UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Agile Social Media Analysis with Neural Networks

SILVIO AMIR ALVES MOREIRA

Supervisor: Doctor Mário Jorge Costa Gaspar da Silva
Co-Supervisor: Doctor Paula Cristina Quaresma da Fonseca Carvalho

Thesis approved in public session to obtain the PhD Degree in
Information Systems and Computer Engineering

JURY FINAL CLASSIFICATION: PASS WITH DISTINCTION

2018

Universidade de Lisboa

Instituto Superior Técnico

Agile Social Media Analysis with Neural Networks

Silvio Amir Alves Moreira

Supervisor: Doctor Mário Jorge Costa Gaspar da Silva
Co-Supervisor: Doctor Paula Cristina Quaresma da Fonseca Carvalho

Thesis approved in public session to obtain the PhD Degree in Information Systems and Computer Engineering

Jury final classification: Pass with Distinction

Chairperson

Doctor Luis Eduardo Teixeira Rodrigues, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee

Doctor Mário Jorge Costa Gaspar da Silva, Instituto Superior Técnico, Universidade de Lisboa

Doctor Luis António Diniz Fernandes de Morais Sarmento, Google UK

Doctor João Miguel da Costa Magalhães, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Doctor Maria Luísa Torres Ribeiro Marques da Silva Coheur, Instituto Superior Técnico, Universidade de Lisboa

Funding Institution

Fundação para a Ciência e Tecnologia (FCT) Ph.D. scholarship SFRH/BD/89020/2012.

2018

Abstract

This thesis proposes an agile framework to accelerate the development of Social Media Analysis (SMA) systems by tackling some of the fundamental challenges of processing user generated content and the main limitations of current methodologies. The noise, brevity and ambiguity of social media pose challenges to traditional NLP methods, often forcing analysts to rely on sub-optimal methods or devote extensive manual efforts in the development of specialized models. On the other hand, social media users are not a representative sample of the population. Yet, current approaches tend to ignore the inherent biases of social media, and thus the outcomes of the analyses might not reflect broader trends. The proposed framework relies on a novel method to derive low-resource supervised neural networks in two steps: (i) learning unsupervised neural embeddings for words and users; (ii) constructing *minimalist* neural architectures that yield low-capacity models, which can be trained with scarce labeled data. By reducing the efforts of developing specialized models, the framework facilitates the deployment of more sophisticated SMA systems. In this work, it is used to implement methodologies to sample demographically representative *digital cohorts* of social media users. These cohorts can be leveraged to conduct demographically controlled studies, thereby mitigating sampling biases and allowing analysts to extrapolate findings gleaned from imperfect datasets and affording deeper insights. The evaluation of the framework was conducted over two case-studies, one involving the development of bespoke classifiers for social sciences studies, and the other concerned with the deployment of DEMOS, a novel digital epidemiology system to track public health discussions and monitor the prevalence of mental illnesses.

Resumo

Esta tese propõe uma plataforma ágil para acelerar o desenvolvimento de sistemas de Análise de Social Media (ASM), abordando alguns dos principais desafios inerentes ao processamento automático de conteúdos gerados por utilizador e as principais limitações das metodologias atuais. O ruído, a brevidade e a ambiguidade destes conteúdos colocam desafios aos métodos tradicionais de PLN, muitas vezes forçando os analistas a utilizar a métodos subótimos ou a dedicar grandes esforços ao desenvolvimento manual de modelos especializados. Por outro lado, os utilizadores de social media não são uma amostra representativa da população. No entanto, as abordagens atuais tendem a ignorar os vieses inerentes a estes dados, e como tal os resultados das análises podem não refletir as tendências do público em geral. A plataforma proposta assenta num método novo para derivar redes neuronais supervisionadas com recursos limitados em duas etapas: (i) induzir vectores neuronais não-supervisionados para palavras e utilizadores; (ii) construir arquiteturas neuronais *minimalistas* que produzam modelos de baixa capacidade, que podem ser treinados com poucos dados anotados. Com a redução dos esforços de desenvolvimento de modelos especializados, a plataforma facilita a produção de sistemas de ASM mais sofisticados. Neste trabalho, esta é usada para implementar metodologias de amostragem de *coortes digitais* demograficamente representativas de utilizadores de social media. Estas coortes podem ser usadas para conduzir estudos demograficamente controlados, mitigando assim os típicos vieses de amostragem, o que permite aos analistas extrapolar resultados obtidos a partir de dados imperfeitos e derivar ilações mais profundas. A avaliação da plataforma foi feita em dois casos de estudo, o primeiro envolveu o desenvolvimento de classificadores especializados para estudos no domínio das ciências sociais, e o outro consistiu na implementação do DEMOS, um novo sistema de epidemiologia digital para rastrear discussões relacionadas com saúde pública e monitorar a prevalência de doenças mentais.

ACKNOWLEDGMENTS

I would like to thank my supervisors, Mário Silva and Paula Carvalho for all the guidance and support along the way. Mário in particular always managed to pulled me back to the right path whenever I got too lost or sidetracked. I would also like to thank Rámon Astudillo for guiding me through the world of neural networks and machine learning — I appreciated all the brainstorming and last-minute submissions to shared task competitions. The work we have done together was one of the catalysts of my research. During my research I had the opportunity to visit American universities and companies. I would like to express my gratitude to Byron Wallace for hosting me at the University of Texas at Austin and inviting me to Northeastern University. I appreciated all the interesting discussions and the nice work we did together. I would also like to thank him for the funding and the help with all the bureaucratic hurdles that came along the way. I would like to thank Glen Coppersmith for offering me a research internship at *Qntfy* — it was a really fun and inspiring experience, so I extend my gratitude to the rest of the team. Finally, I would like thank Mark Dredze and John Ayers for the opportunity to do an internship at *Good Analytics*. I appreciated the opportunity to apply the methods that resulted from my research in innovative applications to improve public-health practices. It is my hope that this work leads to more research and development of computational technologies aimed at improving human lives at a personal and societal level.

Table of Contents

Page

List of Tables

Table **Page**

LIST OF FIGURES

FIGURE **Page**

Chapter 1

Introduction

The emergence of social media networks profoundly transformed the ways in which people communicate, interact and access information[1]. Nowadays, billions of people use these platforms daily to connect and share anything with close friends or the public at large — from personal experiences to reactions to worldly issues and events. These digital interactions leave behind a wealth of data about personal interests, behavior and moods, which provide a lens through which human and societal dynamics can be investigated at unprecedented scales (Helbing and Balietti, 2011). The availability of large volumes of personal data has the potential to revolutionize disciplines concerned with human activities that have lacked enough data to afford data-driven methods, such as the social sciences (Lazer et al., 2009), public health and epidemiology (Paul and Dredze, 2011; Salathe et al., 2012), and clinical psychology (Coppersmith et al., 2014a), among others. Traditional data collection methodologies for these disciplines (e.g. polls, surveys and direct interviews) are hampered by several limitations, such as the difficulty in reaching participants, the lag time between campaign design, data collection and data availability, as well as the significant associated costs. Such methodologies are thus difficult to scale to large samples, leading experts to rely on incomplete and outdated surveillance data (e.g. epidemiologists can only assess the impact of health interventions long after the fact).

In contrast, social media analysis systems can tap into much larger datasets to keep track of specific issues in real-time and, once deployed, can support longitudinal analyses

[1] http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

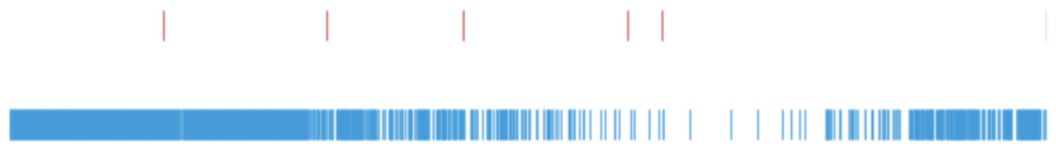


Figure 1.1: An example of a person's interactions with the health care system (red hashes), and Facebook posts (blue) over a period of four years. The ability to extract signals from social media feeds could inform clinicians about general patients well-being, provide complementary information for diagnosis, and help tracking disease progression or intervention outcomes. This plot is a reproduction from (Coppersmith et al., 2017).

with minimal maintenance costs. Moreover, these methods can potentially reach swaths of the population that tend to be excluded or underrepresented in traditional estimates, e.g. people that do not have access to the health-care system (see Figure 1.1).

1.1 Social Media Analysis

Social Media Analysis is based on the assumption that the topics and emotional tone of social media chatter reflects the *zeitgeist* and correlates with external events and circumstances. Therefore, by mining this data one can get insights into current states of affairs — e.g. public opinion on relevant issues (O'Connor et al., 2010) and socio-demographic indicators (Mitchell et al., 2013) —, and even predict the outcomes of future events, e.g. stock-market movements (Bollen et al., 2011b), movie box-office revenues (Asur et al., 2010) or electoral results (Tumasjan et al., 2010). In practice, SMA systems are operationalized as data processing pipelines, as depicted in Figure 1.2, to realize the following main steps:

1. **Data Collection:** find and retrieve the relevant content (e.g. tweets about a target politician);

2. **Natural Language Processing:** extract specific linguistic signals from the data, leveraging Natural Language Processing methods (e.g. opinions about said politician). These methods can be operationalized with lexicon-based approaches or with machine learning models, and produce analytics over posts, individual users, or groups of users.

3. **Data Analysis**: aggregate and process the extracted linguistic signals so as to transform these data into statistics, which can then be interpreted directly or as

proxies for other measurements (e.g. voting intentions).

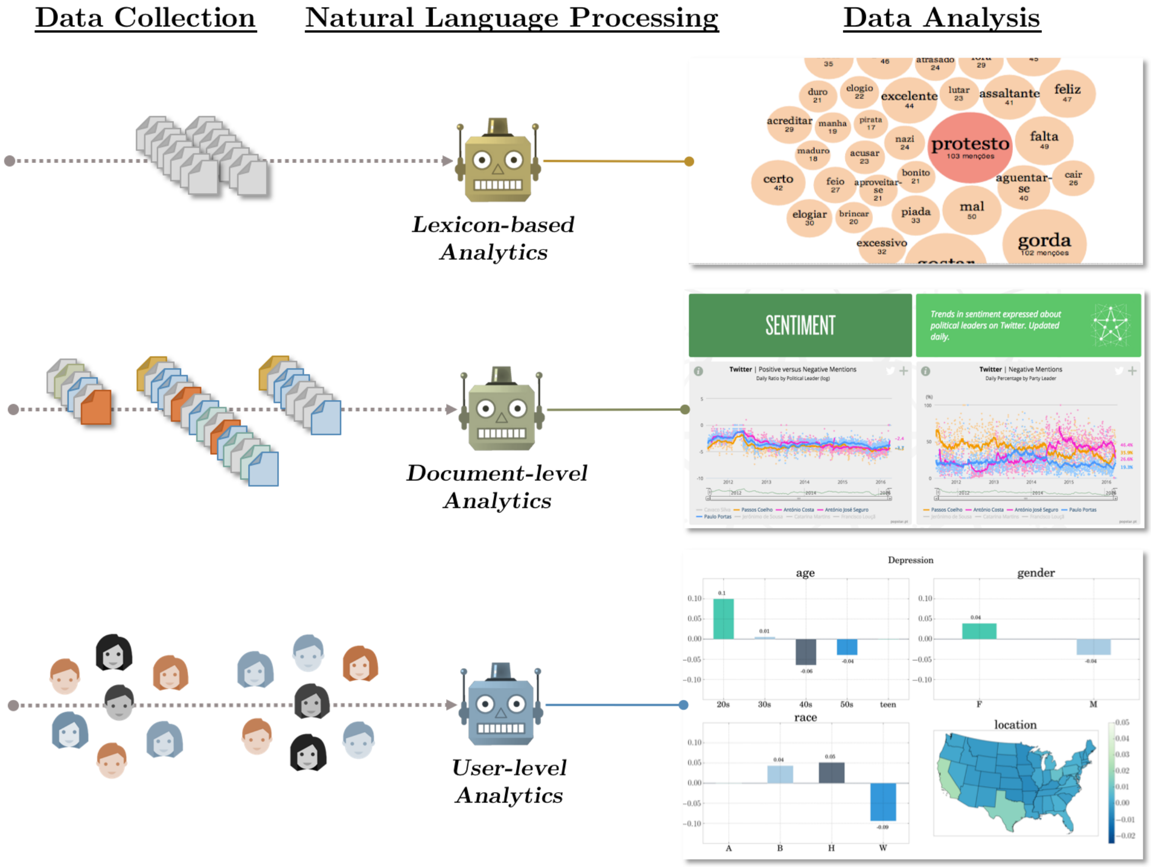


Figure 1.2: High-level depiction of the main steps of a social media analysis pipeline. The plots were taken from real-world analyses conducted in this thesis. The topmost shows the reactions on Twitter to a visit of the German Chancellor Angela Merkel to Portugal, measured with a lexicon-based system — the plot shows sentiment-bearing words used in tweets about the event, as bubbles of size proportional to their frequency. The middle is from POPMine, a political opinion mining for the web, that uses document-level supervised models to track the popularity and sentiment about Portuguese political entities, over time (Chapter 2). The bottommost is from DEMOS, a digital epidemiology system focused on mental-health issues, that will be described in Chapter 6 — the plot shows the relative prevalence of depression across demographic groups, estimated with user-level models over a cohort of US Twitter users.

The potential impact of social media analysis across a wide range of domains fostered significant research in these technologies, yielding promising results for a variety of applications. Nevertheless, in many cases, claims about the merits of the proposed methods have been somewhat exaggerated. For example, approaches claiming that could have predicted the outcome of an election with simple methods based on counting the mentions of political candidates and rudimentary sentiment analysis, performed no better than a random estimator when applied to other elections (Chung and Mustafaraj, 2011; Metaxas et al., 2011; Jungherr et al., 2012). These works have been widely criticized for oversimplifying complex problems, ignoring the inherent biases of social media, failing to justify arbitrary methodological decisions (e.g. excluding smaller parties from the analyses), and relying on simplistic NLP methods to analyze subjective and nuanced communications (Gayo-Avello, 2013). The peculiarities of user generated content and social media platforms poses new problems to traditional NLP methods and requires more sophisticated data analysis methodologies.

This thesis aims to address these problems and lay the necessary groundwork go beyond the somewhat simplistic SMA methodologies that have been employed thus far. First, by shedding light into the main challenges of building accurate social media mining systems and their inherent limitations; and second, by applying the resulting insights into tools and methodologies to improve the deployment of SMA systems for various applications. Let us begin then, by taking a closer look into the main tasks involved in building SMA systems and their associated challenges. As a motivating example, we will consider the development of a political opinion mining system for Twitter data but this discussion is equally relevant for other domains and applications.

1.1.1 Data Collection

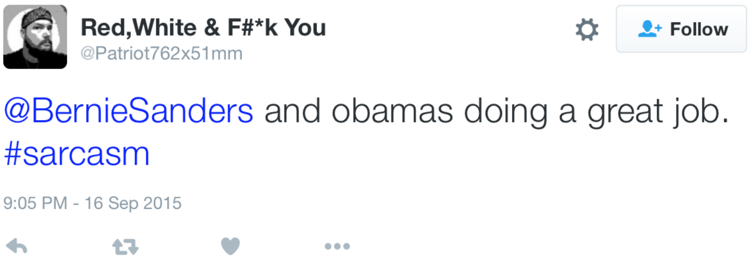
The first step of building a SMA system is setting up data collection methods and defining a set of search keywords to retrieve relevant content. Selecting the right set of search terms, however, might not be a trivial task, e.g. a politician might be referred by various names, nicknames or titles (e.g. *the prime minister*). This is even harder for more abstract topics (e.g. *the economy*) — too broad terms might yield a lot of unrelated content but too narrow terms might miss a great deal of the discussion. On the other hand, Twitter serves multiple purposes and hosts content created by billions of users, some of which correspond to individual citizens but others belong to organizations, brands, media outlets, celebrities and *social bots*, among others. Incidentally, the latter usually generate much more content, hence we can easily run into the problem of the 'vocal' minority overshadowing the 'silent' majority. Particularly, if these entities have a vested interest in the matter at hand or there are incentives to manipulate the public perception, which raises concerns regarding the credibility of the data. Indeed, we have seen concrete examples of this issue in the 2016 elections in the United States and France, and recent research shows an increase in socket-puppet Twitter accounts (Bessi and Ferrara, 2016; Ferrara, 2017).

Finally, it should also be noted that the population of Twitter users is not a representative sample of the general population (e.g. urban young adults tend to be overrepresented). As such, trends or insights gleaned from these data might not be generalizable to the general public. All these issues must be taken into account when trying to use social media to assess the public's attitudes about a given matter or event. Social media datasets are inherently noisy and thus näive data collection strategies, e.g. ignoring *self-selection* biases and inadvertently including unrelated content, can severely thwart the validity of the subsequent analyses.

1.1.2 Model Development

The second step concerns with the design and implementation of NLP systems to infer relevant signals from the data. However, the informal and spontaneous nature of social media communications poses significant challenges to standard NLP methods, which were created for conventional language and editorial domains. Social media content tends to be *noisy*, frequently containing misspellings, internet slang and other extra-linguistic markers (e.g. emoji and emoticons); and short, providing little context to appreciate their intended meaning. Therefore, generic text analysis methods that have been successful for many objective tasks (e.g. topic detection or spam filtering) are insufficient for inferences over subjective social media. On the one hand, manually crafted lexicons fail to capture the informal style of social media language. On the other hand, simple statistical models based on bag-of-word representations are hampered by feature sparsity issues, arising from the large vocabularies needed to keep track of all the lexical diversity and the brevity of individual posts. Furthermore, social media users often express subjective views on complex issues with humor and sarcasm (Carvalho et al., 2009). As an example, consider the tweet in Figure 1.3 — without the #*sarcasm* marker it would be difficult to conclude with certainty whether the remark was intended sarcastically or in earnest. Subjective and figurative language are more nuanced and ambiguous hence requiring more sophisticated modeling approaches, e.g. taking the *context* of an utterance into account is critical to discern ironic intent (Wallace et al., 2014).

