes related to change, showing that different users play different roles in the introduction and diffusion of linguistic innovations, and that it is possible to predict the success of an innovation based on the users who adopt it.

The three studies in Part One, by investigating different aspects of linguistic variation in online communities, jointly contribute to provide a multifaceted account of the target phenomenon, and to advance its **theoretical understanding**. In the first place, they provide large-scale confirmation of the sociolinguistic theories they build on, and that I introduced in Chapter 2. While these theories were proposed to account for variation in small offline communities, my studies show that they can explain variation also in large communities in online setups. Furthermore, the empirical findings reported in this part of the thesis also contribute new perspectives on the existing theories, and make new aspects of theoretical interest emerge. For example, Chapter 5 shows that the accounts of the role of innovator by Milroy and Milroy (1985) and by Labov (1972a), traditionally considered as competing, can actually provide a complementary characterization of innovators in online communities. Also, the study in Chapter 4 raises theoretical questions concerning the relation between change of context of use, semantic shift, and referential phenomena that are relevant for the field of CL, and that have been addressed by subsequent studies (Boleda, 2020).

In a nutshell, hence, the main finding of Part One is that the sociolinguistic processes defined by the reference theoretical frameworks can be effectively identified and described in online setups by means of computational tools. In Part Two I build on this finding, and move from the description of the relevant sociolinguistic processes to their exploitation for practical purposes. In particular, I leverage the concept of **homophily**, that is, the fact that similar individuals share similar linguistic and extra-linguistic practices. In the two studies Part Two consists of, my goal was to create user representations that capture homophily relations among users on social media platforms, and to use such representations to improve the performance of NLP models for text classifi-

cation. In Chapter 6, I focused on RQ-3, introducing a new methodology to represent homophily based on the connections among users in the social graph. Building on the idea that not all the connections are equally important in a specific communicative situation, I introduced a model which dynamically explores the social connections of a user in the social graph, identifies the most relevant ones for the target task, and creates the representations of the user accordingly. In Chapter 7, I addressed RQ-4, proposing a methodology to encode homophily in language-based user representations. This methodology relies on the assumption that the way a person uses language reflects the ideas and beliefs of that person. Since persons spreading fake news tend to share the same set of ideas and beliefs, I introduced a model that identifies in the linguistic production of social media users the relevant linguistic cues indicating their tendency to spread fake news, and exploits this information to improve the performance of an NLP model for fake news detection. Finally, I jointly considered the linguistic production of the users and their connections in the social graph and introduced a methodology to measure the Echo Chamber effect, that is, the situation whereby connected users mutually reinforce their ideas. I showed that such an effect is at play in the communities under scrutiny, and described its main characteristics.

Overall, Part Two provides relevant contributions to the field of NLP, as it introduces novel **neural architectures** and **methodologies** that, by efficiently creating and exploiting user representations, help to improve the performance of NLP models in several downstream tasks. It is worth noticing that, while in this thesis user information was leveraged to improve model performance in some specific tasks, the methodologies I introduced can be used for any task involving the automatic understanding of texts produced or shared by users in online setups. Finally, from a more general perspective, the studies in this part of the thesis provide further evidence to the basic assumption underlying the predictive approaches for text classification described in Section 2.2.3, namely, that social information about users has a crucial role in the understanding and modeling of the language derived from online communication.

8.2 Current Limitations and Possible Extensions

Part One As mentioned in the previous section, the studies in Part One have the merit to shed light on many aspects of linguistic variation in online setups. However, they also suffer from some limitations. For example, the study in Chapter 3 shows that significant differences exist in terms of meaning variation between communities at different levels of the hierarchical structure proposed by Clark (1996). It is unclear, however, which are the relations among these communities, and how linguistic innovations flow across the different levels of the structure. A related question arises from Chapter 5. The chapter describes how linguistic innovations are introduced and spread *within* online communities, but it does not consider the following phase, that is, the one in which such innovations go *beyond* single communities, are accepted in other communities and, possibly, in the general language. Also, none of these studies analy-

ses the characteristics and the role played by users belonging to multiple communities. While the study in Chapter 6 models this kind of users, it does not provide a description of the variation patterns they adopt, nor of their role in the process of innovation and spread of new linguistic practices.

In Chapter 4, an extended analysis of the linguistic phenomena related to short-term meaning shift is presented. However, the study does not consider the relation between the observed linguistic phenomena and the ones related to long-term meaning shift. Thus, many interesting questions remain open, for example: Which linguistic phenomena do short and long-term meaning shift have in common? Is the type of linguistic process observed in the short term related to the possibility of an innovation to be permanently adopted at a later stage?

Finally, in Chapter 5 I decided to focus on new linguistic *forms*, rather than new *meanings*, since the former are much easier to spot and track compared to the latter. While this choice allowed me to analyze the spread of a very large set of innovations, and, accordingly, to draw robust conclusions from my experiments, I still consider this a sub-optimal choice, as an analysis of the spread of new meanings would have been more coherent with the other studies included in Part One.

Part Two The main limitations of the chapters introduced in Part Two are related to the kind of data used in the experiments and to the processing of such data. In Chapter 6, I proposed a model that allows for the creation of dynamic user representations, that is, representations that can change depending on the communicative situation. While my experiments show that these representations are effective, I could not assess how the representation of a *single* user changes in different situations, due to the fact that in the data I used, almost all the users are involved in just one communicative situation (i.e., they are the authors of just one tweet). In Chapter 7, I had to face the fact that a significant number of users in the dataset spread a large number of news. For this reason, I considered unique users only, that is, users who only spread one news. By making this methodological choice, I ensured the model would not just memorize specific users. However, I also forced a dichotomous distinction between real and fake news spreaders, in which the information about users who spread *both* real and fake news is lost.