Overcoming these issues requires considerable manual efforts in the annotation of large labeled datasets and the research and development of specialized features for supervised machine learning models. This often requires custom NLP tools and resources which are language and domain dependent and thus cannot be easily transferred



Red,White & F#*k You
@Patriot762x51mm

@BernieSanders and obamas doing a great job. #sarcasm

9:05 PM - 16 Sep 2015

Figure 1.3: A sarcastic tweet.

across applications. On the other hand, creating these assets from scratch is a laborious endeavor, often requiring extensive domain knowledge and subject matter expertise.

1.1.3 Data Analysis

The final step entails applying statistical analysis and data mining techniques (e.g. aggregation, smoothing, clustering) to distill structured information from the output of the aforementioned inference models. This information can then be used to build information visualization tools allowing for analysts to explore, interpret and generate new insights from the data. The question then becomes how to map the inferred signals into an objective measurement of interest, e.g. how can the opinions inferred from a set of tweets be mapped into a measure of vote intention? Regardless of the mapping of choice however, it might still be difficult to interpret the meaning or the implications of a single measurement aggregating many noisy signals, as it can hide much of the complexity and nuance of the subject matter. Time-series can be more informative by showing how a measurement evolves over time, but without context it can still be difficult to know *why* a certain change is observed. It can then be tempting to process the data in a way that matches prior expectations, which can distort the results. This can also lead social media analysts to only ask very narrow or pre-conceived questions about complex issues, rather than allowing the data to dictate the relevant questions, which can also introduce confirmation biases. In summary, current social media analysis approaches can capture general trends but often fail to provide adequate context to understand said trends and produce valid insights.

1.2 Research Goals and Contributions

Despite all the enthusiasm and research around social media analysis, in practice, building and deploying these systems remains a complex and costly endeavor. In part, due to the multidisciplinary nature of many applications, often involving subject matter experts, computer scientist and statisticians, but also because these systems are composed of several disparate components (e.g. data crawlers, NLP tools, data analysis packages). The development of each component usually involves non-trivial (and often subtle) decisions that impact the overall behavior of the system. However, the major bottleneck of this process is the development of custom NLP models.

1.2.1 Agile Social Media Analysis

Suppose we wanted to use social media to investigate the reactions to the results of the 2016 US Presidential election, which was marked by the unexpected victory of the republican nominee Donald Trump. Building a system for this purpose would require the development of a pipeline of NLP models tailored for this task. However, this analysis would only be relevant immediately after the fact, hence these models would need to be deployed quickly. These elections were highly controversial and involved a variety of unusual factors and circumstances. Therefore, to get a better understanding of the subject we could want to identify key topics or issues related this event (e.g. the interference of the Russian government) and then infer the attitudes about those issues. However, this would require the development of additional models to further filter and categorize the data. The tremendous manual efforts required to develop task-specific models can prevent analysts from conducting time-sensitive analyses on a timely manner. It can also restrict studies of complex or polarizing subjects to mere superficial analyses. A fair investigation into such subjects often requires taking various perspectives into account, but this requires multiple specialized models. Moreover, this constrains the application of SMA to a limited set of languages and domains. Unlocking the true potential of social media analysis requires agile methodologies to ease the efforts of building of specialized NLP models.

The main goal of this thesis is to develop methods to support the agile deployment of SMA systems for various applications and improve the quality of the analyses. First, by investigating methods to tackle the fundamental challenges of processing subjective user generated content (i.e., the noise, brevity and ambiguity) and build specialized models for this content in low-resource settings. Second, by leveraging the low-resource models to operationalize methodologies to deal with the inherent biases of social media data, thereby improving the validity of studies conducted over these data.

1.2.2 Agile Modeling with Low-Resource Neural Networks

To go beyond generic bag-of-words classifiers there are currently two main options: researching and developing highly specialized features, or adopting representation learning methods to automatically learn such features from data. *Deep learning* models gained notoriety in recent years due to the ability to learn internal *distributed* representations, thus freeing practitioners from the burden of manually designing task-specific features (Goodfellow et al., 2016). However, deep neural networks are more flexible and complex (in terms of free model parameters) and thus require larger training datasets to be optimized. Therefore, the manual efforts that could be saved in feature engineering must now be devoted to producing larger annotated datasets, as well as searching for the best architecture amongst a myriad of possible components and a huge configuration space.

To address this issue, this thesis introduces a method to train supervised neural networks in low-resource settings, by exploiting two key ideas:

1. **unsupervised neural embedding learning**: exploiting large amounts unlabeled data to train neural embedding representations for words and users. The resulting embeddings can be used as generic representations for various models and provide an effective way to complement small hand-crafted resources.

2. **low complexity neural architectures**: constructing *minimalist* neural architectures (i.e. with a single small hidden layer) that yield low-complexity models, but with enough flexibility to learn and exploit specialized representations for the task at hand. This helps to constrain the capacity of downstream models thus reducing the chances of overfitting to small training datasets.

Decoupling the process of learning generic representations and specialized ones allows for the same generic embeddings to be reused and adapted for various applications. Furthermore, this method makes no assumptions with regards to what the embeddings represent or how they were computed, thus providing leeway for different types of models to be built — e.g., word embeddings can be used to derive models that operate over words and sentences, and user embeddings can be leveraged to induce models that make predictions with respect to users. For the former, one of the many recently proposed word embedding models can be used (e.g. Mikolov et al. (2013b)'s Skip-Gram). For the latter, a novel neural language model is introduced to estimate user embeddings from posting histories, that capture latent personal aspects and soft notion of *homophily*. These user representations are then used to construct *Content and User Embeddings Neural Networks* — novel family of deep neural networks that combine representations of the content and of the author of a post, to support contextualized inferences over social media. These methods were transposed into an agile modeling framework to accelerate the development of SMA models by reducing manual efforts in feature engineering and data annotation, and allowing linguistic resources and components to be adapted and re-used.

1.2.3 Demographically Controlled Social Media Analysis

Social media analysts often seek to use social media to investigate human-related phenomena and extrapolate the findings and insights to the general public. However, the universe of social media users is not a representative sample of the population. Yet, current SMA methodologies assume that all the data samples that can be collected about a topic contribute equally to the analysis. Therefore, the findings gleaned from these methods may not reflect broader trends. Moreover, this can distort the interpretations of the outcomes — given the diversity of users and the plethora of use cases of these platforms, it often becomes difficult to understand the underlying causes of an observed measurement, and control for inherent biases or intentional manipulation efforts.

To address this problem, the aforementioned agile modeling framework is leveraged to operationalize a methodology to automatically collect, process and organize social media data, such that it can subsequently be used to sample demographically representative *digital cohorts*. Specifically, the framework is used to develop demographic inference models to infer attributes of social media users, such as the age, gender and race/ethnicity. The representative cohorts can then be used to conduct demographically controlled studies, thereby ameliorating both the selection and confirmation biases, and improving the validity of analyses concerning national trends. Furthermore, demographic information will allow analysts to ask more fine-grained questions, regarding not only the topics that are discussed on social media, but also how these vary across the population, thus fostering deeper insights into the subjects of study.

1.2.4 Main Contributions

The main novel contributions from this doctoral research can be summarized as follows:

1. **Low-resource Learning for Noisy NLP**: A method to transform generic embedding representations into task-specific ones, with scarce and noisy labeled datasets (Astudillo et al., 2015b). This approach allows us to construct low-resource neural networks for various applications and induce specialized models operating at different levels of granularity: document-level, word-level (Amir et al., 2016a) and user-level (Amir et al., 2017).

2. **Unsupervised User Embedding Models**: A neural language model to estimate vector representations for social media users, which can be used to inform downstream predictive models for various tasks, including mental-health inference and demographic inference (Amir et al., 2016c).

3. **Contextualized Social Media Analysis Models**: A family of neural network architectures for textual inferences that combine lexical representations of the contents of a post with representations of the author, which allows models to take contextual information into account. These models can better deal with ambiguous content (e.g. involving sarcasm) and support personalized inferences (Amir et al., 2016c).

4. **Shared Task Evaluations**: Participation in international shared task competitions including: various systems for SemEval Twitter Sentiment Analysis competition, all of which ranked amongst the top four submissions (Amir et al., 2014; Astudillo et al., 2015a; Amir et al., 2016b); the top ranking submissions for RepLab 2013 Online Reputation Monitoring shared task (Filgueiras and Amir, 2013) and SemEval 2015 Twitter Sentiment Lexicon Induction shared task (Amir et al., 2015).

5. **Social Media Analyses**: A methodology to sample digital cohorts of social media users to support demographically controlled studies, and a publicly available software toolkit with the models and linguistic resources developed in this thesis[2]. This software has already been used to build real-world SMA applications for computational journalists, political scientists, and epidemiologists.

[2] http://github.com/samiroid/ASMAT

1.3 Thesis Summary

The remainder of this dissertation is organized as follows.

Chapter 2

This work stands at the intersection of various fields of study, thus this chapter aims to acquaint the reader with the main concepts, ideas and challenges of this research. First, by introducing the previous work done on social media analysis with particular emphasis on applications related to public health and the social sciences. Then, it overviews the current approaches to build NLP models for inferences over subjective content, such as lexicon-based models and supervised machine learning approaches including linear and neural network models. Finally, it outlines the main activities involved in the development and deployment of SMA systems, and briefly describes some of the preliminary work that paved the way to this dissertation.

Chapter 3

This chapter develops a method to build specialized models with low-resource supervised neural networks. The main idea is to cast the problem of inducing specialized models as that of extracting tailored representations from generic pre-trained embeddings. To that end, the idea of *embedding subspace* is introduced as method to learn tailored representations with scarce and noisy labeled datasets. Then, a predictive model based on this concept is formalized — the Non-linear Subspace Embedding model (NLSE).

The model was first used to construct specialized document-level classifiers for a set of popular academic Twitter sentiment analysis and opinion mining datasets. The results showed that this method consistently outperforms simple *off-the-shelf* baselines and other more sophisticated neural models (which require larger training datasets). Then, this approach was validated in a case-study of real-world social media analyses for the social sciences. The experiments were conducted over a range of datasets collected and curated by computational social scientists at the *Centre for the Analysis of Social Media*[3] (CASM) to investigate a myriad of contemporary issues, from attitudes towards politicians to reaction to natural disasters, as expressed by Twitter users. Again, the results showed that leveraging specialized classifiers induced with the NLSE yields much better results, particularly for subjective tasks.

[3] 4.

Chapter 4

This chapter develops the low-resource modeling approach further — first, this method is leveraged to build specialized word-level models. These models can then be used to automatically expand small labeled lexicons and adapt them to social media environments. The evaluation was conducted by training models to predict the subjective ratings assigned by human judges to seven well-known lexicons, describing 14 aspects (e.g. sentiment polarity and happiness score). The results demonstrate that the NLSE significantly outperforms other baselines based on word embeddings. Second, it investigates the effect of the subspace projection on the representation space; and measures the low-resource learning ability of the NLSE.

The word-level models are then used to derive a large-scale sentiment lexicon for downstream social media analysis applications. To assess the impact of the such lexicons in practice, a set of experiments is conducted to compare the performance of lexicon-based Twitter sentiment analysis systems based on small and expanded lexicons. The results show that the latter can substantially improve the quality of lexicon-based models provided that words that do not bear sentiment (i.e. *neutral* words) are removed.

Chapter 5

This chapter extends low-resource neural networks to support the development of user-level and contextualized (document-level) models. To that end, it introduces the *User2Vec*, a neural language model to estimate user embeddings which capture latent user aspects and soft notion of *homophily*. The low-resource user-level models were evaluated over a mental-health inference task, aiming to discriminate between users affected by depression and post-traumatic stress disorder from matching controls, given their Twitter posts. The results show that these models offer state-of-the-art performance and allow for insightful analyses leveraging the user similarities captured by the embeddings. Then, the user embeddings are leveraged to construct *Content and User Embedding Neural Networks* which can be leveraged to derive contextualized models. The models combine lexical representations extracted with a set of neural Convolutional filters and user embedding representations, to support inferences over highly ambiguous and subjective content. The evaluation was conducted over a sarcasm detection task, and showed that these models can outperform a state-of-the-art sarcasm detection model that relies on a heavily engineered set of features

Chapter 6

This chapter deals with the practical aspect of deploying social media analysis campaigns, and builds upon the methods introduced in the previous chapters to improve over current methodologies. First, the aforementioned methods are transposed into an agile modeling framework to accelerate the development of document-level, word-level, user-level and contextualized models. Second, this framework is leveraged to operationalize a methodology to conduct demographically controlled analyses. To that end, a process is developed to automatically collect, process and organize social media data, such that it can subsequently be used to sample demographically representative digital cohorts of social media users.

These tools and methodologies are then validated in another case-study of real-world social media analysis, now with focus on public health and epidemiology applications. The case-study aims to assess the impact of demographic information in social media studies; and provide a proof concept of a SMA system built with the methods introduced in this thesis. The study consisted in the deployment of DEMOS, a *Digital Epidemiology and Mental-Health Observation System*, designed to track discussions around specific public health issues, and monitor the prevalence of mental illnesses over Twitter data. To validate the system, two pilot studies of demographically controlled public health surveillance over Twitter were conducted. The first, measuring how different demographic groups engage with different health related issues. These analyses reveal differences in the topics that are more relevant for different parts of the population (e.g young people discuss pregnancy more than contraception). The second, uses the cohort to investigate how mental illnesses affect different demographic groups — the results shows that mental-health issues also manifest differently across the population (e.g. depression and PTSD is more prevalent amongst women and racial minorities). This case-study demonstrates that leveraging digital cohorts allows for more insightful interpretations of social media analysis outcomes.

Chapter 7

This chapter, first presents the conclusions, main takeaways and limitations of this work. Then, it briefly highlights some of my ongoing research that did not make it into the thesis and points out relevant and promising avenues for future work.

Chapter 2

Background and Preliminaries

This thesis aims to improve the quality of data-driven social media analysis pipelines for various applications, and thus it stands at the intersection of several disciplines and research areas. This chapter aims to provide an overview of the relevant research topics, discuss previous work and highlight the current open challenges. Section 2.1 discusses the seminal works using social media as a lens into current trends and events, points out the main challenges of processing user generated content and the limitations of current social media analysis methodologies. Section 2.2 describes the main approaches to build NLP models for subjective language, reviews the literature on two subjective tasks — Twitter sentiment analysis and sarcasm detection — and briefly discusses methods to infer supervised machine learning models in low-resource settings. Section 2.3 discusses neural representation learning algorithms and introduces the main concepts of neural network models for NLP tasks. Section 2.4, addresses the methodological aspects of deploying social media analysis pipelines for real-world applications. First, by outlining the high-level process that governs the development of these systems and describe the typical methodology that is used to materialize this process. Second, by describing how this process was used to build a political opinion mining system for the web. Section 2.5 summarizes and concludes the chapter.

2.1 Predicting the Future (and the Present) with Social Media Analysis

With the rise of web-based social networks, researchers from various fields have been exploring the idea of using this data to tap into the 'wisdom of crowds' in hopes to get insights into current and future worldly events and circumstances. Much work has been done on the analysis of the structure of the networks themselves, e.g. to understand the properties and emerging characteristics of large real-world social networks (Mislove et al., 2007). The focus of this research, however, is on the use of NLP methods to analyze the contents generated by the members of such social networks.

2.1.1 Predicting the Future

The possibility of using social media the outcomes of future events is fascinating and the challenges it brings about has attracted significant attention from researchers and businesses. A classic example is trying to predict electoral results from signals gleaned from social media, such as the volume of messages about a candidate and the expressed sentiment (Tumasjan et al., 2010; Bermingham and Smeaton, 2011; Marchetti-Bowick and Chambers, 2012; Tjong Kim Sang and Bos, 2012). Others have used similar methods to predict movie box-office revenues (Asur et al., 2010) and anticipate stock-market movements (Bollen et al., 2011b; Zhang et al., 2011).

However, several researchers pointed out serious methodological flaws on previous works claiming to be able to predict electoral results with basis on Twitter data (Jungherr et al., 2012; Metaxas et al., 2011; Gayo Avello et al., 2011). For example, Jungherr et al. (2012) replicated Tumasjan et al. (2010) method on a comparable dataset and found that the predictions were very sensitive to experimental parameters, such as the period of analyses and inclusion or exclusion of smaller parties. Yet, in the original paper no adequate justification is given for these choices. By analyzing results from a number of different elections Metaxas et al. (2011) concluded that Twitter data is only slightly better than chance when predicting elections. It should be noted that chance is not a realistic baseline, since incumbency alone is a very strong predictor. Bermingham and Smeaton (2011) analysis of elections in Ireland also showed that the predictive power of Twitter was not competitive with traditional polling methods. Metaxas et al. (2011) suggests that methods claiming predictive power of elections on basis of Twitter data should meet at least the following requirements:

1. propose a clearly defined algorithm, e.g. formalize what constitutes a 'vote' (use all the users? drop users with few tweets? drop users with a lot of tweets?)

2. take into account the demographic differences between Twitter and the actual population

3. provide explainable predictions, i.e. black-box approaches ought to be avoided.

Stock market prediction and political forecasting have been widely studied and are known to be inherently difficult problems. These types of applications raise a host of additional problems having to do with complexity and uncertainty of these domains (this will be elaborated in Section 2.1.3). It does not seem reasonable that social media alone can directly and consistently predict such complicated and non-linear processes. Therefore, this research will be mostly concerned with applications using social media to 'predict the present', that is, as a lens to better understand current states of affairs.

2.1.2 Predicting the Present

Social media analysis has also been used to investigate specific events both in real-time and in retrospective, e.g. to monitor the public reactions to televised political debates (Shamma et al., 2009; Diakopoulos and Shamma, 2010), gage the opinions about social policies and reforms (Speriosu et al., 2011) and manage crisis-response efforts in real-time (Nagy and Stamberger, 2012; Mandel et al., 2012). Other works entail extracting continuous indicators from social media as a proxy for other real-world phenomena, e.g. Bollen et al. (2011a) found correlations between the public mood (as measured by an emotion lexicon), and various socio-economic trends; and O'Connor et al. (2010) correlated measures of social media sentiment with economic confidence and presidential approval rates estimated by traditional polls. Along these lines, Dodds et al. (2011) deployed the *Hedonometer*[1], a system that uses a *happiness* lexicon to compute daily measures of happiness from social media and align those measurements with relevant worldly events. Mitchell et al. (2013) used a similar methodology to connect happiness levels per geographic region (using geo-located tweets) with socio-demographic characteristics of those places.

There is also a growing body of work on using social media to investigate aspects related to personal health and well-being (McCaughey et al., 2014; Coppersmith et al., 2017). For example, it has been shown that signals from social media correlate with

[1] `http://hedonometer.org/index.html`

mental-health status, suggesting that these data could potentially be used to help prevent and diagnose mental illnesses such as depression (Hao et al., 2013; Schwartz et al., 2014; Coppersmith et al., 2014a), post-traumatic stress disorder (Coppersmith et al., 2014b), attention deficit disorder (Coppersmith et al., 2015a), and suicidal ideation (De Choudhury et al., 2016), among others. A related line of research has been looking at using social media tool to improve public health practices (Salathe et al., 2012; Dredze, 2012). For example, Paul and Dredze (2011) used Twitter data to track illnesses over time and location, measure behavioral risk factors and medication usage; and others sought to understand public perception about vaccines (Dredze et al., 2016b,a; Huang et al., 2017). These methods have also shown great promise for epidemics surveillance (Culotta, 2010; Broniatowski et al., 2013) and pandemics prediction (Ritterman et al., 2009). See J. Paul and Dredze (2017) for a thorough review of social media analysis applied to public health.

2.1.3 How *Not* to Make Predictions from Social Media

Using social media to build models that make predictions about the world is certainly an appealing idea. Such models must necessarily make simplifications about the phenomena they try to capture, however complicated problems should not be oversimplified. Gayo-Avello (2013) presented a comprehensive meta-analysis of electoral prediction methods based on Twitter data. In it, the author voices the same concerns discussed in Section 2.1.1 and presents core research challenges that need to be addressed to make accurate prediction of elections feasible. I argue that some of these concerns and challenges also apply to other domains and applications aiming to leverage social media analysis to investigate human-related issues. This section reiterates and comments on these challenges.

- **Demographic Biases**: Social media is not a representative and unbiased sample of the voting population — some strata are underrepresented while others are overrepresented (e.g. urban young adults). These biases should be acknowledged and predictions corrected on its basis. Similar strategies have already proven to be effective for online surveys, which can have comparable validity to other survey modalities simply by controlling for basic demographic features such as the location, age, ethnicity and gender (Duffy et al., 2005). While demographic information might not be available on social media platforms there is a growing body of work on methods to infer demographic traits from social media data (Cesare et al., 2017).

- **Self-selection Biases**: A minority of users are responsible for most of the political chatter and, thus, their opinions will drive what can be predicted from social media (Mustafaraj et al., 2011). Generally speaking, inferences drawn from social media will tend to overemphasize the concerns and views of the members that are most actively engaged. Therefore, SMA systems should ensure that data collection methods include samples from the long tail of less vocal users. This is particularly relevant for controversial issues, given that social networks often exacerbate homophily effects and foster 'echo chambers' of highly polarized views (Conover et al., 2011; Colleoni et al., 2014).

- **Credibility**: A substantial amount of social media data is not trustworthy and should be automatically discarded (Castillo et al., 2011). Recent research also points to the rise of sock puppet accounts to spread propaganda and misinformation, which already had a concrete impact in recent democratic processes (Bessi and Ferrara, 2016; Ferrara, 2017). Some work has been done on methods to measure data credibility, e.g. the *Truthy* system proposed by Ratkiewicz et al. (2011) to detect *astroturf* in political campaigns to discredit, or to simulate widespread support for a candidate.

- **Näive Models**: The use of simplistic sentiment analysis methods for complex and ambiguous data should be avoided. For example, political discourse is plagued with humor, double meanings and sarcasm, thus models should be robust enough to deal with this type of communication. Therefore, efforts should be devoted to methods for the particular case of sentiment analysis of political content and other such domains.

2.1.4 Social Media Analysis Methodologies

Conducting social media analyses always involves at least three steps: (i) data collection; (ii) inference over the data (iii) data analysis. In practice, studies are usually operationalized with one of the following methodologies: Manual Analysis, Automated Generic Analyses and Automated Tailored Analyses.

Manual Analysis

Manual analysis of the data by researchers, akin to the traditional research methods in the social sciences. This entails collecting a dataset about an issue and manually coding it with respect to the targets of analysis. For example, Carvalho et al. (2011) analyzed how users expressed opinions about a political debate in the comments section of an online newspaper; Tjong Kim Sang and Bos (2012) measured the sentiment expressed in tweets mentioning political candidates to the Dutch senate; and Ayers et al. (2017) used Twitter to investigate the reasons for the popularity of electronic cigarettes. This approach allows researchers to engage closely with the data and uncover underlying patterns, but it also suffers from the same limitations of traditional surveillance methods — i.e., it is restricted to relatively small datasets, it is laborious and time-consuming.

Automated Generic Analyses

Leveraging automated methods but with little or no manual inspection of the data and without tailoring the analysis to the specifics of the situation (Tumasjan et al., 2010; O'Connor et al., 2010; Mitchell et al., 2013). This involves applying some off-the-shelf method as a 'blackbox' (e.g. a general sentiment lexicon) to a dataset or data stream, and then aligning the results with observed real-world events. For example, counting the mentions to politicians in social media to predict election results intentions or correlating sentiment measures obtained with a generic sentiment lexicon with public opinion polls. These methods can often capture general trends but are also limited in that generic models and resources are unable to cope with the nuances of particular domains. Therefore, this methodology should not be used for detailed analyses over complex and nuanced data.

Automated Tailored Analyses

Approaches that employ automatic data analysis with tailored methods, e.g. bespoke machine learning models designed for the task (Bermingham and Smeaton, 2011; Marchetti-Bowick and Chambers, 2012). The output of the models is then aggregated and interpreted by subject matter experts to draw insights and conclusions about the observations. Developing the models involves manual data inspection and annotation efforts, and can also require additional research and development efforts to improve model performance. This is the process that involves the most manual efforts, but it is also the one that best exploits the full potential of social media as source of knowledge about the world. However, this methodology still faces various limitations that prevent the widespread adoption of SMA as a reliable instrument. This work aims to improve this methodology by addressing some of the most critical of those limitations.

2.1.5 Challenges of Processing Social Media

Social media analysis models (SMA) are usually operationalized with NLP methods and techniques. However, the characteristics of social media pose significant challenges to traditional methods that were developed for more conventional types of text. The main challenges of processing social media are the following:

- **noise**: social media tends to be *noisy*, due to the prevalence of web slang, typos and unconventional word spellings;
- **brevity**: posts are very short, thus providing little context to appreciate the meaning of some utterances;
- **ambiguity**: social media serves various purposes and is often used to express subjective views on different topics and issues. Subjective content is more ambiguous than factual assertions and thus more difficult to process automatically.

Therefore, lexicon-based models require more comprehensive lexicons to account for typos and alternative word spellings. Standard text classifiers also require larger vocabularies to keep track of all the lexical diversity, which generates very high-dimensional feature vectors. This in turn, yields more complex models (in terms of free parameters) which require more labeled data to be properly trained. Another pitfall of dealing with noisy text is the propensity to a large number of out-of-vocabulary words, i.e. tokens that were not seen during model training but do occur at inference time, which further degrades the performance. Moreover, bag-of-word models represent each word as a different symbol, hence they cannot learn that some similar words refer to the same concept (e.g. the words ‘coool’,‘c00l’ and ‘cool’).

Several approaches have been proposed to deal with token diversity by explicitly reducing the vocabulary size. These range from simple pre-processing heuristics such as clustering non-informative tokens into unique symbols (e.g. replacing URLs with the token `URL`) to more sophisticated approaches that try to normalize the text via stemming, spell-checkers (Han et al., 2013), dictionaries to expand abbreviations (Kouloumpis et al., 2011), or discarding the less discriminative tokens with feature selection techniques (Pak and Paroubek, 2010). However, these approaches are often too brittle and can end up compromising the model’s generalization, as some of the discarded words may appear at inference time — this is especially problematic given the brevity of social media posts. The brevity of the messages also implies that most words will not occur in a given post; as such, the resulting feature vectors tend to be rather sparse, thus carrying a weak signal. This causes additional problems to generic models, which need to be augmented with additional information about the content and context of a post usually via task-specific, manually crafted features. Finding and implementing the appropriate set of features involve a laborious and time-consuming process. Moreover, some of these features might depend on external tools (e.g. a POS tagger) which are may not available for most languages other than English. This makes it particularly difficult to deploy these systems for low-resource languages.

2.2 Modelling Subjective Language

One of the primary goals of this thesis is building computational models for the analysis of opinions, emotions and reactions expressed in social media. Emotions, opinions and sentiment are all related but distinct concepts — a *sentiment* is an attitude or thought prompted by a feeling, whereas an *opinion*, is a judgment or appraisal about something. *Emotions* are harder to define precisely because it is a complex human aspect that has been studied from multiple perspectives (e.g. neurological, psychological, sociological and philosophical). However, in everyday life we usually refer to emotions as our subjective thoughts and feelings (Liu, 2012). One common characteristic of these types of expression is that they are **subjective**, that is, *referring to internal states that are not open to observation or verification* (Pang and Lee, 2008).

Unlike objective expressions, i.e. factual assertions about the world, subjective language has an added degree of ambiguity which makes it more challenging to process automatically. Note however that these differences are often subtle, and not consensual amongst researchers and practitioners, e.g. objective statements can also imply sentiment if they refer to desirable or desirable facts. Nevertheless, for our purposes, we can abstract these nuances by treating all these problems as prediction tasks. More precisely, we are concerned with approaches to build models

$$\hat{y} = f_\theta(\phi(d)) \tag{2.1}$$

that given a document $d = \{w_1, \ldots, w_N\}$ composed of words w, predict a numeric response $\hat{y} \in \mathbb{R}$ or a category $\hat{y} \in \mathcal{Y}$, corresponding to some high-level aspect of the document. For example, a sentiment analysis model takes in d and tries to predict the sentiment polarity $\hat{y} \in \mathcal{Y} = \{\texttt{positive, neutral, negative}\}$ expressed in d. These models operate over formal representations of the input $\phi(d)$ and are characterized by a set of parameters

θ that encode some form of linguistic or domain knowledge that informs the mapping between inputs and outputs.

Language is a complex human phenomenon, in which sequences of discrete symbols (i.e., words) are used to communicate and reason about anything from concrete and tangible objects to abstract ideas. If we tried to explicitly encode all the ways in which words can be combined and influence each other to convey meaning, the resulting models would be cumbersome, and difficult to estimate and interpret. Therefore, in practice, our choices for $f_\theta(\cdot)$ will always involve simplifying assumptions to derive models that are useful and tractable. The main choices for building predictive models for subjective NLP tasks are either lexicon-based or supervised machine learning approaches. The former, are informed by labeled lexicons and operationalize the mapping via a set of manually specified rules and heuristics. The latter, requires a set of manually labeled training examples to learn a discriminative function that given a new document d predicts the corresponding label y. Note that both types of models are parametrized by θ which depends on some form of supervision. Indeed, recent work has shown that under certain conditions these approaches are equivalent, specifically, lexicon-based classifiers are a special case of Näive Bayes classifiers (Eisenstein, 2017). An important distinction is that lexicon-based models depend on word-level annotations, whereas supervised models require document-level annotations.

2.2.1 Lexicon-Based Models

Lexicons are dictionaries mapping pre-specified keywords to prior semantic annotations (e.g. sentiment). These dictionaries can be used to inform algorithms that extract specific signals from the text with basis on the presence of these keywords (usually disregarding the context). For example, one can infer the overall sentiment of a document by looking at the proportion of words from lists of *positive* and *negative* words (Pang et al., 2002). Assuming we have a lexicon $\mathcal{L} : w \in \mathcal{V} \to \mathbb{R} \in [-1;1]$ associating sentiment polarity scores to a set of words $\mathcal{V}$, we can make predictions by aggregating the *sentiment scores* of individual words and use a threshold t to decide on the output label. More formally, given a post $d = \{w_1, \ldots, w_n\}$ with n words, the overall sentiment polarity is inferred as:

$$\hat{y} = \begin{cases} \texttt{positive}, & \text{if sentiment}(d) > t \\ \texttt{negative}, & \text{if sentiment}(d) < t \end{cases} \tag{2.2}$$

$$\text{sentiment}(d) = \frac{1}{n} \sum_{w_i \in d} \mathcal{L}(w_i)$$

These models are very popular due to their simplicity, ease of implementation, and interpretability. For domain experts, creating these word lists is intuitive, and provides a quick way to build a reasonably accurate classifier. If the performance is unsatisfactory the models can be easily debugged and iteratively improved by refining the lexicons.

However, relying on the semantics of (a few) words devoid from context is obviously far from sufficient to capture the subtleties and nuances of subjective language. Consider the sentence: *is he a **good** candidate?* — despite containing a ‘positive’ word, it does not convey any sentiment. On the other hand, the intended meaning of a word strongly depends on the context, e.g. *scary* might be a positive aspect of a movie or negative trait of a person. Furthermore, words can have different connotations in different domains and communities (Silva et al., 2012; Hamilton et al., 2016) — e.g., the expression *he is so sick*, can be referring to an ill person or a great person[2]. This implies that specific analyses should be carried out with lexicons tailored for the task and domain of application.

2.2.1.1 Lexicon Induction

Another major challenge of building lexicon-based models is in obtaining high quality and comprehensive lexicons. These resources are usually created manually by linguists or subject matter experts who must, first, choose a collection of relevant terms; and second, provide judgments for the terms either themselves, or collect these judgments through annotation campaigns. This process can become rather expensive and time-consuming and thus manually crafted lexicons are usually incomplete. This is particularly problematic for applications over user generated content which often contains abbreviations, misspellings and other orthographic variations. Recently, researchers have begun exploring crowdsourcing approaches to create much larger lexicons at lower costs (Mohammad and Turney, 2013; Warriner et al., 2013) but these approaches can only go so far. For example, reliable crowdsourcing services are not available in every country, and it is not reasonable to ask annotators to assign labels to all spelling variations of a word. Furthermore, it might not be easy to crowdsource tasks that require highly specialized domain knowledge, thus these methods are more suited for general and intuitive properties. Therefore, in order to be effective, lexicon-based SMA methods require strategies to automatically induce large-scale lexicons at low-costs.

The automatic extraction of lexicons is a well-known and widely studied problem, with most proposed solutions being predicated on the idea that *similar* words should have *similar* labels. Proposed approaches differ essentially along two axes: first, in how

[2] `https://www.urbandictionary.com/define.php?term=sick`

the word similarities are defined and measured; and second, how the label assignment is operationalized, e.g. with heuristics (Kim and Hovy, 2006), supervised classifiers (Silva et al., 2012) or graph-based label propagation algorithms (Baccianella et al., 2010; Velikovich et al., 2010). Word similarity measures can be inferred from word relations, such synonymy and antonymy as described in a manually curated knowledge base (Hu and Liu, 2004; Rao and Ravichandran, 2009). However, these methods are limited in that such knowledge bases are not available for most languages, and do not usually encompass the informal language that is characteristic of social media. Alternatively, one can leverage a data-driven approach and exploit statistics from corpora analysis. Context is one strong indicator for word similarity, as related words tend to occur in similar contexts (Firth, 1961). Therefore, word similarities can be inferred via information theory metrics (Turney and Littman, 2003; Kiritchenko et al., 2014) and with methods based on distributional similarities induced with Latent Semantic Analysis (Bestgen and Vincze, 2012) or neural word embeddings (Tang et al., 2014).

2.2.1.2 Sentiment Lexicon Expansion with Word Embeddings

Along these lines, we also developed a system based on word embeddings to participate in SemEval Twitter sentiment lexicon induction shared task (Rosenthal et al., 2015). The goal was to assign sentiment polarity labels between 0 and 1 to a set of words and phrases collected from Twitter. Instead of computing word similarities directly, we used Ling et al. (2015a) Structured Skip-Gram embeddings as features for a regression model trained to predict the labels of a smaller lexicon (Amir et al., 2015). The model consisted of a Support Vectors Machine (Vapnik, 2000) with a Radial Basis Function kernel of the form $\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = e^{(-\gamma|\mathbf{x}_i - \mathbf{x}_j|^2)}$ with $\gamma > 0$, where $\mathbf{x}$ denotes a feature vector. Hence, it learns a linear function in the space induced by the kernel and the data, which corresponds to a non-linear function in the original space. Despite the simplicity, this approach produced the top-ranking submission of the competition.

2.2.2 Supervised Linear Classifiers

Lexicon-based models are often too simplistic and brittle to capture the nuances of subjective expression. A more powerful modelling approach are machine learning methods that make predictions based on learned correlations between instances and target responses. Generally speaking, this approach entails three steps: first, choosing a form for the predictive model $f_\theta(\cdot)$; second, defining a representation for the input $\phi(\cdot)$; and third, choosing a learning algorithm to estimate the parameters θ from training data. The simplest and most popular choice for $f_\theta(\cdot)$ are linear models, which assume a linear relationship between the observed instances (more precisely, a set of *features* describing those instances) and the labels. This relationship is characterized by a set of parameters $\theta = \{\mathbf{w} \in \mathbb{R}^k, b \in \mathbb{R}\}$, specifying the relative contribution of each feature to predict a response.

The model makes predictions with a linear combination of the observed features and parameters as

$$\hat{y} = f_\theta(\phi(d)) = \sigma(\mathbf{w} \cdot \mathbf{d} + b) \tag{2.3}$$

where $\mathbf{d} \in \mathbb{R}^k$ is a feature vector describing document d and $\sigma(\cdot)$ is a convenience function that transforms the model output into the desired response. This is an univariate model but it can be easily extended to multi-class classification by learning (conditionally) independent weight vectors $\mathbf{w}$ for each class, and constructing a weight matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times k}$ as

$$\mathbf{W} = \begin{bmatrix} — & \mathbf{w}_1 & — \\ & \vdots & \\ — & \mathbf{w}_{|\mathcal{Y}|} & — \end{bmatrix} \tag{2.4}$$

In this case, the model produces a vector of estimates as

$$\hat{\mathbf{y}} = \sigma(\mathbf{W} \cdot \mathbf{d} + \mathbf{b}) \tag{2.5}$$

where $\mathbf{b} \in \mathbb{R}^{|\mathcal{Y}|}$ is a bias vector. Then, if the goal is just classification, we can set $\sigma(\mathbf{x}) = \underset{i}{\operatorname{argmax}}\ \mathbf{x}_i$ to return the index of the class with the highest score. To interpret the univariate model predictions as a probability, the *sigmoid* function can be used to map the output to a real number in the range $[0;1]$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

The generalization of *sigmoid* to k-class classification is known as the *softmax* function

$$\operatorname{softmax}(\mathbf{x}) = \frac{e^{\mathbf{x}}}{\sum_i^{|\mathcal{Y}|} e^{\mathbf{x}_i}} \tag{2.7}$$

which induces a proper probability distribution over the outputs via exponentiation (ensuring non-negative numbers) and normalization (ensuring that all values sum to one). Thus, we can define a probabilistic classifier as

$$P_\theta(\hat{\mathcal{Y}}|d) = \operatorname{softmax}(\mathbf{W} \cdot \mathbf{d} + \mathbf{b}) \tag{2.8}$$

This is model usually referred as Logistic Regression or Maximum Entropy model.