Also in this case, while my studies provide answers to specific research questions, they raise new, relevant questions. The first one regards the complementarity of the methodologies introduced in Chapters 6 and 7. These two chapters proved that both social connections and linguistic production provide relevant information about the users. However, it would be interesting to verify if these two sources of information can be used *together* to create even more informative user representations. At the end of Chapter 7, I partially investigated the relation between the two sources of information, but I did not address this relevant question. The investigation of the complementarity of different sources of information could then be extended by considering other sources, such as, for example, demographic information. As explained in Section 2.2.1, this

kind of information is highly valuable, but not always available, while, as we have seen, social connections are available for all the users on social media platforms. It might be worth, hence, to jointly model social connections and demographic information, for example, by using the former to propagate the latter. A further step along this line would concern the investigation of the complementarity of the social and linguistic sources of information used in my experiments with visual ones, such as, for example, the profile picture provided by the users. In general, I believe that the joint modeling of different sources of information would lead to the creation of more informative user representations, and, at the same time, would enable a more accurate analysis and characterization of the social standing of the users in online setups.

8.3 Ethical Considerations

The discussion about the social impact and the ethical aspects of the studies in the field of NLP has become increasingly important in the community, as witnessed by the set of resources available on the "Ethics in NLP" page of the Wiki of the Association for Computational Linguistics.[1] This is particularly true for the studies that, like the ones included in this dissertation, deal with data derived from human interactions. In this section, I point out some aspects of my work that are related to relevant ethical and social concerns.

In all my experiments, I leveraged user-generated data coming from online social media. A first relevant issue was how to manage these sensitive data. Several studies have been concerned with the ethical treatment of user-generated data, both in NLP and related fields (Vitak et al., 2016; Leidner and Plachouras, 2017; Schmaltz, 2018; Olteanu et al., 2019). These studies focused on different aspects, and proposed several good practices, which, to the best of my ability, I tried to follow. In the first place, I collected and used only data made publicly available by the users, obtained by using the APIs of the social media platforms introduced in Section 2.2.1. Since these data are public, no approval and informed consent from the users were needed. Secondly, when modeling the social information of the users, I only used the users and posts IDs assigned by the social media platform, and in no case I did try to trace it back to the real identity of the users. Finally, I tried to avoid preprocessing practices that could introduce biases. For example, I randomly extracted the Reddit posts used to create representations for global language in Chapters 3 and 4; similarly, I randomly sub-sampled the unique users considered in Chapter 7.

Despite my effort in trying to follow the good practices outlined above, I am aware of the fact that some biases which I cannot control for exist, and that they could affect the possibility to generalize the results of my studies. The most relevant of such biases is the **demographic bias**, that is, the fact that the datasets I used represent only specific sections of the population. In particular, Hovy and Spruit (2016) point out

[1] https://aclweb.org/aclwiki/Ethics_in_NLP.

that most of the available datasets in NLP include language produced by western, educated, industrialized, rich, and democratic individuals (*WEIRD*). This situation hinders the possibility to draw conclusions that apply to other social groups. Another potential problem related to my experiments, and, in particular, to the experiment on fake news spreaders presented in Chapter 7, is **overexposure**. My study finds that users belonging to a specific social group, characterized by a peculiar use of language, are more prone to misbehave (i.e., spread fake news) in online setups. In case other studies find further evidence about their misbehavior, these users would be overexposed, and this could potentially lead to their discrimination.

Given this risk, I would like to conclude this section by clarifying, once again, the main motivation of the study in Chapter 7. The idea for the study was born by the simple observation that some persons (even very close to me personally) do not seem to be able to distinguish real news from fake ones. This observation found theoretical grounding in studies which show that, while there are malicious users who consciously spread fake news for different (usually unethical) reasons, others do it simply because they are unable to spot fake news (Pennycook and Rand, 2017; Kumar and Shah, 2018). My goal, hence, was to implement a system that could help to automatically identify these vulnerable users, not to hold them up to public disdain but, rather, to warn them of the risk to be involuntarily involved in a harmful process.

8.4 Final Remarks

As mentioned in the Introduction, I consider the current dissertation as belonging to the research area of Computational Sociolinguistics, since all the experiments it includes focus on the interplay between society, language, and computation. This is a relatively young research area in the wider panorama of NLP and CL, as it only developed in the last few years, when the data and tools underpinning its existence became available. The raise of Computational Sociolinguistics coincided with the years of my PhD. This was a very lucky coincidence: I have always been interested in the social aspects of language, and I had no doubt about carrying out my investigation in this research area. My hope is that the ideas, results, and tools introduced in this thesis will help Computational Sociolinguistics to grow and will be of inspiration to future studies dedicated to the investigation of the relation between language and society.

Appendix A

Appendix to Chapter 4

I provide here the instructions given to the `LiverpoolFC` members who took part in the annotation of the dataset used for the experiments in Chapter 4.

What is this about? Words can acquire new meanings in short periods of time. Think about the word 'insane': until recently, it was only used in a negative way to say someone or something was mad. In the last couple of years, it has flipped its sense, and you can find it in sentences like: 'Salah scored an insane goal', meaning that the goal was amazing. Changes of meaning like this are very frequent, especially in the slang language used in online communities. The goal of this survey is to identify words that have recently changed their meaning in the r/LiverpoolFC subreddit.

Your task You will be presented with a set of target words. For each word, you will be shown a few posts from r/LiverpoolFC where the word is used. Some of the posts date back to 2011-2013, while the others have been written in 2017. We ask you to indicate whether the meaning of the word (the way the word is used) has changed between 2011-2013 and 2017. There are no right or wrong answers: we are interested in your opinion given you personal experience as a member of the subreddit and your observation of the sample posts. Please make a choice, even if it is difficult or unclear. If you have comments about a word, feel free to include them in the comment box. In the next page you will see some examples, coming from other communities of football fans, so you can practice, and then the actual survey will start. You can exit the survey at any moment pressing the 'finish' bottom.

Example 1

Target word: **CAN**

2011-2013

- We begin by drinking a warm **can** of Diet Coke
- What? Opened the **can** and poured it out
- Not even an empty beer **can**

2017

- In today's match we are wearing the **can**
- The apple pie, whipped cream, followed by downing a **can** of beer
- Hapoel-Inter 3-2: players suck when playing with that **can** kit...