The model can be trained with gradient based methods (e.g Gradient Descent) by finding the parameters θ that minimize the inverse likelihood of the training data

$$\theta \leftarrow \min_{\theta} - \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|} \log P_\theta(y^{(i)}|d^{(i)}) \tag{2.9}$$

where $P_\theta(y|d)$ is estimated with Eq. 2.8. However, this is general model which has been rediscovered in many fields and thus it gives rise to various popular classifiers. For example, the Näive Bayes classifier is also a linear model with but it follows a generative approach and thus it is trained with Maximum-Likelihood Estimation. The linear formulation of Support Vector Machines also corresponds to this model. The difference is that the training procedure imposes additional constraints to ensure that the classifiers learn the hyperplane that separates the classes with the largest possible margin, thereby improving the robustness of the model (Cortes and Vapnik, 1995). See Manning et al. (1999) for detailed explanations and derivations of these and other classical NLP models.

2.2.2.1 Representations

The most common choice to represent words in NLP is using the *one-hot* encoding — each word is represented as a vector with the size of the *vocabulary*, i.e. a dictionary mapping all the known words to numeric indices, with zeros in all the positions and the value 1 in the position corresponding to that word. However, this approach has two inherent limitations. First, one-hot word vectors are all orthogonal to each other, meaning that every word is treated as a different symbol unrelated to every other word. This means that the models cannot generalize to unseen words even if they are similar to other known words (e.g. *hotel* and *hotels* refer to the same concept but are represented with two distinct symbols). Second, the size of the vectors is proportional to the size of the vocabulary, implying that (for linear models) the complexity scales linearly with vocabulary size. This means that models operating over large vocabularies have a large number of parameters, thus requiring a large number of training examples be estimated without overfitting — this is a problem known as the curse of dimensionality.

Documents can be represented by summing the one-hot vector representations of individual words to obtain a vector $\phi(d) = \mathbf{d} \in \{0,1\}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the vocabulary. This is known as a *bag-of-words* model (BOW) and treats documents as unordered collections of words. Despite the success of BOW models in several text analysis applications (e.g. spam filtering and topic detection), completely disregarding word order information gives only a crude approximation of the contents of a document. Furthermore, in tasks such topic classification the presence of certain words is very indicative of the class e.g. a document about sports contain very specific words that are unlikely to appear in a document about, say politics. However, that is not case for inferences over subjective text, due to the inherent ambiguities and subtleties, e.g. 'positive words' can be used to convey negative judgments. Thus, many applications rely on additional *features* (i.e., characteristics) of the document that better capture its semantics. The process of manually designing features, usually referred as *feature engineering*, is highly dependent on the task and data at hand, and often requires significant domain expertise, ingenuity and experimentation.

2.2.3 Low Resource Learning

The success of supervised systems largely depends on the amount and quality of the available training data, oftentimes, even more than the particular choice of learning algorithm (Banko and Brill, 2001). However, creating high-quality manually labeled datasets for supervised systems is an expensive and time-consuming endeavor, which motivated the research in techniques to learn models in low-resource settings. These include complementing small datasets with other sources of supervision, such as: *transfer learning* methods to re-use labeled data from some other task/domain to build models for applications with scarce resources (Pan and Yang, 2010); and *semi-supervised learning* methods to combine large amounts of unlabeled data with few labeled instances to induce better models (Zhu, 2005). However, these methods assume that the out-of-domain data has some similarities with the target data, which is difficult to guarantee and not trivial to even ascertain. An alternative are *unsupervised feature learning* methods that exploit large volumes of unlabeled data to infer correlations, statistical regularities and other underlying patterns and learn high-level representations of the input (Ghahramani, 2004; Bengio et al., 2013). For example, one can learn word features by exploiting the fact that related words tend to occur in similar *contexts* (Firth, 1961). Approaches that are based on this concept include, Latent Semantic Analysis, where words are represented as rows in the low rank approximation of a term co-occurrence matrix (Deerwester et al., 1990) and word clusters obtained with hierarchical clustering algorithms based on Hidden Markov Models (Brown et al., 1992). In the next section, we discuss another approach based on this idea, consisting of learning continuous word vectors via neural language models.

For the extreme case where there are no labeled resources available, researchers have resorted to *distant supervision* to obtain large labeled datasets for 'free'. The idea is to leverage simple heuristics (e.g. the presence of specific words or meta-data attributes) to automatically assign *noisy* (i.e. imperfect) annotations to unlabeled data. Despite the low-quality of the annotations, this approach has been used to amass large labeled datasets for various tasks over Twitter data, e.g. sentiment analysis (Go et al., 2009), fine-grained emotion classification (Purver and Battersby, 2012), sarcasm detection (Bamman and Smith, 2015), and mental-health inference (Coppersmith et al., 2015b). Part of the research work presented in this thesis was conducted on datasets obtained in this fashion.

2.2.4 Applications

2.2.4.1 Twitter Sentiment Analysis

In recent years, the topic of sentiment analysis over user generated content has received significant attention from researchers, due to the wide range of applications. Indeed, Twitter sentiment analysis has become a research field in its own right and was popularized by international shared tasks such as SemEval (Nakov et al., 2013). As such, several variations of this task have been addressed in the literature (e.g. aspect-based sentiment analysis and opinion mining) — the main concepts and related tasks have already been thoroughly reviewed and summarized elsewhere (Pang and Lee, 2008; Liu, 2012, 2015). Hence, here we will briefly review supervised methods proposed for the general sentiment analysis task, as defined by SemEval, i.e. judging if a document expresses an overall positive, neutral or negative sentiment polarity.

Predicting the sentiment expressed in text is fundamentally a text classification task, thus the obvious choice are representations based on n-grams — Go et al. (2009) were the first to report on experiments with supervised classifiers for this task, using features based on unigrams and bigrams. A substantial amount of work has also been done in improving message representations with more sophisticated lexical features based on Part-of-Speech tags (Barbosa and Feng, 2010), disambiguated word senses (Miura et al., 2014), linguistic cues captured with morphological and syntactic patterns (Davidov et al., 2010a) and tree kernels (Agarwal et al., 2011). Kouloumpis et al. (2011) tried to take into account the informal style of social media with features based on the presence of hashtags, excessive punctuation, emoticons and word casing. Others have proposed methods that try to explicitly capture the nuance of sentiment classification, e.g. Paltoglou and Thelwall (2010) compared different feature weighting schemes (e.g.

binary, term frequency) and found that weighting schemes used in Information Retrieval to be more effective for sentiment analysis; and Kiritchenko et al. (2014) proposed a set of sophisticated features based on five sentiment lexicons. Most of the top performing systems that participated on the first editions of SemEval competition consisted of linear models based on the aforementioned features. The results of these competitions showed that simple BOW models perform very poorly, and that the most predictive features are the ones derived from sentiment lexicons, particularly when extracted from lexicons tailored for the social web.

Exploiting Unlabeled Data for Twitter Sentiment Analysis

In preliminary work of this thesis, we also developed Twitter sentiment analysis models to participate on international shared task competitions. The goal was to investigate methods to exploit unlabeled data and reduce the sparsity of the feature vectors. We experimented with compact document representations obtained with Concise Semantic Analysis (Li et al., 2011) and by mapping words into dense word vectors obtained with Brown word clusters (Brown et al., 1992) and Skip-Gram word embeddings (Mikolov et al., 2013a). The word clusters and embeddings provide a general way to include unlabeled data into the model. The lexical representations were then complemented with features extracted from multiple sentiment lexicon, as described by Kiritchenko et al. (2014). Some of the lexicons were automatically extracted and thus providing another mechanism to improve the model with unlabeled data. This approach allowed us to build a state-of-the-art system which ranked amongst the top submissions in two international shared tasks: RepLab Polarity for Reputation Classification task (Filgueiras and Amir, 2013) and SemEval's Twitter Sentiment Analysis (Amir et al., 2014).

2.2.4.2 Sarcasm Detection

Social media users often resort to humor and figurative language, such as irony and sarcasm, to express their views on complex issues (Carvalho et al., 2009). Like other figurative devices, irony and sarcasm are difficult to define precisely, but verbal irony is commonly defined as the rhetorical process of deliberately saying something, while trying to convey the opposite meaning of what is being said (Colston and Gibbs, 2007) — sarcasm can be viewed as a special case of irony, where the positive literal meaning is perceived as an indirect insult (Dews et al., 1995).

Early approaches to detect the use of figurative language in text used features similar to those used in sentiment analysis. Carvalho et al. (2009) analyzed comments posted by users on a Portuguese online newspaper and found that oral and gestural cues, such as emoticons, onomatopoeic expressions for laughter, heavy punctuation marks, quotation marks and positive interjections are indicative of irony. Others used features based on word and character n-grams, sentiment lexicons, surface patterns and textual markers (Davidov et al., 2010b; González-Ibáñez et al., 2011; Reyes et al., 2013; Lukin and Walker, 2013). Some work has also been done on features to capture specific patterns that are suggestive of ironic intent, such as the expression of contrasting sentiments in the same utterance, the presence of ambiguous words or co-occurrence of frequent and rare tokens (Riloff et al., 2013; Barbieri and Saggion, 2014).

Modelling Context

However, all the aforementioned approaches rely predominantly on features *intrinsic* to texts — and yet, lexical clues alone are often insufficient to discern ironic intent. Appreciating the *context* of utterances is critical for this; even for humans (Wallace et al., 2014). Therefore, recent work has been focused on augmenting lexical representations with features to capture contextual information, e.g. Wallace et al. (2015) exploited the expected biases of certain communities (e.g., people tend talk favorably about the politicians they support and unfavorably otherwise); Khattri et al. (2015) extended the notion of *contrasting sentiments* beyond the textual content at hand. In particular, they analyzed previous posts to estimate the author's prior sentiment towards specific *targets* (i.e., entities). A tweet is then predicted to be sarcastic if it expresses a sentiment about an entity that contradicts the author's (estimated) prior sentiment regarding the same. Bamman and Smith (2015) investigated a rich set of features to capture information about the interactions between author and the audience of a post; and Rajadesingan et al. (2015) operationalized theories of sarcasm expression from psychology and behavioral sciences to model a range of behavioral aspects of the author of a post (e.g. mood and writing style). The major downside of these and related approaches, are that they often rely on platform specific information and demand significant manual effort to collect additional data and derive the adequate feature sets.

2.3 Neural Networks for Natural Language Processing

As we have seen in the previous section, the process of designing features is very dependent on the task at hand and relies heavily on domain knowledge. Moreover, there is no principled or systematic way to discover such features and hence feature engineering is usually a trial and error process. Artificial neural networks can alleviate this process by incorporating the feature extraction process into the model. The model is then trained to map inputs into the desired outputs and also how to best represent those inputs for the prediction task. In other words, the feature extraction function $\phi(x)$ will also be learned from data, thus obviating the need to manually devise task-specific representations. Neural networks are thus defined as a composition of feature extraction functions and a prediction function — these functions are also known as *layers*. The two main types of neural models are: (i) *feedforward neural networks*, which take as input a fixed sized vector, perform a sequence of computations and outputs a single prediction vector (Rosenblatt, 1961; Rumelhart et al., 1988); and (ii) *recurrent neural networks* which operate sequentially over inputs of variable length (e.g. sentences). At each step, an internal representation is formed based on the previous and current inputs, and a local prediction can be made (Elman, 1990; Hochreiter and Schmidhuber, 1997).

2.3.1 The Multilayer Perceptron

The simplest possible feedforward neural network is the Perceptron (Rosenblatt, 1961), which only has an input and output layer, and thus it is equivalent to the LR model (Eq. 2.8). The Multilayer Perceptron (MLP) extends the Perceptron by introducing a set of *hidden* layers (i.e. feature extraction functions) in between the input and output layers

$$P(\hat{\mathcal{Y}}|x) = \text{softmax}(\mathbf{W} \cdot \phi_{\theta^n}(\ldots \phi_{\theta^1}(\ \phi_{\theta^0}(x)\ )\ldots) + \mathbf{b}) \tag{2.10}$$

By training the MLP for a specific prediction task, the hidden layers learn to transform the inputs into internal representations that are suited for that problem. These are known as *distributed representations* because they describe each concept with multiple *features* and each feature can be involved in describing multiple concepts (Hinton, 1986). In this case individual features have no concrete meaning, unlike *symbolic* representations (e.g. the *one-hot* encoding) which associate each feature to only one concept. The MLP defines the feature extraction function as an affine transformation followed by a pointwise non-linear function, such as the sigmoid (Eq. 2.6)

$$\phi_\theta(\mathbf{x}) = \sigma(\mathbf{H} \cdot \mathbf{x} + \mathbf{c}) \tag{2.11}$$

Here, $\mathbf{H} \in \mathbb{R}^{l \times j}$ and $\mathbf{c} \in \mathbb{R}^{l}$ are the weights and bias — note that this function is identical to the Logistic Regression model. The non-linear transformation is often called an *activation* function in the context of neural network models.

The MLP consists of a stack of LR layers, each of which takes the output of the previous layer as input, transforms it, and passes it to the next layer — this is known as *forward pass*. Another way to look at these models is as computational graphs (specifically, a Directed Acyclic Graphs) wherein each layer consists of a set of nodes, which are connected to the nodes in the next layer by a set of weighted edges, as depicted in Figure 2.1. The choice of size and number of hidden layers is referred to as the *architecture* of the network. In theory, given a large enough hidden layer, the MLP can approximate any function (Cover, 1965). However, there are no guarantees that the current learning algorithms are able to find the correct parameters in a reasonable amount of time. Indeed, the main challenge of building neural networks is how to efficiently estimate the parameters from data.

The MLP is as old as the Perceptron itself but it was not feasible for practical applications until Rumelhart et al. (1988) invented of the backpropagation algorithm. Rumelhart et al. (1988) showed that it was possible to minimize the global error of the network by minimizing the local errors of each layer, and proposed an algorithm to efficiently estimates the error gradients with respect to all the network parameters. The idea was to apply the chain rule of derivatives to the computational graph and leverage the fact the quantities that are used to compute the gradients of the upper layers, can be reused to compute the gradients on the lower layers — the errors are thus propagated backwards trough the graph. Even though the backpropagation algorithm made it possible for models to learn representations automatically, the hidden layers exponentially increase the number of model parameters which makes the models more prone to overfit and much slower to train. Moreover, the objective function that neural networks optimize is no longer convex, hence gradient based solvers are not guaranteed to find the optimal solution.

2.3.2 *Deep* Neural Networks

Neural networks with more than one hidden layer, are said to be *deep*. Researchers found that for the same number of parameters, better results could be obtained by favoring

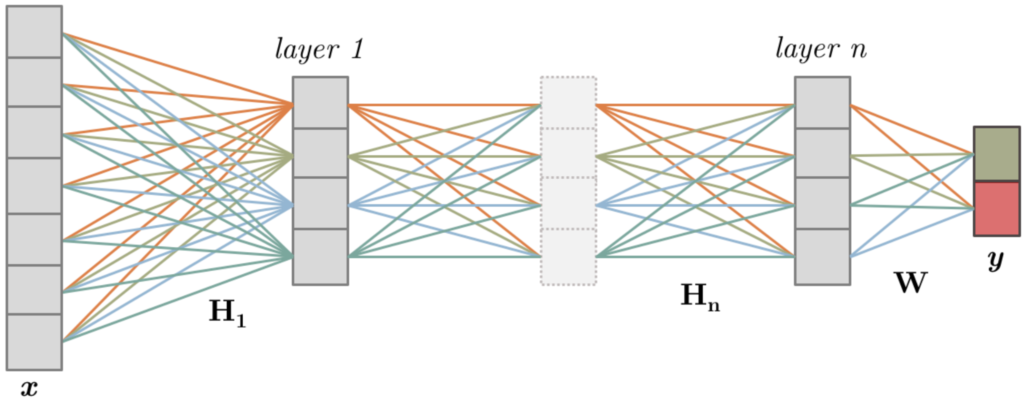


Figure 2.1: Schematic depiction of the Multilayer Perceptron. Feedforward neural networks process information through a sequence of layers that transform the input $\boldsymbol{x}$ in such way that a classifier can better predict the respective label $\boldsymbol{y}$. Each layer consists of a set of nodes (the 'neurons') that perform a non-linear operation (e.g. the sigmoid function). The nodes on a given layer are connected to those of the succeeding layer by a set of learned weights (the 'synapses') that are used to process the input with an affine transformation.

'tall' architectures — i.e. with multiple small hidden layers — over 'wide' architectures, i.e. with few large hidden layers. One of the reasons is that wide layers are so expressive that they can simply memorize all the mappings that were seen during training. This hampers the generalization of the models. In contrast, deep neural networks are able to learn hierarchical representations, such that a concept can be described at different levels of abstraction, each level building upon the representations from the previous levels. A good example is image classification: in this case the lower layers learn to detect basic patterns in an image, such as lines and edges; the layers above learn to detect patterns of lines and edges to form basic shapes; and the top layers learn to detect patterns of shapes to identify objects composed of those shapes (Lee et al., 2009). Moreover, researchers noted that even though the optimization procedure could get stuck in local minima, in practice the parameters that were learned were generally good solutions for the problem (Lecun et al., 1998).

Another advantage of deep networks is that they can be built incrementally with a technique known as *layerwise pre-training* (Hinton et al., 2006). We can start with shallow network with just one hidden layer ϕ_{θ^1} and train the network until it converges. Then, θ^1 can be used to set the parameters of the first layer of a 2-layer network; θ^1 is fixed and the model is trained again, thereby learning the parameters of the second layer ϕ_{θ^2}. This process can then be repeated to build a very deep neural network, while estimating only a smaller number of parameters at each step. More recently, Eldan and Shamir (2016) has formally shown that increasing depth adds exponentially more 'value' to the network than increasing width — e.g., there are limits to how well a 2-layer feedforward network can approximate functions that can be easily expressed with a 3-layer network.