[X] change: there is at least one post in 2017 where the meaning of the word is novel and different from 2011-2013
[ ] **no change**: the meaning is the same in 2011-2013 and 2017

In the example above, it makes sense to select change because in 2011-2013 the target word 'can' is always used with its standard meaning, while in posts 1 and 3 of 2017 it is used to talk about the third kit of the Team, whose colors recalled those of a Sprite can. Due to this similarity, the fans started to call that kit just the 'can'. Note: it's enough for one or two of the examples from 2017 to show a novel meaning for you to choose the change option.

Example 2

Target word: **TRANSFER**

2011-2013

- One thing Jose will improve for sure is **transfer** policy
- United believe they will be able to negotiate **transfer** fees down

2017

- Yeah we've nailed last **transfer** window
- they're both strikers arriving on a free **transfer** who are at the end of their careers

[ ] **change**: there is at least one post in 2017 where the meaning of the word is novel and different from 2011-2013
[X] no change: the meaning is the same in 2011-2013 and 2017

In this case, no change would be the appropriate answer. The meaning of the word "transfer" (move to another team) seems approximately the same in 2011-13 and 2017.

Example 3

Target word: **KID**

2011-2013

- Come on bro, I am a real fan, since I was a **kid**
- He's only 19 wow, people are worried about him not getting play time but this **kid** is still so young, his chance will come

2017

- I was there! The **kid** scored first if not mistaken
- should have been 1-3, the **kid** missed an easy chance

[X] change: there is at least one post in 2017 where the meaning of the word is novel and different from 2011-2013
[ ] **no change**: the meaning is the same in 2011-2013 and 2017

In this case, the target word shows a change. In 2011-2013, 'kid' is used with its standard meaning (child or youngster). However, in the posts from 2017 the word has become similar to a nickname: 'kid' is the word fans in this group used to call a specific player.

Example 4

Target word: **BEER**

2011-2013

- I was watching the game accompanied only with a bottle of **beer** and I fell asleep
- Have a great day and fingers crossed we'll smash these off, have a **beer** for me!

2017

- **Beer** and Sheva: what a wonderful night for Milan fans!
- Inter had three **Beer** and get drunk in Europa League
- And for Inter: **BEER BEER BEER**!!

[X] change: there is at least one post in 2017 where the meaning of the word is novel and different from 2011-2013
[ ] **no change**: the meaning is the same in 2011-2013 and 2017

Another change example. In the posts from 2011-2013, 'beer' is used with its standard meaning (to refer to the drink). But this is not so in the posts from 2017: when InterFC unpredictably lost the Europa League game to Hapoel Be'er Sheva fans of the other Italian teams transformed 'Be'er' into 'beer' and made a lot of jokes with this word, which was suddenly transformed into a *meme*.

Example 5

Target word: **BOX**

2011-2013

- We can deal with any cross they put into the **box**
- just don't let them play passes in and around the **box**

2017

- Good player, he's a true **box** to **box**
- Pretty realistic and outside-the-**box** thinking on my part

[ ]**change**: there is at least one post in 2017 where the meaning of the word is novel and different from 2011-2013
[X] no change: the meaning is the same in 2011-2013 and 2017

In this case, no change would be appropriate. In the posts from 2011-13 and in post 1 from 2017, 'box' is used to indicate the penalty area. Therefore, its meaning has not changed. Note: In post number 2 from 2017, 'box' is used as part of a common English expression. This use is not novel and hence does not indicate a change

Appendix B

Appendix to Chapter 5

Table B.1 includes the full results of the experiment presented in Section 5.5.2.
Tables B.2 and B.3 include the full results of the experiment presented in Section 5.6.

subreddit	**k=1**		**k=2**		**k= 3**		**k=4**		**k=5**		**k=6**	
	s	**w**	**s**	**w**	**s**	**w**	**s**	**w**	**s**	**w**	**s**	**w**
Android	0.5	0.45	0.63	0.45	0.65	0.44	0.73	0.45	0.76	0.44	0.81	0.44
apple	0.58	0.47	0.7	0.48	0.73	0.48	0.8	0.47	0.8	0.47	0.81	0.47
baseball	0.62	0.55	0.73	0.55	0.71	0.54	0.76	0.56	0.75	0.55	0.75	0.54
beer	0.51	0.49	0.62	0.49	0.68	0.49	0.69	0.48	0.72	0.48	0.73	0.48
boardg.	0.71	0.6	0.83	0.6	0.8	0.58	0.88	0.6	0.7	0.56	0.9	0.6
cars	0.63	0.57	0.73	0.58	0.77	0.56	0.85	0.57	0.8	0.53	0.88	0.57
FinalF.	0.61	0.58	0.59	0.57	1.0	0.56	-	-	-	-	-	-
Guitar	0.57	0.53	0.7	0.54	0.7	0.52	0.81	0.53	0.89	0.53	0.77	0.53
harryp.	0.53	0.51	0.5	0.5	1.0	0.51	-	-	-	-	-	-
hockey	0.74	0.61	0.8	0.61	0.89	0.62	0.76	0.58	0.83	0.61	0.94	0.62
Liverpool	0.61	0.52	0.61	0.51	0.67	0.53	0.67	0.49	-	-	-	-
Patriots	0.65	0.64	0.62	0.63	0.6	0.65	0.67	0.65	-	-	-	-
pcgaming	0.68	0.58	0.78	0.58	0.74	0.56	0.85	0.58	0.87	0.58	0.88	0.58
photo.	0.65	0.57	0.73	0.57	0.82	0.57	0.84	0.58	0.75	0.57	0.73	0.57
pokemon	0.54	0.48	0.66	0.48	0.69	0.48	0.69	0.48	0.71	0.47	0.71	0.47
poker	0.57	0.57	0.7	0.57	1.0	0.57	-	-	-	-	-	-
reddevils	0.56	0.54	0.54	0.53	0.6	0.52	0.6	0.5	0.4	0.5	1.0	0.49
running	0.6	0.55	0.67	0.56	0.71	0.55	0.82	0.55	1.0	0.55	1.0	0.54
StarWars	0.61	0.5	0.77	0.51	0.82	0.51	0.88	0.5	0.9	0.5	0.91	0.5
subaru	0.49	0.53	0.53	0.51	0.92	0.52	0.8	0.48	-	-	-	-

Table B.1: Probability of increase in dissemination of a linguistic innovation after being used by a strong (**s**) or weak (**w**) tie for **k** consecutive time bins. Missing values indicate no such condition was found in the community.

subreddit	**k=3**		**k=6**		**k=12**		**k=24**		**k=48**	
	ts	**wb**	**ts**	**wb**	**ts**	**wb**	**ts**	**wb**	**ts**	**wb**
Android	0.5	0.39	0.54	0.38	0.57	0.4	0.66	0.38	0.76	0.41
	0.6	0.52	0.62	0.5	0.65	0.53	0.73	0.51	0.81	0.53
apple	0.59$^{\#}$	0.64	0.71	0.61	0.8	0.63	0.82	0.62	0.84	0.61
	0.52$^{\#}$	0.55	0.64	0.52	0.75	0.53	0.76	0.54	0.79	0.52
baseball	0.64$^{\#}$	0.64	0.73	0.65	0.8	0.65	0.81	0.64	0.83	0.66
	0.54$^{\#}$	0.54	0.65	0.54	0.71	0.55	0.73	0.53	0.76	0.55
beer	0.48$^{\#}$	0.47	0.58	0.48	0.63	0.5	0.65	0.48	0.69	0.47
	0.49$^{\#}$	0.49	0.57	0.49	0.61	0.5	0.62	0.5	0.68	0.48
boardg.	0.55$^{\#}$	0.49	0.63	0.48	0.72	0.48	0.84	0.49	0.86	0.47
	0.56$^{\#}$	0.51	0.64	0.48	0.71	0.48	0.83	0.5	0.86	0.49
cars	0.64	0.58	0.72	0.6	0.77	0.6	0.79	0.59	0.83	0.6
	0.57	0.51	0.67	0.53	0.7	0.53	0.73	0.52	0.79	0.54
FinalF.	0.63	0.58	0.63$^{\#}$	0.6	0.8	0.59	0.82	0.6	0.86	0.61
	0.57	0.51	0.57$^{\#}$	0.54	0.75	0.52	0.77	0.53	0.83	0.55
Guitar	0.64	0.57	0.7	0.57	0.77	0.61	0.79	0.58	0.83	0.58
	0.6	0.5	0.65	0.51	0.72	0.54	0.75	0.52	0.8	0.53
harryp.	0.54	0.47	0.56	0.47	0.56	0.48	0.51	0.47	0.59	0.47
	0.57	0.5	0.6	0.49	0.57	0.52	0.53	0.5	0.6	0.5
hockey	0.73	0.62	0.73	0.62	0.73	0.64	0.8	0.62	0.9	0.62
	0.66	0.53	0.64	0.53	0.64	0.54	0.73	0.53	0.86	0.53

Table B.2: Classification results for the first half of the communities. **k=** length of the tie strength vector used for the prediction; **ts / wb=** results obtained using tie-strength information and weighted baseline. Difference between **ts** and **wb** is always significant (p <0.01) except when marked with $^{\#}$.

subreddit	k=3		k=6		k=12		k=24		k=48	
	ts	**wb**	**ts**	**wb**	**ts**	**wb**	**ts**	**wb**	**ts**	**wb**
Liverpool	0.57	0.52	0.64	0.49	0.57	0.48	0.71	0.52	0.79	0.51
	0.56	0.51	0.64	0.49	0.54	0.49	0.69	0.51	0.77	0.49
Patriots	0.71	0.67	0.78	0.69	0.82	0.64	0.83	0.64	0.85	0.61
	0.62	0.57	0.72	0.58	0.76	0.52	0.78	0.55	0.8	0.52
pcgaming	0.72	0.66	0.78	0.66	0.81	0.65	0.8	0.65	0.85	0.67
	0.65	0.55	0.72	0.56	0.76	0.55	0.73	0.53	0.79	0.56
photo.	0.69	0.58	0.71	0.59	0.76	0.6	0.79	0.6	0.87	0.62
	0.63	0.51	0.65	0.51	0.69	0.52	0.73	0.52	0.84	0.53
pokemon	0.65	0.51	0.6	0.51	0.68	0.51	0.72	0.5	0.77	0.53
	0.62	0.49	0.55	0.5	0.64	0.5	0.68	0.5	0.74	0.52
poker	0.55$^{\#}$	0.57	0.61$^{\#}$	0.58	0.62	0.61	0.7	0.58	0.76	0.61
	0.47$^{\#}$	0.51	0.52$^{\#}$	0.51	0.52	0.53	0.6	0.53	0.68	0.53
reddevils	0.56$^{\#}$	0.55	0.68	0.57	0.73	0.56	0.8	0.57	0.86	0.57
	0.52$^{\#}$	0.5	0.64	0.5	0.68	0.49	0.75	0.5	0.83	0.52
running	0.53	0.57	0.65	0.58	0.69	0.59	0.74	0.57	0.78	0.57
	0.5	0.51	0.59	0.53	0.63	0.53	0.68	0.51	0.73	0.52
StarWars	0.6$^{\#}$	0.59	0.6$^{\#}$	0.58	0.64	0.57	0.7	0.58	0.77	0.56
	0.54$^{\#}$	0.53	0.53$^{\#}$	0.5	0.56	0.52	0.61	0.51	0.71	0.5
subaru	0.65	0.6	0.69	0.58	0.73	0.61	0.76	0.58	0.77	0.62
	0.58	0.52	0.61	0.51	0.64	0.53	0.67	0.51	0.69	0.54

Table B.3: Classification results for the second half of the communities. **k=** length of the tie strength vector used for the prediction; **ts / wb=** results obtained using tie-strength information and weighted baseline. Difference between **ts** and **wb** is always significant (p <0.01) except when marked with $^{\#}$.