2.3.2.1 Structured Architectures

The layers of an MLP are *fully connected*, meaning that the neurons on a given layer are connected to all the neurons on the succeeding layer (Figure 2.1). To reduce the capacity of the networks researchers began to explore custom architectures that exploited the characteristics of specific tasks or simplifying assumptions to reduce the number connections and to share some of the weights. For example, using symmetric weights for the input and output connections of a layer (Hinton and Salakhutdinov, 2006). Recurrent Neural Networks, exploit the temporal structure of sequential data to inform the architecture. In this case, the network uses a fixed sized feature extractor that slides through the input; at each step, the network extracts local features and computes an internal state that represents the sequence up to that point. This can also be viewed as a special case of a feedforward neural network where each layer has the same weights.

Another example are Convolutional Neural Networks (CNN), which were initially proposed for the task of optical character recognition by Lecun et al. (1998). One challenge of image recognition in general, is ensuring that models are invariant to shift, scale and distortion, e.g. a cat should be recognized whether it appears in the center of the image or in a corner. This usually requires non-trivial preprocessing and heavily engineered features to account for those variations. CNNs address this problem with three architectural ideas:

- *Local Receptive Fields*: Variables that are spatially and temporally nearby are highly correlated, e.g. the pixels that form or a line or the words that form a phrase. Therefore, the feature extractors (also known as *filters*) act on fixed size patches of the input and 'slide' across the data to extract local features, which form a feature map. These local features are then combined by the upper layers to detect higher-order features.

- *Subsampling*: Once a feature is detected only the relative position with respect to other features is important (not its exact location). Therefore, the aforementioned feature map is passed to a *sub-sampling layer* that performs local averaging and sub-sampling to reduce the resolution of the map, thereby reducing the sensitivity of the output to shifts and distortions.

- *Shared weights*: By training, each feature extractor becomes sensitive to specific patterns (e.g. edge detectors) but the same patterns can occur at different points in space, thus by re-using the same weights across the input the model ensures shift invariance. A target pattern will be detected regardless of its position on the input.

A convolutional layer is thus composed of a set of filters so that multiple features can be extracted from the same location. A CNN is created by stacking a set of these convolutional layers and sub-sampling layers and then adding a stack of fully connected layers on top. The former extract rich image representations and the latter further transform the representations and computes predictions. The state-of-the-art models for computer vision that achieve super human performance are based on these same principles (Krizhevsky et al., 2012; He et al., 2016). In recent years, CNN models have also showed excellent results for natural language processing tasks (Collobert et al., 2011; Kim, 2014; Kalchbrenner et al., 2014). Chapter 5 provides a formal definition of a CNN for text classification.

The most recent developments in deep learning were enabled in part by the availability of more powerful hardware and larger training datasets. But also due to a better understanding of the impact of the activation functions on learning (e.g. vanishing and exploding gradients), which lead to a host of new such functions; and the use of regularization techniques, both the classical techniques from statistical learning (e.g. norm based regularizers) to methods specific to neural networks such as Dropout (Srivastava et al., 2014). These factors allowed practitioners to train very deep neural networks without overfitting the data.

Deep neural networks are able to learn models for many non-trivial prediction tasks which, despite using little explicit human intervention, outperform well established baselines and in some cases even human themselves. These remarkable results may give the impression that these models just work out-of-the-box for any problem, but in practice that is not the case. Building neural models requires not only much more labeled data but also involves choosing amongst a myriad of possible neural architectures and a huge configuration space. This choice depends on factors such as the amount and quality of the available data; the task difficulty — e.g. predicting the topic of document is easier than inferring emotional tone or sarcastic intent; and the task complexity, as measured by the size of the response space (i.e. the number of labels). Broadly speaking, more difficult or complex tasks tend to require more complex architectures (i.e. with more parameters), which in turn require more training data.

2.3.3 Neural Language Models

One of the first successful large-scale applications of deep learning to NLP tasks was a probabilistic model for the task of language modeling. Language models aim to predict the probability of a sentence, i.e. the likelihood that a sequence of words forms a plausible sentence. These models are a core component for many downstream NLP tasks, e.g. knowing the probability of sentence can be used to reject ungrammatical sentences produced by machine translation or speech recognition systems. The main limitation of linear models used for this task was the difficulty to represent long sequences, due to sparsity and the aforementioned curse of dimensionality problem. Bengio et al. (2003) proposed to address this problem with a deep neural network trained to predict the occurrence of a word given a context of preceding words. The high-dimensional sparse vectors were first embedded in a lower dimensional dense space (i.e. the first hidden layer) allowing words to be represented as continuous vectors. N-grams could then be represented by aggregating individual word representations into a fixed sized vector, which was the input to a classifier that tried to predict the probability of the next word. Bengio et al. noted that, after training, the internal representations learned by the word embedding layer captured relevant latent aspects of words. Moreover, 'similar' words were associated to similar representations which allowed the model to generalize to variations of the same words or to related words that were not seen during training. Later work showed that these word vectors could be used as features to improve state-of-the-art models for various NLP tasks (Turian et al., 2010) and to initialize the parameters of downstream neural networks (Collobert et al., 2011). Even though this approach did not outperform the state-of-the-art model of at the time, this work was highly influential and spawned significant research on neural language models as a general way to learn word representations — also known as *word embeddings*.

2.3.4 Learning Word Embeddings

Harris (1954) hypothesis that the meaning of a word is a function of the contexts of other words where it occurs, has given rise to several successful computational linguistic models (e.g., Latent Semantic Analysis (Dumais et al., 1988) and Brown et al. (1992) word clustering algorithm). This is known as the *distributional hypothesis* and it is best known by the aphorism *you shall know a word by the company it keeps* (Firth, 1961). Neural language models are the latest operationalization of this idea and thus the word embeddings they produce are able to capture semantic and syntactic properties of words. The task of predicting the occurrence of words, given their contexts, requires models to have a basic understanding of syntax, e.g. to correctly predict the missing word $\boldsymbol{x}$ in the sentence *'you shall $\boldsymbol{x}$ a word by the company it keeps'*, the model needs to know that, in this context, $\boldsymbol{x}$ is probably a verb. By learning statistical correlations between words and the contexts where they occur, the models also capture latent semantics (in the distributional sense). From a machine learning standpoint, we have that word vectors that are used to predict the same words (i.e. that occur in the same contexts) will converge to similar solutions. As a result, similar words end up having similar representations which helps to improve model generalization.

The main problem of Bengio et al. (2003) neural model was that it was computationally expensive and thus could not be trained with very large datasets or vocabularies. Hence, various methods have been proposed to improve the scalability e.g. by approximating the training objective with simpler functions (Morin and Bengio, 2005; Mnih and Hinton, 2009); others have abandoned the goal of learning a language model, instead focusing on learning the word vectors directly (Collobert et al., 2011). This shift opened the door to more efficient, discriminative learning algorithms and simpler neural architectures, in turn allowing models to be trained with much larger datasets. Indeed, the state-of-the-art neural word embedding models do not have hidden layers essentially corresponding to linear models trained on word prediction tasks (Mikolov et al., 2013a; Pennington et al., 2014). Nevertheless, the general approach remains the same: associating words with parameter vectors, which are then optimized to predict other words that occur in the same contexts. The widely popular SKIP-GRAM model (Mikolov et al., 2013a), operationalizes this approach by sliding a *window* of a pre-specified size across an unlabeled corpus; at each step, the center word is used to predict the probability of one of the surrounding words as

$$P(y = d_c|d) \propto \exp(\mathbf{W}_c \cdot \mathbf{E} \cdot \mathbf{d}) \tag{2.12}$$

where $\mathbf{d}$ is a one-hot vector for word w, $\mathbf{E} \in \mathbb{R}^{e \times |\mathcal{V}|}$ correspond to the embedding matrix, transforming sparse word vectors into a dense real valued space of size e, and $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times e}$ is an output matrix projecting the dense representation to a vector with the size of the vocabulary. This approach was later expanded to learn representations for paragraphs (or, more generally, sequences of words), by trying to predict the center word of the sliding window, given the surrounding words and the paragraph (i.e., their respective vectors) (Le and Mikolov, 2014).

2.4 Building Social Media Analysis Pipelines

We now turn our attention to the methodological aspects of deploying social media analysis applications in the real world. Building these systems is essentially a software engineering endeavor wherein the goal is to produce a computational artifact to fulfill some business need or a set of requirements. As such, the specific methods and practices are largely influenced by factors such as the organization structure (e.g. small independent teams vs large hierarchical organizations), budget and expected outcomes (e.g., a bespoke pipeline for exploratory analyses or a monitor for longitudinal studies), among others. A key difference, however, is that typical software engineering projects aim to produce a piece of software, whereas SMA applications aim to produce new knowledge and insights; the developed software is just a means to attain that goal. Moreover, SMA projects are usually multidisciplinary, involving subject matter experts (SME), computer scientists and data scientists; and have a strong *R&D* component requiring experimentation and analysis throughout the process — even formulating the exact question to be addressed usually requires exploration and preliminary analyses (Wibberley et al., 2013). Nevertheless, there is a common process and a set of high-level activities that govern any such endeavor, namely:

1. Requirements Analysis
2. Data Collection
3. Data Annotation
4. Model Development
5. Data Analysis
6. Validation and Interpretation

These activities tend to have a sequential dependency, i.e. each step depends on deliverables from the previous activities and, at the same time, the feedback from each step can be used to further refine and improve the preceding activities. This naturally suggests a spiral development process, where each step is revisited and refined with basis on the evaluation of the previous version of the system or specific components thereof, as illustrated in Figure 2.2. It is worth noting that within this process there are often closed loops between adjacent activities, for example, the *requirements analysis* and *data collection* steps provide immediate feedback to each other, in that the latter is informed by keywords specified in the former; on the other hand, inspecting the data might show that the search terms are too broad or ambiguous. Similarly, the *model development* and *data annotation* steps are closely tied, as are the *data analysis*, and *validation and interpretation* steps. See tables 2.1 , 2.2 and 2.3 for further details on these activities including the main roles, goals, validation strategies, inputs and outputs.

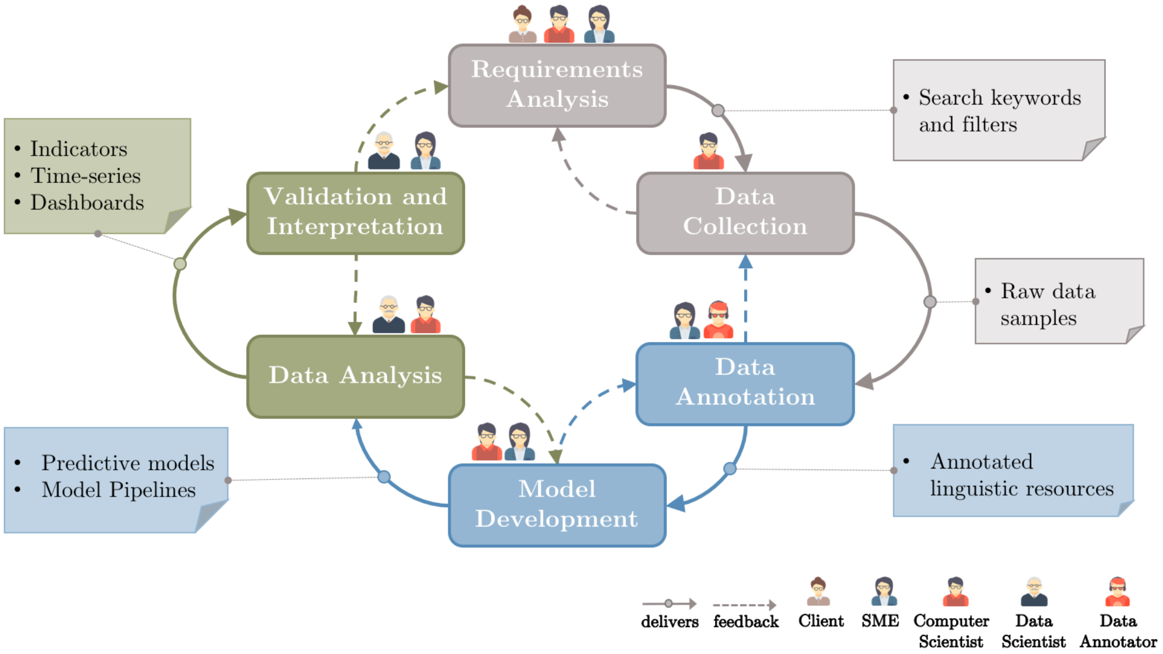


Figure 2.2: Diagram of the high-level process of building social media analysis pipelines

REQUIREMENTS ANALYSIS

Roles	Client, Subject Matter Expert and Computer Scientists
Goals	Formalize the business/information needs, questions or hypotheses driving the project, which must be scoped and translated into concrete pipeline components. A critical part of this step is specifying how should the system be validated. Even though each pipeline component will be subjected to some form of evaluation, it is crucial to have a criterion to validate the final output the whole system. The decisions and deliverables produced in this step directly inform all the subsequent activities, e.g. the relevant data sources, analytics and the intended outputs.
Output	Data sources, search terms and filters for the data collection, the labels or responses to be measured, the indicators to be produced, and the systems validation metrics.

DATA COLLECTION

Roles	Computer Scientist
Goals	Define methods for data retrieval (e.g. web scrappers, database clients or web API consumers) and storage, taking into account the goals and scale of the project, and factors such as the privacy and sensitivity of the data (e.g. anonymization might be required).
Validation	Manually exploring the data to get a better sense of the signals that are present. This step might reveal that the current search terms are too broad or too restrictive, thus prompting a revision of the requirements.
Input	Search keywords and filters
Output	Raw data samples.

Table 2.1: Requirements Analysis and Data Collection tasks

DATA ANNOTATION

Roles	Subject Matter Expert and Data Annotators
Goals	Produce annotations for words, phrases, documents or users that are representative of the task at hand. This entails defining guidelines and policies to ensure that all the data is annotated in a meaningful and consistent fashion. Depending on the complexity of the task or sensitivity of the data, the annotations might be obtained from SME or through data annotation campaigns conducted by annotator teams hired and trained for that purpose. These campaigns can be realized via interviews, crowdsourcing activities or gamification strategies.
Validation	The quality of the annotations can be ascertained by measuring the inter-annotator agreement. This step might also expose new topics or response categories that were not known upfront, thus motivating a revision of the data collection process or the project requirements.
Input	Raw data samples and data annotation guidelines
Output	A set of annotated linguistic resources to inform the development of predictive models.

MODEL DEVELOPMENT

Roles	Computer Scientist and Subject Matter Expert
Goals	Develop NLP methods to infer relevant signals from the data. This usually involves establishing simple baselines to ascertain that the signals of interest can be measured with the current technologies. Depending on the performance of the baselines, this step might also involve *R&D* activities to develop novel models, techniques and linguistic resources — e.g. leveraging domain knowledge to design task-specific features for machine learning models.
Validation	Models predictive performance on a held-out dataset. It might also become apparent that the annotated resources are insufficient or inappropriate for the intended goals or performance requirements, thus prompting a new data annotation campaign.
Input	Labeled data
Output	One or more predictive models, which can be composed into 'chains' or 'cascades' to e.g. filter and decompose data streams or operationalize complex decision processes.

Table 2.2: Data Annotation and Model Development tasks

DATA ANALYSIS

Roles	Data Scientist and Computer Scientist
Goals	Distill structured information from the models output via statistical analysis and data mining techniques (e.g. aggregation, smoothing, clustering).
Validation	The distilled information can then be used to build information visualization tools allowing for analysts to explore, interpret and generate new insights from the data. This step has an influence on the choice of predictive models, particularly in determining how ‘interpretable’ should the models be[3] — in some cases, it might be preferable to have a less accurate but more interpretable model.
Input	Raw signals inferred by the predictive models
Output	Indicators, time-series and other relevant statistics.

VALIDATION AND INTERPRETATION

Roles	Data Scientist and Subject Matter Expert
Goals	Validate the output of the system and interpret the outcomes of the analyses. This might involve comparing the output of the system with gold standard data, testing hypotheses or framing the interpretation of the results in light of some theory. These findings and evaluation results can also provide feedback to further refine the requirements and the data analysis step.
Input	Indicators, time-series and other statistics
Output	New insights about the topic of study, which can be used to inform policies, actionable decisions or be disseminated to specific audiences of the general public.

Table 2.3: Data Analysis, Validation and Interpretation tasks

2.4.1 *POPMine*: A Political Opinion Mining for the Web

After describing the high-level process of building SMA systems, we now look at a real-world application. The preliminary phase of this research involved the development of several such systems for computational journalists and political scientists. This section, describes the development of *POPMine*, a political opinion mining system to track the popularity and attitudes expressed about the main Portuguese politicians in conventional and social media, over time (Saleiro et al., 2015). The project was carried out by a development team of computer scientists and computational linguists; and an analysis team of political scientists and economists. The former were responsible for building and deploying the system and the latter were responsible for the analysis and interpretation of the results. The system was deployed as a pipeline, depicted in Figure 2.3, to operationalize the following tasks:

1. collect texts from web-based conventional media and social media;

2. process those texts to recognize mentions to political entities and the expressed opinion polarities;

3. generate longitudinal indicators of both frequency of mention and sentiment across media sources.