Appendix C

Appendix to Chapter 6

Tables C.1, C.2 and C.3 include, for each task, the results for LING, LING+PV, LING+N2V and LING+GAT as reported in Chapter 6, together with the standard deviation values computed on the ten runs of each model. Additionally, the precision, recall and F-score for each class are reported.

Model	**Av. Rec.**	**Negative**			**Neutral**			**Positive**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
LING	0.676 $\pm$0.005	0.585	0.656	0.618	0.684	0.678	0.680	0.737	0.694	0.712
LING+PV	0.671 $\pm$0.004	0.584	0.639	0.609	0.677	0.679	0.678	0.734	0.693	0.713
LING+N2V	0.672 $\pm$0.004	0.584	0.639	0.609	0.681	0.679	0.680	0.734	0.699	0.715
LING+GAT	0.666 $\pm$0.01	0.599	0.597	0.596	0.666	0.691	0.677	0.730	0.691	0.709

Table C.1: **Sentiment Analysis**: Average Recall across the three classes, plus precision, recall and F1 per class.

Model	**Av. Ag.**	**Against**			**Neutral**			**Positive**		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
LING	0.569 $\pm$0.01	0.730	0.625	0.672	0.355	0.462	0.399	0.446	0.490	0.466
LING+PV	0.601* $\pm$0.02	0.739	0.673	0.701	0.353	0.380	0.362	0.479	0.536	0.501
LING+N2V	0.629$^{*\diamond}$ $\pm$0.01	0.761	0.697	0.727	0.380	0.369	0.370	0.488	0.588	0.531
LING+GAT	0.640$^{*\diamond\dagger}$ $\pm$0.01	0.749	0.725	0.734	0.380	0.316	0.330	0.507	0.600	0.545

Table C.2: **Stance Detection**: Average F1 of the Against and Favor classes, plus precision, recall and F1 per class.

Model	**F1 Hateful**	**Normal**			**Hateful**		
		P	**R**	**F1**	**P**	**R**	**F1**
LING	0.624 $\pm$0.01	0.968	0.989	0.978	0.773	0.526	0.624
LING+PV	0.667* $\pm$0.02	0.974	0.983	0.979	0.730	0.621	0.667
LING+N2V	0.656* $\pm$0.008	0.972	0.986	0.979	0.742	0.589	0.656
LING+GAT	0.674$^{*\diamond\dagger}$ $\pm$0.005	0.973	0.989	0.980	0.765	0.605	0.674

Table C.3: **Hate Speech Detection**: F1 for the Hateful class, plus precision, recall and F1 per class.

Bibliography

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. Socioeconomic dependencies of linguistic patterns in twitter: a multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1125–1134, 2018.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557, 2019.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5):e19009, 2011.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J Silva, and Bryon C Wallace. Quantifying mental health from social media with neural user embeddings. In *Machine Learning for Healthcare Conference*, pages 306–321, 2017.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518. ACM, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

Jin Yeong Bak, Suin Kim, and Alice Oh. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 60–64. Association for Computational Linguistics, 2012.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

David Bamman, Chris Dyer, and Noah A Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 828–834, 2014a.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014b.

Federica Barbieri. Patterns of age-based linguistic variation in american english 1. *Journal of sociolinguistics*, 12(1):58–88, 2008.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. Simply the best: minimalist system trumps complex models in author profiling. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 143–156. Springer, 2018.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Shane Bergsma and Benjamin Van Durme. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, 2013.

Michelle A Blank and Donald J Foss. Semantic facilitation and lexical access during sentence processing. *Memory & Cognition*, 6(6):644–652, 1978.

Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 2020.

Constantinos Boulis and Mari Ostendorf. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 435–442. Association for Computational Linguistics, 2005.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies-Volume 1, pages 773–782. Association for Computational Linguistics, 2011.

Justine Cassell and Dona Tversky. The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, 10(2):00–00, 2005.

Jack K Chambers and Peter Trudgill. *Dialectology*. Cambridge University Press, 1998.

Kyunghyun Cho, Bart Van Merri
"enboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Nat- ural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics., 2014.

Daejin Choi, Selin Chun, Hyunchul Oh, Jinyoung Han, et al. Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, 10(1):1–10, 2020.

Eve V Clark and Herbert H Clark. When nouns surface as verbs. *Language*, pages 767–811, 1979.

Herbert H Clark. Inferring what is meant. *Studies in the perception of language*, pages 295–322, 1978.

Herbert H Clark. Making sense of nonce sense. *The process of language understanding*, pages 297–331, 1983.

Herbert H. Clark. *Using language*. Cambridge University Press, 1996.

Herbert H Clark and Richard J Gerrig. Understanding old words with new meanings. *Journal of verbal learning and verbal behavior*, 22(5):591–608, 1983.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. Predicting twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research*, 55:389–408, 2016.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM, 2013.

Hank Davis and S Lyndsay McLeod. Why humans value sensational news: An evolutionary perspective. *Evolution and Human Behavior*, 24(3):208–216, 2003.

Anna De Fina. Discourse and identity. *The Encyclopedia of applied linguistics*, pages 1–8, 2012.

Ferdinand De Saussure. *Course in general linguistics*. Columbia University Press, 2011.

Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. Tracing metaphors in time through self-distance in vector spaces. In *CLiC it 2016 - Third Italian Conference on Computational Linguistics*, 2016.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3): 554–559, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Penelope Eckert. Communities of practice. In *Encyclopedia of language and linguistics*. Elsevier, 2006.

Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100, 2012.

Penelope Eckert and Sally McConnell-Ginet. Communities of practice: Where language, gender, and power all live. In Kira Hall, Mary Bucholtz, and Birch Moonwomon, editors, *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99, 1992.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369, 2013.

Jacob Eisenstein. Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, and D. Watt, editors, *Handbook of Dialectology*. Wiley, 2015.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, 2014.

Heba Elfardy and Mona Diab. Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, 2013.

Noémie Elhadad, Shaodian Zhang, Patricia Driscoll, and Samuel Brody. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 516. American Medical Informatics Association, 2014.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079, 2010.

Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657, 2016.

Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Alessandro Provetti. The role of strong and weak ties in facebook: a community structure perspective, 2012.