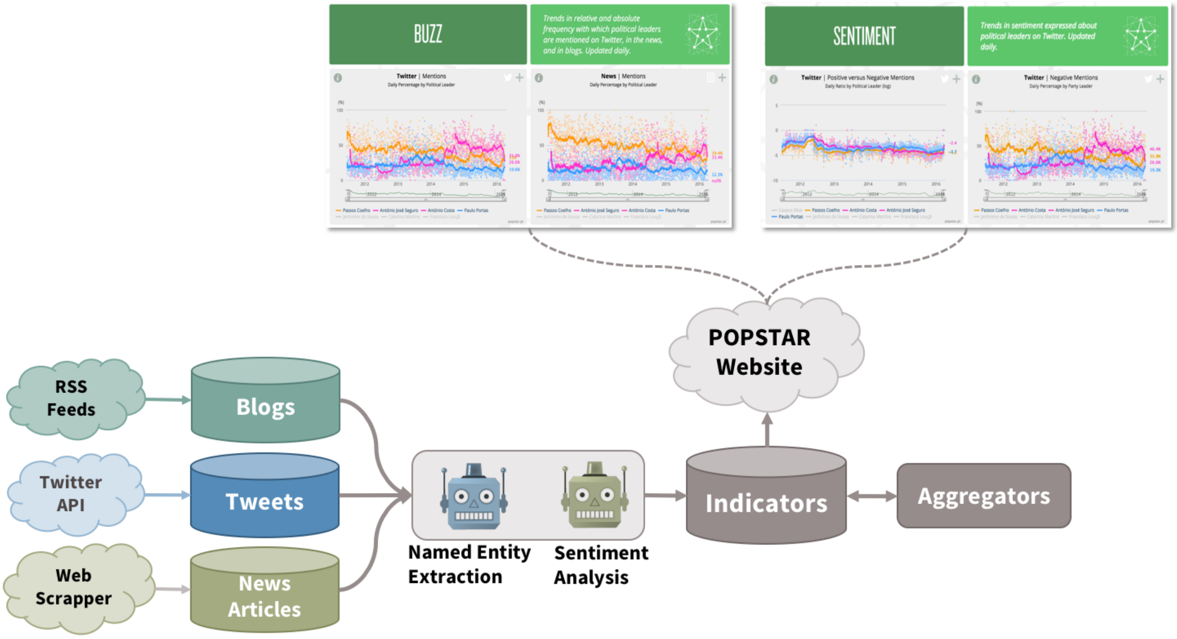

BUZZ
Trends in relative and absolute frequency with which political leaders are mentioned on Twitter, in the news, and in blogs. Updated daily.
Twitter | Mentions
Daily Percentage by Political Leader
News | Mentions
Daily Percentage by Political Leader
SENTIMENT
Trends in sentiment expressed about political leaders on Twitter. Updated daily.
Twitter | Positive versus Negative Mentions
Daily Ratio by Political Leader (log)
Twitter | Negative Mentions
Daily Percentage by Party Leader
Passos Coelho
António José Seguro
António Costa
Paulo Portas
POPSTAR Website
RSS Feeds
Twitter API
Web Scrapper
Blogs
Tweets
News Articles
Named Entity Extraction
Sentiment Analysis
Indicators
Aggregators

Figure 2.3: Diagram of the POPMine system

2.4.1.1 Data Collection

The first step involved setting up data collectors to retrieve and filter the relevant content — we collected content from blogs (via RSS feeds), digital newspapers (via web scrappers), and tweets from Portuguese users with Bošnjak et al. (2012)'s Twitter crawler. Politicians are often referenced in the media by different surface names, current office titles (e.g. *the prime minister*) and nicknames[4]. Therefore, we manually compiled and curated lists of alternative names used to mention politicians from public knowledge bases (e.g. the POWER ontology (Moreira et al., 2011) and the *Sapo Verbetes* API[5]) to extract search terms and help with entity recognition and disambiguation. The goal was to measure opinions expressed by ordinary social media users, hence we also manually compiled a black list of accounts belonging to politics experts or news agencies, and created filters to discard posts created by those accounts.

2.4.1.2 Model Development

The analytics pipeline consisted of two predictive models: a named entity disambiguator and a sentiment classifier. The former worked as a relevancy filter that given a message and the detected entity, predicts if the message is related to that entity — some politicians have very common names so a mention might be referring to another person or to something else entirely[6]. The latter was used to extract opinion polarities. The models were deployed as a 'chain' , i.e. the first model acts as filter and discards irrelevant messages; the rest are passed to the sentiment classifier.

This was the most challenging and laborious part of the process. It took months of dedicated effort by two teams devoted to research and develop effective models for those tasks. Both models attained state-of-the-art results in international shared tasks: see (Saleiro et al., 2013) and (Filgueiras and Amir, 2013; Amir et al., 2014) for details on the entity extraction and sentiment analysis models. However, we were only able to deploy stripped down versions of these models, as some of the features depended on tools that were not available in Portuguese. Moreover, the labeled datasets that were produced were significantly smaller than the ones used on the *R&D* phase (around 1.5k examples). The sentiment analysis dataset was also highly skewed containing only 10% of positive

[4] e.g. former Portuguese prime minister José Sócrates was often referred as "Sócras" in the social media

[5] `https://store.services.sapo.pt/en/cat/catalog/other/free-api-information-retrieval-verbetes`

[6] e.g. Portuguese politician Paulo Portas — 'portas' means 'doors' in Portuguese which was the source of a lot of false positives

examples[7] and contained a large number of repeated messages (e.g. re-tweets) which made it difficult to train the classifier — subsampling was not option here given the already reduced size of the dataset.

2.4.1.3 Data Analysis

We created indicators of sentiment and *buzz* (i.e. the frequency with which political leaders are mentioned), over time. These indicators were then aligned with polls and surveys collected from professional pollsters and published via a rest API. However, the social media indicators are much noisier, i.e. they have more variance, than conventional polls and thus we used a Kalman Filter to estimate smoother trends from these measurements. Unsurprisingly, most of the messages were classified as either neutral or negative which prompted us to create different indicators focusing on volume and share of negative messages, per politician. The API was then used to feed a near-real time visualization tool for the political scientists and economists to interpret the measurements in light of what was happening at the moment[8].

2.4.2 Discussion

It is not hard to see that this process is rather laborious, expensive, and time-consuming and thus, completely unsuited for time-critical applications (e.g. crisis management or trauma response). The most demanding steps are the model development and data annotation, as they involve coordination between different teams and experts and can take days to months to complete. Furthermore, building custom models require specialized resources which are difficult obtain and re-use. It was frustrating to use this approach to build tools for journalists to analyze unexpected newsworthy events, because it took days to: build the systems, assess and interpret the resulting data, write the stories and go through the editorial process — by the time the article was complete it was not news anymore. In the case of POPMine there was a limited budget to conduct the annotation campaigns but in the case of time-sensitive applications there might not be enough time to conduct extensive annotations tasks. Overcoming these obstacles required investigating approaches to induce better models (e.g. with feature engineering) in low-resource settings (e.g. exploiting unlabeled data) and experimenting with other publicly available datasets.

[7] Bermingham and Smeaton (2011) also faced the same challenge when collecting a political opinion mining dataset. Only 12% of the messages were labeled as positive.

[8] `www.popstar.pt`

Another challenge was related to the data collection. We found that a great part of the Twitter content related to Portuguese politicians was generated by accounts associated to media outlets and professional analysts. In fact, outside the elections period most discussion about politics was in reaction to some news related event. This raised several questions, such as: if a positive opinion about a news story involving a politician should be considered as support for the politician; should a retweet of a post expressing a negative opinion or with a link to a negative story, be considered as another negative opinion? If so, popular tweets (or mentioning popular news) would be included multiple times, which can distort the results. Other similar questions that do not have a clear-cut answer also appeared throughout the project. The informality and diversity of social media may make even tasks as simple as collecting 'relevant data' illusive, therefore it can be difficult to know exactly *what* is being measured by SMA systems. These issues were the motivation for the research presented in this dissertation.

2.5 Summary

This chapter introduced the main concepts and previous works related to social media analysis. The main methodologies used to build these systems and their limitations were discussed and some of the applications developed for the social sciences and public health domain were briefly described. We highlighted the main challenges of processing user generated content and the common approaches to build predictive models for subjective data. Then, previous work on sentiment analysis and sarcasm detection tasks was reviewed. The increased ambiguity of this type of communications requires more sophisticated models based on heavily engineered features, and larger training datasets. Low-resource learning methods to cope with scarcity of labeled data were also briefly discussed — neural language models in particular, present an effective mechanism to incorporate unlabeled data into pre-existing supervised models, and will be further explored throughout this research. The main concepts and basic neural network models were introduced here and will be also be expanded in the rest of the thesis, since one the main goals of this research is finding methods reduce the costs of building neural networks to improve the development SMA models. Finally, we discussed the high-level process and the main tasks of building SMA systems for real-world applications and described how this process was materialized in the development of political opinion mining system for the web.

Chapter 3

Low-resource Neural Networks for Noisy Text Analysis

Generic BOW models are often insufficient for inferences over subjective social media. To go beyond these methods there are currently two main options: either enriching these models with highly specialized features, or adopting deep neural networks to automatically learn sophisticated models from data. If the first requires extensive manual efforts in feature engineering, the second also requires significant efforts in producing larger annotated datasets, as well as finding the best architecture amongst a myriad of possible components and a huge configuration space. This chapter investigates approaches to reduce the gap between the manual efforts that go into model development and the quality of the resulting model. A method is proposed to build supervised neural networks in low-resource settings, which can be used to reduce the efforts of deploying of specialized models for noisy and subjective content (Section 3.1). This method is then leveraged to derive specialized document-level classifiers for SMA applications. The classifiers were first evaluated over a collection of Twitter sentiment analysis and opinion mining datasets (Section 3.2). Then, a case-study of real-world social media analysis applied to the social sciences is conducted, leveraging a range of datasets collected by computational social scientists. The study consisted in building bespoke classifiers to investigate a myriad of contemporary issues, from attitudes towards politicians to reaction to natural disasters, as expressed by Twitter users (Section 3.3).

3.1 Low-Resource Neural Networks

Deep neural networks are able to learn specialized models for many non-trivial prediction tasks which, despite using little explicit human intervention, outperform well established baselines and in some cases even human themselves (Goodfellow et al., 2016). However, taking full advantage of these methods requires access to large manually labeled training datasets or strategies to automate the creation of such datasets, e.g. synthesizing training data (Jaderberg et al., 2014) and self-training (Silver et al., 2017). Therefore, these methods excel in problems that require a single model for a specific and well-defined task, but become less practical for subjective problems or when multiple 'small' models are required for exploratory studies or multifaceted analyses.

Another well-known technique used to overcome the lack of labeled data is to integrate multiple sources of supervision by *pre-training* some of the layers with large amounts of 'cheap' data (e.g. unlabeled, distantly labeled, or out-of-domain data) and then *fine-tuning* the whole network with the training error gradients from the task-specific labeled data. However, directly updating pre-trained layers might still require estimating a large number of parameters and thus fail when only small training datasets are available. This problem is only exacerbated with noisy data, such as social media, where singletons and out-of-vocabulary words (OOV) are frequent — these words (i.e. their representations) will receive very few updates and hence be poorly adapted for the intended task; even worse, words not seen during training will never be updated. To address this problem, we developed a novel approach to build task-specific NLP models with scarce and noisy training datasets.

Assuming that the pre-trained embeddings induced by neural language models capture meaningful information, the problem of inducing specialized models can be cast as that of extracting tailored representations from generic embeddings. We follow the intuition that for any given prediction problem, only a subset of the latent aspects captured in the representation space is relevant. Therefore, instead of directly updating the embedding layer, we can use the labeled data to learn lower-dimensional adapted representations that capture task-specific information. In turn, leveraging smaller representations helps to reduce the complexity of downstream models thereby reducing the chances of overfitting to small training datasets. Hence, we can derive low-resource neural networks in two steps:

1. exploiting large unlabeled corpora to train neural embedding representations;

2. constructing *minimalist* neural architectures, i.e. with a single small hidden layer.

> The hidden layer acts as an adaptation mechanism that translates generic concepts into domain specific ones;

This results in low-capacity models with enough flexibility to learn and exploit specialized representations for the task at hand, thus eliminating the need for manually designed features. Moreover, decoupling the process of learning generic and specialized representations allows the same embeddings to be reused and adapted for different applications with a small amount of task-specific labeled data, as depicted in Figure 3.1.

3.1.1 Tailored Representations with Embedding Subspaces

Neural embeddings are learned *distributed representations*, i.e. they describe each concept with multiple *features* (vector dimensions can be interpreted as abstract features) and each feature can be involved in describing multiple concepts[1]. By training the embeddings in a prediction task, certain groups of features learn to detect and encode specific aspects of the data (Hinton, 1986). However, if we knew which set of features describes the most relevant properties for a specific task (e.g. sentiment analysis) we could extract compact representations tailored for that task, thereby eliminating noise and irrelevant aspects of the data.

Let $\mathbf{E} \in \mathbb{R}^{e \times |\mathcal{X}|}$ be a pre-trained embedding matrix where columns represent instances from a fixed sized set $\mathcal{X}$ (e.g. words) and e is the number of latent dimensions. Assuming that for any given prediction task, there exists a lower-dimensional *embedding subspace* that encodes the most meaningful aspects for that problem, then we can use labeled data to find a projection into that space. Therefore, we propose to learn task-specific embeddings, by factorizing the input as $\mathbf{H} \cdot \mathbf{E}$, where $\mathbf{H} \in \mathbb{R}^{s \times e}$ is a projection matrix with $s \ll e$. In other words, we keep $\mathbf{E}$ fixed and determine the optimal linear projection of the original embedding matrix into a low-dimensional subspace of dimension s, with respect to the prediction target. Hence, the transformation corresponds to adapting generic representations into *task-specific* ones. The intuition is that by aggressively reducing the representation space, the model is forced to learn only the most discriminative aspects of the input, while preserving the rich information from the original embeddings[2].

[1] In contrast, *symbolic* representations associate each feature to only one concept — e.g. the *one-hot* encoding, represents word i as a zero vector with the value 1 on the i-th dimension.

[2] This is also related to the idea of learning representations with information bottlenecks, e.g with auto-encoders (Hinton and Salakhutdinov, 2006)

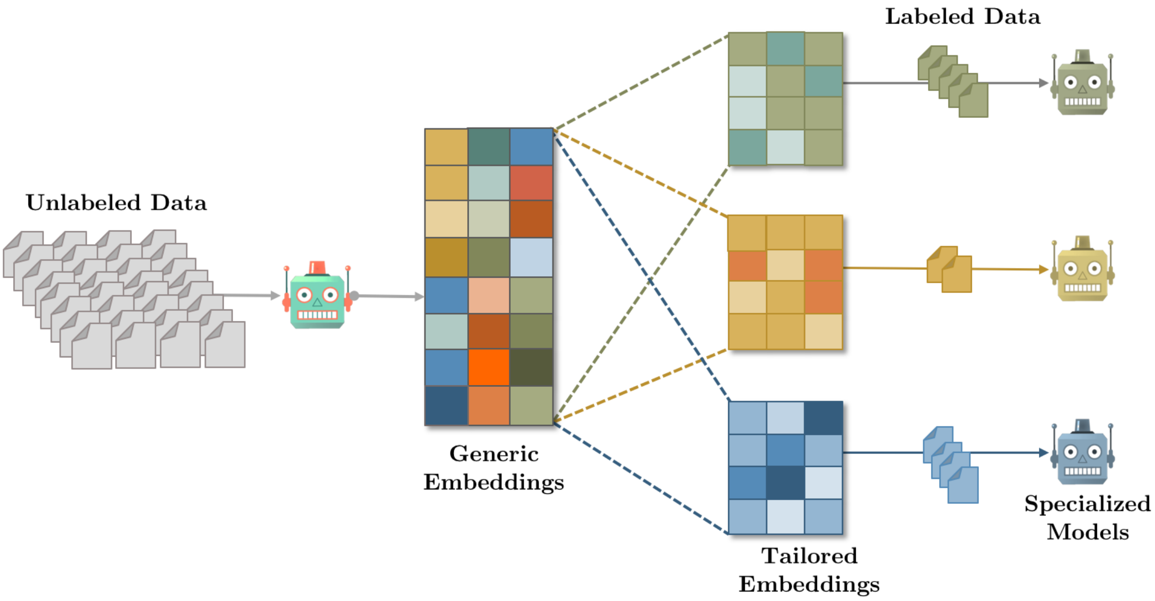


Figure 3.1: Low-resource Neural Networks for Social Media Analysis

3.1.2 Non-Linear Subspace Embedding Model

The concept of embedding subspace can be applied to deep neural architectures or to simpler models. Our goal here is to derive a low-capacity model, thus we now describe the simplest architecture based on this technique, henceforth denoted as the Non-linear Subspace Embedding model (NLSE). Let $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$ be a labeled dataset where $\mathbf{x} \in \mathcal{X}$ denotes an instance represented in *one-hot* form and $y \in \mathbb{R}$ is the respective label, the model induces a function $f_\theta(\cdot)$ parameterized by $\theta = \{\mathbf{w}, b, \mathbf{H}\}$ that given an instance $\mathbf{x}$, estimates a label $\hat{y}$ as

$$f_\theta(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$$
$$\phi(\mathbf{x}) = \sigma(\mathbf{H} \cdot \mathbf{E} \cdot \mathbf{x}) \tag{3.1}$$

where $\mathbf{w} \in \mathbb{R}^{1 \times s}$ and b refer to the predictor weights and $\sigma(\cdot)$ denotes an element-wise non-linear function (e.g. sigmoid), which allows the model to learn the intermediate representations. The model has only two hyperparameters — the size of the subspace s and the form of σ. Notice that at inference time $f(\mathbf{x})$ reduces to a linear predictor[3] with *embedding subspace* features $\phi(\mathbf{x})$.

[3] Note that without the non-linearity the would model essentially be learning a single projection

The NLSE can also be interpreted as a *Multi-layer Perceptron* with one single hidden layer (Rumelhart et al., 1988), but it differs in two key aspects: first, the input layer is factorized into two components: the generic input embeddings **E**, and the estimated subspace projection matrix **H**; second, the size of the subspace is much smaller than that of the original embeddings, with typical reductions above one order of magnitude. Similarly to other neural networks, the parameters θ can be estimated with gradient-based optimization methods and the *backpropagation* algorithm to compute loss gradients with respect to all the parameters.

3.2 Document-level Models

This section leverages the NLSE approach to build specialized document-level classifiers. A known limitation of feedforward neural networks is that the inputs must be represented as single fixed sized vectors. A question that arises when building word embedding based models for textual inferences, is how to transform individual word representations into a single document representation. The simplest approach is to use a bag-of-words assumption and just sum up the individual vectors. For example, the MLP projects *one-hot* document vectors into the input layer as linear combination of the individual word embeddings. However, squashing the vectors implies a loss of resolution and information. Other, more sophisticated, compositional models have been proposed to address this issue, e.g. Convolutional Neural Networks extract (task-specific) local features which are then aggregated to represent the entire document (Collobert et al., 2011); and Recurrent Neural Networks process variable length inputs sequentially, and at each step, form an internal representation that depends on the current and previous words (Hochreiter and Schmidhuber, 1997). However, these models require a lot of training data and the latter are notoriously harder to optimize (Pascanu et al., 2013).

Here, we explore a variation of the bag-of-words approach, which allows the model to preserve the embedding information as much as possible. Instead of squashing the vectors at the input layer, these are first processed in a forward pass through the network and aggregated *after* the output layer. This means the model makes individual word-level predictions with respect to the document label and then aggregates those predictions, as depicted in Figure 3.2. More formally, given a document $d = \{w_1, \ldots, w_n\}$ composed of n words drawn from a predetermined vocabulary $\mathcal{V}$, and a word embedding matrix **E** where each column represents a word $w \in \mathcal{V}$, we represent document d as a *one-hot* matrix

$$\mathbf{D} = \begin{bmatrix} — & \mathbf{x}_1 & — \\ & \vdots & \\ — & \mathbf{x}_n & — \end{bmatrix} \tag{3.2}$$

where each row $\mathbf{x}_i$ denotes a word in *one-hot* form. Then, we predict a probability distribution over the labels of a document as

$$\begin{aligned} P(\hat{\mathcal{Y}}|d) &\propto (\mathbf{W} \cdot \phi(\mathbf{D})) \cdot \mathbf{1} \\ \phi(\mathbf{D}) &= \sigma(\mathbf{H} \cdot \mathbf{E} \cdot \mathbf{D}) \end{aligned} \tag{3.3}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times s}$ is a matrix of classifier weights, $\mathbf{H} \in \mathbb{R}^{s \times e}$ is the subspace projection matrix and $\mathbf{1} \in 1^{n \times 1}$ is a matrix of ones that sums the un-normalized probability scores for all words. Note that, once learned, the subspace projection can be used to induce tailored representations even for words that were not seen in the training data (i.e. out-of-vocabulary words), thereby improving models' generalization and robustness to noisy data.

3.2.1 Evaluation: Sentiment Analysis and Opinion Mining

To evaluate our approach, we developed sentiment analysis and opinion mining classifiers in a low-resource setting, i.e. with fairly small datasets and no manually devised features. We used the datasets described in Table 3.1 wherein each document is classified as positive, negative or neutral. We compared the performance of our model with that of standard off-the-shelf classifiers and other representation learning methods:

- **NB**: Näive Bayes;
- **BOW**: linear models with bag-of-word features;
- **BOE**: linear models with *bag-of-embeddings* features, i.e. a document x is represented as $\phi(x) = (\mathbf{E} \cdot \mathbf{x}) \cdot \mathbf{1}$ where $\mathbf{x}$ is a *one-hot* document vector;
- **MLP**: Multi-layer Perception;
- **CNN**: Convolutional Neural Network with one hidden layer and three filters corresponding to unigrams, bigrams and trigram features, as described by Kim (2014);

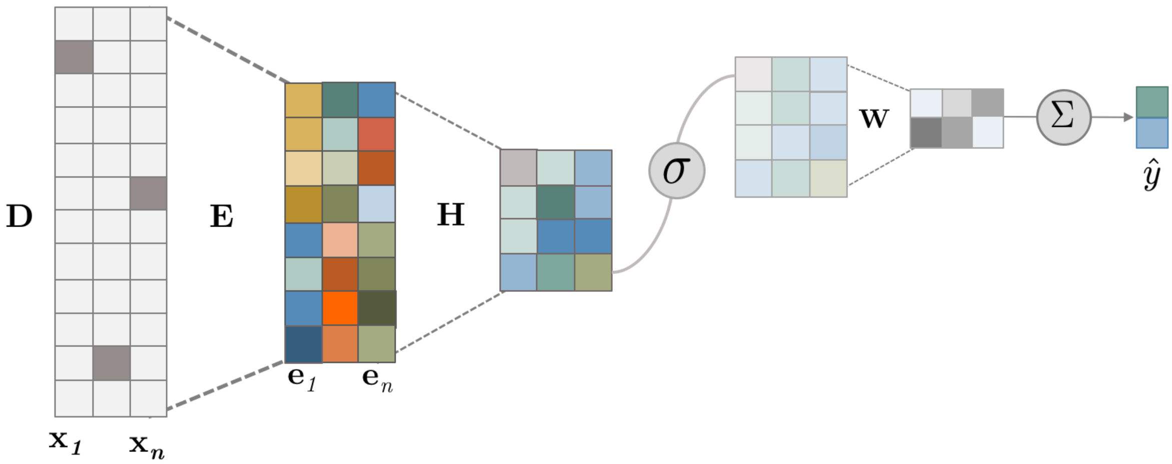


Figure 3.2: Schematic depiction of the Non-Linear Subspace Embedding Model for document-level classification

3.2.2 Experimental Setup

Our method requires a set of unsupervised word embeddings, thus we resorted to the corpus of 52 million raw tweets described by Owoputi et al. (2013) to estimate 400-dimensional vectors with the Structured Skip-Gram model (Ling et al., 2015a). To help reduce the lexical variability, all the datasets were pre-processed with the typical procedure (described in Section 2.1.5). The experiments were conducted using 10-fold cross-fold validation, using the same partitions across all the models. For each partition, the training set was divided into 80%/20% train/validation subsets to facilitate model selection by early-stopping and hyper-parameter selection via grid search.

Regarding the hyper-parameters, we searched the best subspace size for the NLSE model in the range $S = [3,5,10,15,20]$, the number of feature maps and size of the hidden layer for the CNN in the set $M = [50,100,200,400]$ and size of the hidden layer in the same range for the MLP model. The Näive Bayes baseline was trained with Maximum Likelihood Expectation while the rest of the models were trained with *vanilla* Stochastic Gradient Descent to minimize the inverse average log-likelihood of the observed data $\mathcal{D}$

3.2.3 Results

The classification results presented in Figure 3.3 show that the NLSE model largely outperform the linear models. This is intuitive and agreeable with the literature — simple bag-of-word models are insufficient for inferences over subjective content. The

	Dataset	Size	Description
Sentiment	TW-13	3812	Tweets annotated w.r.t overall sentiment by SemEval for their Twitter Sentiment Analysis challenge
	TW-14	1852	
	TW-15	2389	
Opinion	OMD	2385	Tweets with reactions to a debate opposing presidential candidates Barack Obama and John McCain in 2008 (Diakopoulos and Shamma, 2010)
	HCR	2392	Tweets discussing the health care reform in the USA in 2010 (Speriosu et al., 2011)

Table 3.1: Summary of the sentiment analysis and opinion mining datasets used in the social media text classification experiments.

Näive Bayes model is particularly inefficient, due to its inability to cope with the extreme feature sparsity caused by noisy data. Despite using the same sparse document representations, the BOW model performs much better, because it can benefit from regularization and learn to ignore non-informative tokens.

Regarding the neural models, we observe that the MLP results are comparable to the BOW while the BOE models tend be slightly better. The embeddings offer more compact representations and can represent arbitrarily large vocabularies without increasing the model's complexity. This can alleviate the problems of feature sparsity and out-of-vocabulary words but, on the other hand, models based on these representations do not benefit from regularization. In fact, the opposite is true, eliminating the contribution of individual dimensions degrades the expressiveness of the embeddings. Our experimental results confirm this observation; we noted that the hyper-parameter selection process always favored configurations with less regularization. We can also see that the CNN performs very poorly as it requires more training data to be accurately optimized. Compared to these baselines, the NLSE takes the best of worlds — on the one hand, it can represent large vocabularies with compact representations. On the other hand, aggressively reducing the representation space sharply reduces the model's capacity, thus imposing an implicit form of regularization and forcing the model learn more discriminative representations.

3.3 Case Study: Agile Social Science

This section aims to validate the proposed low-resource neural networks in real-world applications, by conducting a case-study of social media analysis applied to the social

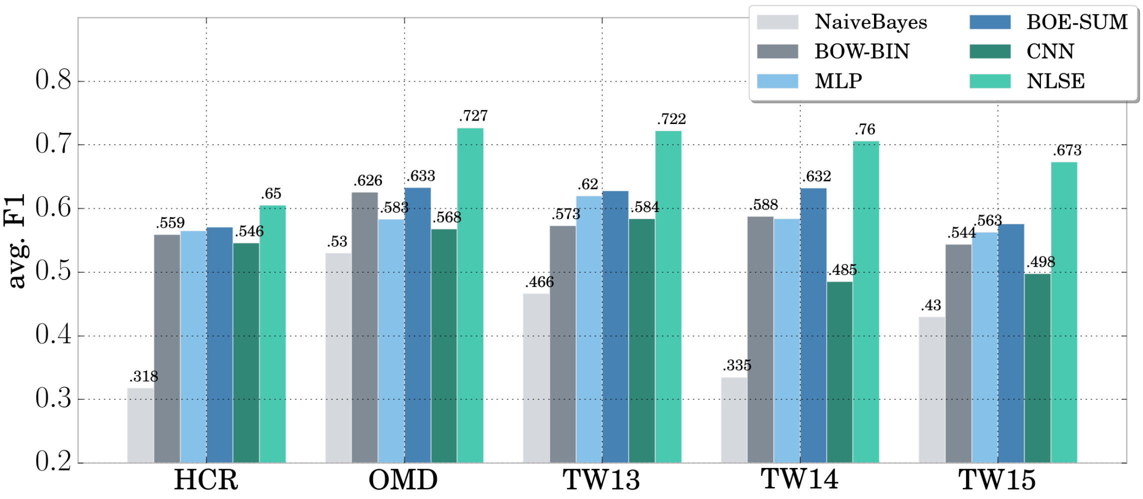


Figure 3.3: Classification results in terms of the average F_1 metric

sciences. We used datasets collected and curated by computational social scientists at the *Centre for the Analysis of Social Media*[4] (CASM) in the UK, for a wide range of exploratory social sciences studies ranging from reactions to political debates to natural disasters. These datasets, described by Kober and Weir (2015) and summarized in Table 3.2, consist of training data for bespoke SMA pipelines containing annotations for either objective tasks, such as topic detection and relevancy filtering (♠), or subjective tasks, e.g. sentiment analysis and opinion mining (♡). The pipelines were created with the modeling framework developed by Wibberley et al. (2013), which is based on the DUALIST — a system proposed by Settles (2011) to reduce annotation efforts in the development of text classifiers. However, the DUALIST was created for conventional text and thus ignores the challenges of processing social media. Critically, it uses Näive Bayes as the underlying model which is particularly inefficient for these data, as we have seen above.

3.3.1 Evaluation

We compared the performance of NLSE classifiers and generic linear classifiers over the aforementioned datasets. The evaluation was conducted using the same experiment protocol and the same pre-trained word embeddings from the previous section. Figures 3.5 and 3.4 show the results for the sentiment analysis and topic detection datasets, respectively. Once again, we observe that the NLSE model consistently outperforms the

[4] `https://www.demos.co.uk/research-area/centre-for-analysis-of-social-media/`

Dataset	Task	Size	Description
boo-cheer	♡	1665	Opinions about a political debate
cameron-1	♠	205	Opinions about former British Prime Minister David Cameron
cameron-3	♡	502	
clacton	♡	930	Attitudes about a controversial by-election triggered by the UK Independent Party
clegg	♡	500	Opinions about British politician Nick Clegg
debate	♡	306	Opinions about a political debate
duggan-1	♠	475	Tweets discussing a controversial story of alleged police brutality in the UK
duggan-3	♡	401	
farage	♡	2614	Opinions about British politician Nigel Farage
flood-1	♠	530	Tweets related to winter floods that occurred in the South of England
flood-2	♠	1615	
miliband-1	♡	927	Opinions about a British politician Edward Miliband
miliband-2	♡	449	

Table 3.2: Summary of the agile social science datasets, described by Kober and Weir (2015). These datasets were collected from different stages of custom classification pipelines created by computational social scientists, and refer to either topic detection (♠) or sentiment analysis (♡) tasks.

other baselines, particularly on the more subjective datasets. We can also see that simple Näive Bayes classifiers perform terribly on most datasets, which is not surprising given their inability to cope with noisy data. The BOE models have very inconsistent results sometimes performing better than BOW but in other cases is much worse.

These results show that just reducing the data annotation efforts is insufficient to improve the quality of SMA applications. Adequate inference of noisy and subjective content also requires specialized models that are able to exploit more sophisticated and fine-grained representations of the input. Simplistic models such as Näive Bayes, ought to be avoided for such inferences as they might erroneously lead us to believe that there is no signal in data. Despite the obvious limitations, these methods are still being used by computational social scientists in real world analyses. In contrast, the NLSE model's ability to learn specialized models for a wide range of tasks and datasets, strongly suggests that this method should be part of any social media analyst toolkit, and constitutes a new baseline against which further developments in rapid deployment of SMA models should be measured. One advantage of Wibberley et al. (2013)'s approach is that it can easily support *dual-supervision*, i.e. leveraging annotations for documents and individual words, which can reduce annotation efforts. How to operationalize this technique with our method remains an open question and a worthwhile topic for future work.

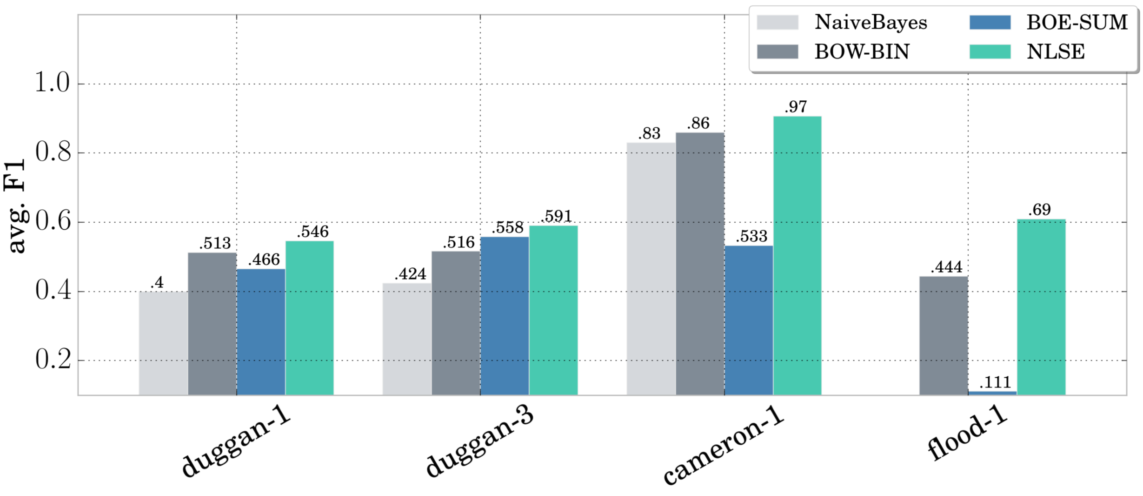


Figure 3.4: Topic classification

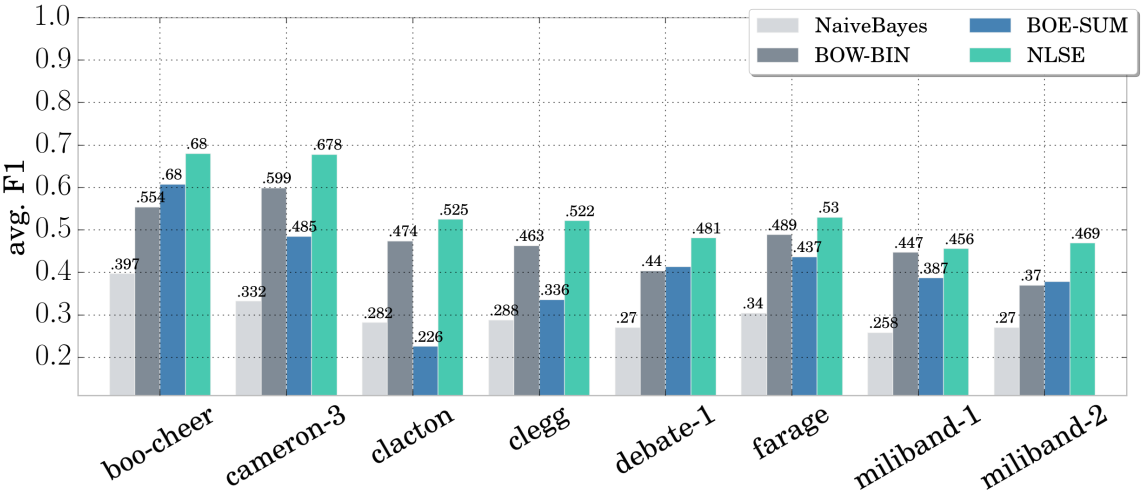


Figure 3.5: Sentiment classification

3.4 Conclusions

This chapter investigated methods to reduce the gap between the manual efforts that go into the development of SMA models and the quality of the resulting models, in terms of predictive performance. To that end, we developed a method to build supervised neural network models in low-resource settings by: first, exploiting large amounts of unlabeled data to learn generic neural embeddings, thereby mitigating the manual efforts in data annotation; and second, constructing low-capacity neural architectures to automatically tailor the embeddings and induce specialized models for the task at hand, thereby ameliorating the manual efforts in feature engineering.

The evaluation was first performed over a set of Twitter sentiment analysis and opinion mining datasets. We compared the performance of the NLSE against that of other off-the-shelf baselines i.e. bag-of-words models without manually crafted features and deep neural network models. The results showed that our model has superior performance across all the datasets. This was expected, since generic BOW models are not able to capture the subtleties of subjective communications, and complex neural models require larger training datasets to be properly optimized. Then, a case-study of real-world social media analyses for the social sciences domain was conducted. We leveraged a collection of datasets compiled by social scientists to build SMA pipelines and evaluated the performance of bespoke classifiers for objective and subjective tasks built with our approach. Again, we found that the NLSE consistently outperforms the baselines. As expected, the gains were particularly pronounced in the more subjective tasks which require more sophisticated representations. These results demonstrate that our approach can be leveraged to quickly deploy document-level classifiers for a variety of problems, with little explicit intervention of subject matter experts or domain knowledge, even in low-resource settings.

This approach offers a reasonable trade-off between the simplicity and interpretability of standard linear models and the benefits of neural representation learning. However, we should emphasize that we are essentially trading variance for bias and this comes with some caveats, particularly when training with small datasets. Low-resource training imposes limitations on what kind of models can be learned — more complex problems will always require more sophisticated models and consequently more labeled data. An underlying assumption of this method is that the pre-trained embeddings already encode the relevant information for a specific task, to some degree. If this is not the case, however, then more training data might be required to learn appropriate representations. Another important limitation of this approach is the assumption that there is a pre-trained embedding available for every possible word; but even embeddings estimated from a very large corpus will certainly not contain all the possible types — words not seen during embedding estimation will not have an associated representation. One way to tackle this issue is by using character-level embeddings as the input and then leverage a composition function to generate representations for any word from the individual characters. We did some preliminary experiments along these lines but we used a very rudimentary character embedding model which was much worse than the word embeddings for known words (Amir et al., 2016b). Nevertheless, I believe that better results can be obtained with more sophisticated models that have been recently proposed (Ling et al., 2015b; Kim et al., 2016).