Clay Fink, Jonathon Kopecky, and Maksym Morawski. Inferring gender from the content of tweets: A region specific example. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, 2016.

Tommaso Fornaciari and Dirk Hovy. Identifying linguistic areas for geolocation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 231–236, 2019.

Kenneth I Forster. Frequency blocking and lexical access: One mental lexicon or two? *Journal of Verbal Learning and Verbal Behavior*, 20(2):190–203, 1981.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922, 2018.

Dirk Geeraerts. *Theories of lexical semantics*. Oxford University Press, 2010.

Aniruddha Ghosh and Tony Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, 2017.

Philip Gianfortoni, David Adamson, and Carolyn P Rosé. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 49–59. Association for Computational Linguistics, 2011.

Eric Gilbert. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1037–1046, 2012.

Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, my echo chamber, and i: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, pages 823–831, 2018.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. Springer, 2016.

Yoav Goldberg. *Neural network methods in natural language processing*. Morgan & Claypool Publishers, 2017.

Johannes Gontrum and Tatjana Scheffler. Text-based geolocation of german tweets. In *Proceedings of the NLP4CMC 2015 Workshop at GSCL, Selbstverlegung, Duisburg*, pages 28–32, 2015.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In *Third international AAAI conference on weblogs and social media*, 2009.

Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6): 1360–1380, 1973.

Alex Graves. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer, 2012.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1): eaau4586, 2019.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501, 2016.

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40(3):52–74, 2017a.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Loyalty in online communities. In *Proceedings of the eleventh International Conference on Web and Social Media*, 2017b.

Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, 2012.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, 2016.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, 2018.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. Cascade: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/C18-1156`.

Libby Hemphill and Jahna Otterbacher. Learning the lingo? gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 305–314, 2012.

Susan C Herring. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, volume 39. John Benjamins Publishing, 1996.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Robert Hopper et al. Couples' personal idioms: exploring intimate talk. *Journal of Communication*, 31(1):23–33, 1981.

Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762, 2015.

Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.

Dirk Hovy, Afshin Rahimi, Timothy Baldwin, and Julian Brooke. Visualizing regional language variation across europe on twitter. *Handbook of the Changing World Language Map*, pages 3719–3742, 2020.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In *BlackboxNLP@ EMNLP*, 2018.

Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

Salud María Jiménez-Zafra, Arturo Montejo-Ráez, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. Sinai at semeval-2017 task 4: User based classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 634–639, 2017.

Simon Jones, Rachel Cotterill, Nigel Dewdney, Kate Muir, and Adam N Joinson. Finding zelig in text: A measure for normalising linguistic accommodation. In *Coling*, volume 2014, page 25th, 2014.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, 2014.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 553–562. ACM, 2016.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 25–30, 2015.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. Association for Computational Linguistics, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR), 2015*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR), 2017*, 2017.

Angelika Kirilin and Micheal Strube. Exploiting a speakers credibility to detect fake news. In *Proceedings of Data Science, Journalism and Media workshop at KDD (DSJM18)*, 2018.

Y Alex Kolchinski and Christopher Potts. Representing social media users for sarcasm detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1121, 2018.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM, 2015.

Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

David Burth Kurka, Alan Godoy, and Fernando J Von Zuben. Online social network analysis: A survey of research applications in computer science. *arXiv preprint arXiv:1504.05655*, 2015.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1384–1397. Association for Computational Linguistics, 2018.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

William Labov. The social motivation of a sound change. *Word*, 19(3):273–309, 1963.

William Labov. The social stratification of english in new york city. *ERIC*, 1966.

William Labov. *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press, 1972a.

William Labov. *Sociolinguistic patterns*. University of Pennsylvania Press, 1972b.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. Multilingual stance detection in social media political debates. *Computer Speech & Language*, page 101075, 2020.

Jean Lave, Etienne Wenger, et al. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jochen L Leidner and Vassilis Plachouras. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, 2017.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008.

Kristina Lerman, Megha Arora, Luciano Gallegos, Ponnurangam Kumaraguru, and David Garcia. Emotions, demographics and sociability in twitter interactions. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Chang Li and Dan Goldwasser. Encoding social information with graph convolutional networks for political perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, 2019.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, 2017.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, 2016.

Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and Alexander Gelbukh. Author profiling with doc2vec neural network-based document embeddings. In *Mexican International Conference on Artificial Intelligence*, pages 117–131. Springer, 2016.

William D Marslen-Wilson and Alan Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63, 1978.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In *LREC*, 2020.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Miriam Meyerhoff. Communities of practice. In *The Handbook of Language Variation and Change*. Blackwell, 2002.

Rada Mihalcea and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy. Methodological issues in using word embeddings in a sociolinguistic perspective: the case of contact-induced semantic variation across canadian twitter corpora. In *Empirical Studies of Word Sense Divergences across Language Varieties*, 2020.

James Milroy and Lesley Milroy. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384, 1985.

James Milroy and Lesley Milroy. Belfast: change and variation in an urban vernacular. In *Sociolinguistic patterns in British English*, pages 19–36. E. Arnold, 1987.

Lesley Milroy. *Language and social networks*. Blackwell, 1987.

Lesley Milroy. Social networks. In *The Handbook of Language Variation and Change*. Blackwell, 2002.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, 2018.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2019a.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, 2019b.

Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 7–16, 2018.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016.

Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics, 2010.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41 (2):10, 2009.

Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

Dong Nguyen and Carolyn P Rosé. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, pages 76–85. Association for Computational Linguistics, 2011.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 115–123. Association for Computational Linguistics, 2011.

Dong Nguyen, Dolf Trieschnigg, A. Seza Dogruöz, Rilana Gravel, Mariet Theune, Theo Meder, and Franciska De Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, 2014.

Dong Nguyen, A Seza Dogruöz, Carolyn P Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational Linguistics*, 2016.

Dong-Phuong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2013.

Bill Noble and Raquel Fernández. Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 29–38, Denver, Colorado, June 2015. Association for Computational Linguistics.

Daisuke Oba, Shoetsu Sato, Naoki Yoshinaga, Satoshi Akasaki, and Masashi Toyoda. Understanding interpersonal variations in word meanings via review target identification. In *Proceedings of the 20th International Conference on Computational Linguistics and Lntelligent Text Processing (CICLing2019)*, 2019a.

Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, and Masashi Toyoda. Modeling personal biases in language use by inducing personalized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2102–2108, 2019b.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

Jukka-Pekka Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

Silviu Oprea and Walid Magdy. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, 2019.

Jahna Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378, 2010.

John Paolillo. The virtual speech community: Social network and language variation on irc. *Journal of Computer-Mediated Communication*, 4(4):JCMC446, 1999.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, 2016.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Gordon Pennycook and David G Rand. Who falls for fake news? the roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity. *SSRN Electronic Journal*, pages 1–63, 2017.

Gordon Pennycook, Jonathan A Fugelsang, and Derek J Koehler. What makes us think? a three-stage dual-process model of analytic engagement. *Cognitive psychology*, 80:34–72, 2015.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

Lawrence Phillips, Kyle Shaffer, Dustin Arendt, Nathan Hodas, and Svitlana Volkova. Intrinsic and extrinsic evaluation of spatiotemporal text representations in twitter streams. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 201–210, 2017.

Barbara Plank and Dirk Hovy. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, 2015.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Who's (really) the boss? perception of situational power in written interactions. In *Proceedings of COLING 2012*, pages 2259–2274, 2012.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, 2015a.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9), 2015b.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 729–740, 2017.

Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking "gross community happiness" from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 965–968, 2012.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, 2017.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106, 2015.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–30, 2014.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44, 2010.

Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, Fabrício Benevenuto, and Erik Cambria. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

Alexander Romiszowski and Robin Mason. Computer-mediated communication. *Handbook of research for educational communications and technology*, 2:397–431, 1996.

Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics, 2011.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463, 2015.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

Rahmtin Rotabi and Jon M Kleinberg. The status gradient of trends in social media. In *Proceedings of the tenth International Conference on Web and Social Media*, pages 319–328, 2016.

Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. Competition and selection among conventions. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1361–1370. International World Wide Web Conferences Steering Committee, 2017.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R Mortensen, and Yulia Tsvetkov. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *arXiv preprint arXiv:2001.07740*, 2020.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. Automatic identification of arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, 2014.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, 2014.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, 2019.

Allen Schmaltz. On the utility of lay summaries and ai safety disclosures: Toward robust, open research oversight. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 1–6, 2018.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in

degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125, 2014.

Fabrizio Sebastiani. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 11–20. ACM, 2015.

Devyani Sharma and Robin Dodsworth. Language variation and social networks. *Annual Review of Linguistics*, 6, 2020.

Eva Sharma and Munmun De Choudhury. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, 2019.

Abdulhadi Shoufan and Sumaya Alameri. Natural language processing for dialectical arabic: A survey. In *Proceedings of the second workshop on Arabic natural language processing*, pages 36–48, 2015.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320. ACM, 2019a.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439, 2019b.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 941–950, 2019.

Ian Stewart and Jacob Eisenstein. Making "fetch" happen: The influence of social and linguistic context on the success of lexical innovations. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Terrence Szymanski. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 448–453, 2017.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics.

Trang Tran and Mari Ostendorf. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, 2016.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska Jong, and Theo Meder. An exploration of language identification techniques in the dutch folktale database. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012)*, 2012.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *The 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Jessica Vitak, Katie Shilton, and Zahra Ashktorab. Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 941–953, 2016.

Svitlana Volkova and Yoram Bachrach. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578, 2016.

Silvio Amir Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. Modelling context with user embeddings for sarcasm detection in social media. *CoNLL 2016*, page 167, 2016.

Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*, 2018.

Lilian Weng, Márton Karsai, Nicola Perra, Filippo Menczer, and Alessandro Flammini. Attention on weak ties in social and communication networks. *arXiv preprint arXiv:1505.02399*, 2015.

Etienne Wenger. *Communities of practice: Learning, meaning, and identity*. Cambridge University Press, 1998.

Christopher C Werry. Linguistic and interactional features of internet relay chat. in-susan c. herring (ed.), computer-mediated communication: Linguistic, social and cross-cultural perspectives, 47–63, 1996.

Paul Werth. Extended metaphor—a text-world account. *Language and literature*, 3 (2):79–103, 1994.

Derry Tanti Wijaya and Reyyan Yeniterzi. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40. ACM, 2011.

Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 955–964. Association for Computational Linguistics, 2011.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015.

Yi Yang and Jacob Eisenstein. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association of Computational Linguistics*, 5(1):295–307, 2017.

Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. Toward socially-infused information extraction: Embedding authors, mentions, and entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1461, 2016.

David Yarowsky. Word sense disambiguation. In *Handbook of Natural Language Processing, Second Edition*, pages 315–338. Chapman and Hall/CRC, 2010.

Yue Ying, Chen Peng, Chao Dong, Yang Li, and Yan Feng. Inferring event geolocation based on twitter. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, pages 1–5, 2018.

Michael Yoder, Shruti Rijhwani, Carolyn Rose, and Lori Levin. Code-switching as a social act: the case of arabic wikipedia talk pages. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 73–82, 2017.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardzic, Nikola Ljubešić, Jörg Tiedemann, et al. Language identification and morphosyntactic tagging: The second vardial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, 2018.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Community identity and user engagement in a multi-community landscape. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

Jichang Zhao, Junjie Wu, and Ke Xu. Weak ties: Subtle role of information diffusion in online social networks. *Physical Review E*, 82(1):016105, 2010.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, 2019.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

Abstract

The same word can be used by different people to mean different things. The observed meaning variation is not random, but determined by the social characteristics of the speakers using it. In particular, a crucial factor in determining the observed variation is the community an individual belongs to. This thesis investigates meaning variation in online communities of speakers with a twofold goal: providing an empirical account of the phenomenon in online setups, and leveraging it to improve the performance of NLP models.