Chapter 4

Word-level Models for Lexicon Induction

Lexicon-based social media analysis approaches are hampered by the scarcity of comprehensive lexicons describing subjective properties of words, as they are used in social media communications. Creating these resources from scratch is laborious and time-consuming, hence, improving lexicon-based SMA requires the ability to easily create comprehensive lexicons tailored for specific languages and applications. The previous chapter, introduced the NLSE model and demonstrated that it can be used to induce specialized document-level models. This chapter further exploits this method to derive specialized word-level models from labeled lexicons. The resulting models can naturally be used to extrapolate such labels to any other word (provided that it has an embedding representation) and thus can be used to automatically expand pre-existing linguistic resources. To evaluate this approach, we developed word-level models to predict the ratings assigned by humans to words in seven well-known lexicons, describing 14 subjective aspects (Section 4.1). Then, we investigated the impact of low-resource training in the performance of the NLSE. Finally, we used a word-level model to induce a large-scale sentiment lexicon and assessed the impact of the expanded lexicons in downstream SMA applications. To that end, we compared the performance of small and expanded lexicons over lexicon-based Twitter sentiment classifiers. We also compared with other baselines leveraging a publicly available large-scale sentiment lexicon and supervised classifiers trained with small datasets (Section 4.2).

4.1 Word-level Models

In this section, we leverage the NLSE to build specialized word-level models to adapt pre-existing manually curated lexicons to social media. Given a lexicon describing subjective aspects of words (e.g. sentiment polarity) the goal is to train models to predict those aspects for other, unseen, words. Most lexicon induction methods proposed in the literature follow the intuitive assumption that *similar* words should have similar labels (see Section 2.2.1). For example, we would like to assign similar labels to different spellings, inflections and synonyms of the same word (e.g. `l0ver`, and `loveerz`), and to semantically related terms (e.g. `happy` and `#feelinggreat`). To operationalize this intuition, we capitalize on two fundamental properties of neural word embeddings: first, they capture latent word semantic aspects, some of which correlate with subjective properties; and second, they encode semantic and syntactic similarities in terms of vector similarities — this will allow us to infer consistent labels for similar and related words. Note that leveraging word embeddings estimated from a large corpus, representative of language used in social media, allows the models to infer labels even for misspelled words and abbreviations. Therefore, this method corresponds to adapting small manually curated lexicons to social media environments.

Previous work by Tang et al. (2014) and Amir et al. (2015) have also investigated the use of word embedding based models to expand Twitter sentiment lexicons. However, both those methods are inherently limited by their choice of word representations — the former uses distantly supervised *sentiment specific* embeddings tailored to capture sentiment information, which constrains the method to sentiment polarity lexicons. The latter, on the other hand, can be used for other lexicon types but it relies on purely unsupervised embeddings, which are sub-optimal for specialized downstream models (Astudillo et al., 2015b). Herein, we capitalize on the NLSE to generalize this approach — the model uses generic embeddings as input, and thus can be used for any lexicon. On the other hand, it also learns tailored representations and thus can be applied to more nuanced properties of words (e.g. emotional tone). Let $\mathcal{D} = \{(w^{(1)}, y^{(1)}), \ldots, (w^{(N)}, y^{(N)})\}$ be lexicon associating words $w \in \mathcal{V}$ to labels $y \in \mathcal{Y}$, and $\mathbf{E} \in \mathbb{R}^{e \times |\mathcal{V}|}$ be an word embedding matrix. Given a word w represented as a *one-hot* vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ we estimate a label as

$$\begin{aligned} \hat{y} = f_\theta(\mathbf{x}) &= \mathbf{W} \cdot \phi(\mathbf{x}) + b \\ \phi(\mathbf{x}) &= \sigma(\mathbf{H} \cdot \mathbf{E} \cdot \mathbf{x}) \end{aligned} \tag{4.1}$$

where $\sigma(\cdot)$ is a sigmoid non-linearity, $\mathbf{W} \in \mathbb{R}^{1 \times s}$ denote the model weights, and $\mathbf{H} \in \mathbb{R}^{s \times e}$ is an embedding subspace projection matrix that induces specialized representations.

4.1.1 Evaluation: Expanding Subjective Lexicons

The evaluation consisted of using labeled lexicons to train predictive models to infer the same labels for unseen words. The experiments were conducted over seven lexicons, summarized in Table 4.1, describing 14 subjective aspects: *subjectivity*, *sentiment polarity*, affective responses (i.e., the *valence*, *arousal* and *dominance*), *happiness*, and Plutchik (1980)'s eight basic *emotions*. Some of these lexicons are categorical, i.e. associating words to specific classes; and others are continuous, assigning real-valued scores to words. For the former, the models were evaluated with respect to the Average F_1 metric, and for the latter, the models were evaluated in terms of the Kendall τ.

As baselines, we considered other methods using word embeddings as the input, including a simple linear model (linear) and a Support Vectors Machine model (Vapnik, 2000) with a non-linear kernel (RBF) as described by Amir et al. (2015). We also compared the performance of linear models with features derived from two other methods that are commonly used to extract more compact representations of the input: (i) using ℓ_1-norm regularization to implicitly prunes the feature space by forcing the weights associated to some input dimensions to be exactly zero ($\boldsymbol{\ell_1}$); and (ii) using PCA to extract lower-dimensional features from the word embeddings (PCA).

The experiments were performed with the same word embeddings described in Section 3.2.2. The labeled data (i.e., the lexicons) was split into 80% for model training and 20% for evaluation. In each experiment, 20% of the training data was reserved for early stopping and hyper-parameter tuning. The linear baselines were trained with the SVM implementation available in the `sci-kit learn` python package, and the NLSE was trained with *vanilla* Stochastic Gradient Descent with fixed learning rate. Regarding the hyperparameters, we picked the optimal misclassification cost from the set $C = \{1e^{-2}, 1^{e-1}, 1, 10, 50, 100, 150\}$ and kernel width from the set $K = \{1e^{-3}, 1e^{-2}, 1^{e-1}, 1, 10\}$ in the RBF baseline; regularization constant from the set $B = \{1e^{-4}, 1e^{-3}, 1e^{-2}, 1^{e-1}, 1, 10\}$ in the regularized baselines; the number of components to keep in the PCA baselines, from the set $S = \{5, 10, 15, 20\}$; and size of the embedding space for the NLSE also in the set S.

4.1.2 Results

The main experimental results with categorical and continuous lexicons are presented in Tables 4.2 and 4.3, respectively. We can see that the NLSE largely outperforms all the other baselines, apart from two exceptions, where the RBF does slightly better. Regarding

	Lexicon	Size	Description
Categorical	OML	6,787	Opinion mining lexicon with words categorized with respect to sentiment polarity (Hu and Liu, 2004)
	MPQA	6,886	Subjectivity lexicon with words rated as objective/subjective and with respect to sentiment polarity (Wilson et al., 2005)
	EmoLex	14,174	Emotion lexicon describing words in terms of their sentiment polarity and association with Plutchik (1980)'s set of basic emotions (Mohammad and Turney, 2013)
Continuous	ANEW	1,040	Affective lexicon with words rated with respect to the degree *valence*, *arousal*, and *dominance* that they evoke (Bradley and Lang, 1999)
	Ext-ANEW	13,915	Expanded version of the ANEW lexicon with ratings for additional words obtained from Amazon's Mechanical Turk workers (Warriner et al., 2013)
	SemLex	1,515	Sentiment lexicon created for SemEval's lexicon induction shared task (Rosenthal et al., 2015). The words are rated in continuous scale representing the degree of association to the positive polarity.
	LabMT	10,000	Happiness lexicon with words annotated in terms of *happiness* score, assigned by Amazon's Mechanical Turk workers (Dodds et al., 2011).

Table 4.1: Summary of the subjective lexicons used in the experiments

the other baselines that try to uncover the relevant information from the embeddings (i.e., PCA and ℓ_1), again we can see that they perform very poorly. This was expected, since the former induces a low-rank approximation of the original embedding, that (tries to) preserve most of the variance. However, since word embeddings are distributed representations, the values of individual dimensions are meaningless. The latter, on the other hand, tries to drop some of the input dimensions, but in doing so degrades the information contained in the word representations. These approaches are particularly inefficient in the more nuanced properties such as fine-grained emotions. Conversely, these are precisely the cases where our approach stands-out, which underlines the benefits of inducing task-specific representations.

		NLSE	SVM			
			linear	ℓ_1	**RBF**	**PCA**
OML	sentiment	**0.882**	0.868	0.686	0.872	0.852
MPQA	sentiment	**0.691**	**0.691**	0.221	0.669	0.555
	subjectivity	0.825	0.819	0.798	**0.833**	0.805
EmoLex	sentiment	**0.676**	0.630	0.404	0.640	0.468
	sadness	**0.509**	0.340	0.167	0.334	0.000
	fear	**0.503**	0.373	0.261	0.394	0.000
	anger	**0.468**	0.353	0.214	0.366	0.000
	disgust	**0.446**	0.343	0.180	0.352	0.000
	joy	**0.440**	0.333	0.148	0.329	0.000
	trust	**0.403**	0.201	0.190	0.167	0.000
	surprise	**0.204**	0.167	0.093	0.119	0.000
	anticipation	**0.240**	0.108	0.151	0.044	0.000

Table 4.2: Results for categorical lexicons in terms of Avg. F_1

		NLSE	SVM			
			linear	ℓ_1	**RBF**	**PCA**
SemLex	sentiment	**0.667**	0.610	0.619	0.630	0.622
LabMT	happiness	**0.640**	0.576	0.573	0.622	0.464
ANEW	arousal	**0.440**	0.365	0.375	0.415	0.389
	valence	**0.683**	0.612	0.604	0.646	0.592
	dominance	**0.546**	0.477	0.456	0.494	0.475
Ext-ANEW	arousal	0.393	0.373	0.371	**0.397**	0.315
	valence	**0.607**	0.567	0.565	0.593	0.494
	dominance	**0.480**	0.445	0.443	0.464	0.405

Table 4.3: Results for continuous lexicons in terms of Kendall τ rank correlation

The overall results show that the word-level models can indeed capture a wide range of semantic properties and be leveraged to induce large-scale subjective lexicons, from small manually labeled samples. In Table 4.4 we present some examples of words from SemLex (top row) along with other similar words with inferred labels (bottom rows). We can see that the model assigns consistent scores to related words and morphological variations thereof, suggesting that this method can be used to accelerate the creation of large lexicons. The quality of the resulting lexicons can be then improved by human experts, by discarding irrelevant words and erroneous labels, which will certainly be less laborious than manually assigning labels to new words. In the next section, we will see that we can also use task-specific labeled data to automatically improve the expanded lexicon.

love	0.93	**fun**	0.883	**lol**	0.625
lovely	0.880	jokes	0.878	lololol	0.671
loving	0.875	comedy	0.878	lolol	0.650
loved	0.874	funny	0.814	lolo	0.642

Table 4.4: Predictions from a word-level model trained with the SemLex lexicon — the top row shows words and scores that were present in the training data and the bottom rows shows cherry picked examples of predictions made by the model. We can see that the model assigns consistent labels to related and similar words.

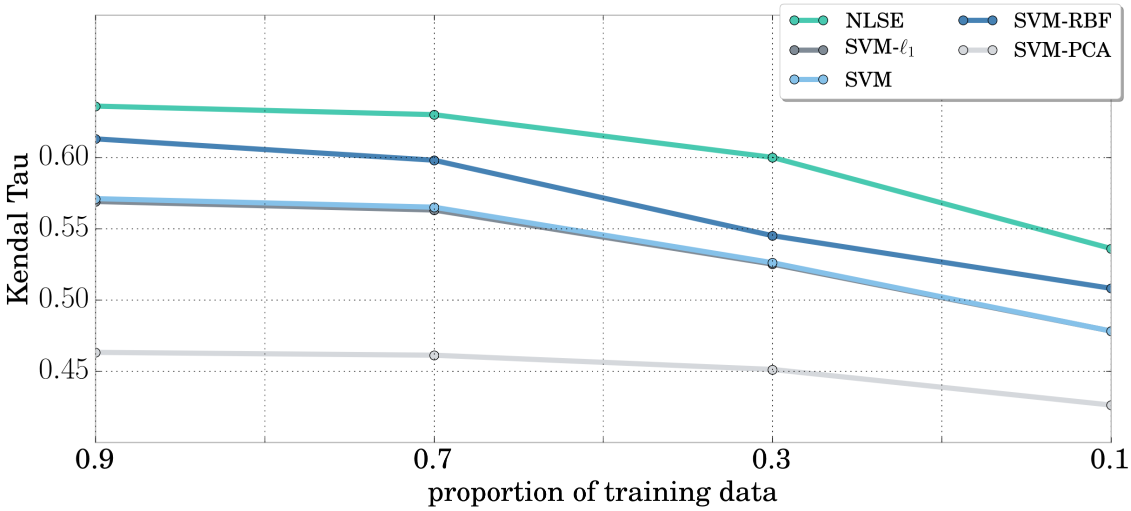


Figure 4.1: Performance of the different baselines in predicting the *happiness score* of words, as a function of the size of the training data.

To take a closer look at the low-resource learning capabilities of our method, we used the LabMT lexicon and repeated the experiments with smaller fractions of the training data. We then plotted the performance of the different baselines as a function of the training data size (Figure 4.1). As expected, the performance of all models decreases monotonically as the training set size decreases but we observe that the NLSE performance decays slower than the RBF baseline (the second-best method). Notably, when trained with 30% of the data, the NLSE model attains the same performance of the SVM RBF baseline trained with 70% of the data, again reinforcing the benefits of leveraging specialized low-dimensional representations in low-resource settings.

Finally, we investigated the effect of the subspace projection on the word representation space. To that end, we used Van der Maaten and Hinton (2008)'s T-SNE algorithm to project the embeddings associated to words from SemLex into two-dimensions, and plot-

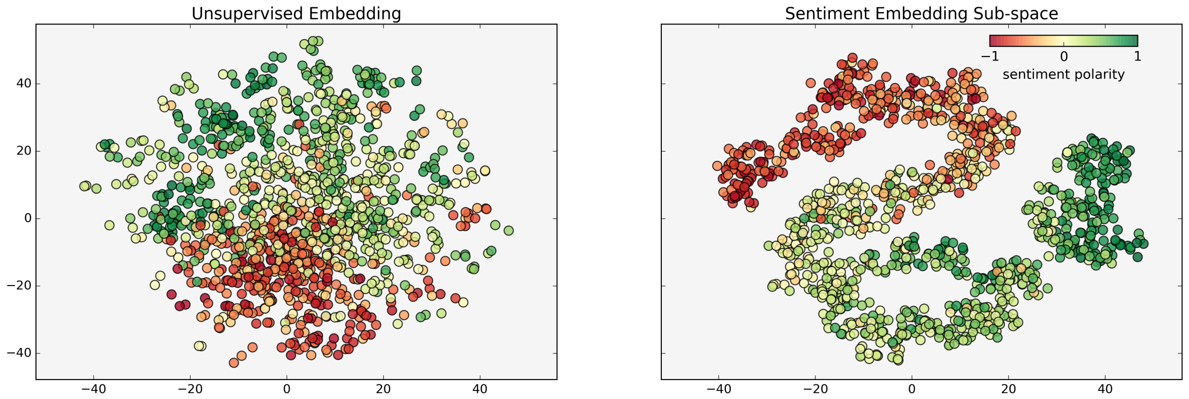


Figure 4.2: T-SNE projection of the embeddings associated to words from SemLex, to two dimensions. The points are colored according to their sentiment polarity. The left plot, shows the words represented as 600-dimensional unsupervised embeddings. The right plot, shows the same words represented with *task-specific* embeddings induced with NLSE model.

ted the resulting vectors colored according to their sentiment score. Figure 4.2 compares the structure of the generic embedding space with the task-specific embedding subspace induced by the NLSE. On the left-hand side, we can see that the generic embeddings can naturally capture sentiment information — words with similar sentiment scores tend to be closer to each other. On the right-hand side, we see that in the space induced by the sub-space projection, not only are similar words (w.r.t to sentiment) drawn even closer but also, quite interestingly, the vectors become arranged in what seems to be a *continuum* from the most negative to the most positive sentiment polarity. This demonstrates that the subspace projection is indeed able to uncover some underlying substructure of the embedding space.

4.2 Improving Lexicon-based Social Media Analysis

In this section, we assess the impact of expanded lexicons in practical applications by developing simple lexicon-based sentiment classifiers, using Eq. 2.2.

4.2.1 Calibrated Lexicons

Large-scale lexicons can improve models predictive performance by accounting for a number of words that would otherwise be missed by manually curated resources — in other words, they help to reduce the *bias* of the model. On the other hand, the majority of words do not convey any subjective information (e.g. only a fraction of all the possible words express positive or negative sentiments), therefore the main strength of large-scale lexicons is also their weakness. Because we have labels for all the words in a vocabulary, they all contribute to the predictions which increases the models *variance*[1].

Dodds et al. (2011) also noted this issue and addressed it by manually specifying a threshold score to restrict the lexicon only to words that strongly convey sentiment (i.e. neutral and objective words were discarded). However, they were using a manually curated lexicon of 10k words whereas our lexicons are at least one order of magnitude larger, which makes their manual inspection unpractical. Ideally, automatically created lexicons should be manually inspected and validated against domain experts knowledge. If this is not possible, a reasonable approach is to use some task-specific labeled data to inform the pruning of irrelevant words. The simplest methodology to accomplish this is with a grid search to find the upper and lower score thresholds, that optimize the predictive performance with respect to a specific task/dataset.

4.2.2 Evaluation

The evaluation was conducted over the same sentiment analysis and opinion mining datasets summarized in Table 3.1. These datasets are labeled with respect to the 3 classes but here we restricted the experiments to binary classification, i.e. considering only positive versus negative documents. We compared the performance of lexicon-based classifiers built on top of the following sentiment lexicons:

- *SemLex*, a small manually labeled Twitter sentiment lexicon
- *SemLex-XL