I build on theoretical frameworks introduced in Linguistics and Sociolinguistics which describe meaning variation in offline communities. In order to investigate variation using digital data derived from online communities, I leverage the tools and methodologies developed in the fields of Natural Language Processing and Computational Linguistics.

The thesis consists of two main parts. The first part focuses on the general research question: how to identify and represent meaning variation in online communities of speakers? This part includes three descriptive studies that address this question from different points of view. Initially, I investigate meaning variation from a synchronic perspective, introducing a methodology to represent how word meaning varies in online communities. Subsequently, I consider the diachronic dimension, focusing both on the process of meaning shift which leads to the observed variation, and on the social dynamics underpinning this process. In the second part, I take a task-oriented approach, as I address the research question: how can social information be used to improve the performance of NLP models? I address this question in two studies. In the first one, I show how it is possible to leverage the information coming from the connections of a user on a social media platform, in order to obtain better results in tasks involving the classification of user-generated texts. In the second study, I show that the language produced by users on social media provides highly valuable information for the task of fake news detection.

Overall, this dissertation presents an extensive study of meaning variation in online communities of speakers, making two main contributions: On the one hand, it

contributes empirical confirmation of the findings of traditional sociolinguistic studies and provides new theoretical insights about meaning variation in online communities of speakers. On the other hand, it introduces new models and methodologies which, by leveraging information about the social context where language is produced, help to improve the performance of NLP systems for text classification.

Samenvatting

Eén en hetzelfde woord kan verschillende dingen betekenen als het door verschillende mensen gebruikt wordt. De variatie die je ziet in betekenis is niet willekeurig, maar wordt bepaald door de sociale kenmerken van sprekers. Een cruciale factor bij het bepalen van de waargenomen variatie is de gemeenschap waartoe een individu behoort. Dit proefschrift onderzoekt betekenisvariatie van sprekers binnen internetgemeenschappen met een tweeledig doel: het geven van een empirische benadering van dit fenomeen in internetomgevingen, en de bevindingen gebruiken om de prestaties van natuurlijke-taalverwerkingsmodellen te verbeteren.

Ik bouw voort op theoretische raamwerken ontwikkeld binnen de taalkunde de socialelinguïstiek, die betekenisvariatie in niet-digitale gemeenschappen beschrijven. Om betekenisvariatie in digitale data van internetgemeenschappen te onderzoeken gebruik ik gereedschappen en methodes afkomstig uit de onderzoeksvelden natuurlijke-taalverwerking en computerlinguïstiek.

Het proefschrift bestaat uit twee hoofddelen. Het eerste deel legt de nadruk op de algemene onderzoeksvraag: hoe kunnen we betekenisvariatie in internetgemeenschappen identificeren en representeren? Dit deel omvat drie beschrijvende studies die deze vraag vanuit verschillende standpunten benadert.

Eerst onderzoek ik betekenisvariatie van een synchroon perspectief, en introduceer een methodologie om de wijze waarop woordbetekenis varieert binnen internetgemeenschappen te representeren. Vervolgens bekijk ik de diachrone dimensie, waar ik de nadruk leg op het proces van betekenisverandering die leidt tot de waargenomen variatie, en op de sociale dynamiek die dit proces ondersteunen.

In het tweede deel hanteer in een taakgerichte aanpak bij het behandelen van de onderzoeksvraag: hoe kan sociale informatie worden gebruikt om de prestaties van natuurlijke-taalverwerkingsmodellen te verhogen? Ik richt me op deze vraag met twee studies. In de eerste laat ik zien hoe het mogelijk is om informatie over verbindingen van een gebruiker op een social media platform te gebruiken om tot betere resultaten te komen voor het automatisch classificeren van door gebruikers gegenereerde teksten. In de tweede studie laat ik zien dat de taal die wordt gebruikt door gebruikers op

social media erg waardevolle informatie bevat voor het automatisch detecteren van nepnieuws.

Over het geheel genomen presenteert dit proefschrift een uitgebreide studie naar betekenisvariatie tussen sprekers van internetgemeenschappen. Er worden twee belangrijke bijdragen geleverd: aan de ene kant levert het een empirische bevestiging van traditioneel onderzoek binnen de sociolinguistiek en biedt het nieuwe theoretisch inzichten over betekenisvariatie tussen sprekers van internetgemeenschappen. Aan de andere kant introduceeert het nieuwe modellen en methoden, die door gebruik te maken van informatie over de sociale context waar taal wordt gegenereerd, de prestaties van taaltechnologische systemen voor tekstclassificatie kunnen verhogen.

List of my Publications

- Mario Giulianelli, **Marco Del Tredici** and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

- **Marco Del Tredici**, Diego Marcheggiani, Sabine Schulte im Walde and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Dominik Schlechtweg, Anna Hätty, **Marco Del Tredici** and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

- **Marco Del Tredici**, Raquel Fernández and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

- Pushkar Mishra, **Marco Del Tredici**, Helen Yannakoudakis and Ekaterina Shutova. 2019. Abusive Language Detection with Graph Convolutional Networks. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

- **Marco Del Tredici** and Raquel Fernández. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

- Pushkar Mishra, **Marco Del Tredici**, Helen Yannakoudakis and Ekaterina Shutova. 2018. Author Profiling for Abuse Detection. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

- **Marco Del Tredici** and Raquel Fernández. 2017. Semantic Variation in Online Communities of Practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

- **Marco Del Tredici**, Malvina Nissim and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-IT)*.

- **Marco Del Tredici** and Núria Bel. 2016. Assessing the Potential of Metaphoricity of Verbs Using Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

- Alexandra Anna Spalek and **Marco Del Tredici**. 2016. Towards a Methodology for Automatically Clustering Selection Restrictions of Predicates. In *Proceedings of the XVII EURALEX International Congress*.

- **Marco Del Tredici** and Núria Bel. 2015. A Word-Embedding-based Sense Index for Regular Polysemy Representation. In *Proceedings of Workshop on Vector Space Modeling for NLP at NAACL 2015*.

- **Marco Del Tredici** and Malvina Nissim. 2014. A Modular System for Rule-based Text Categorization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies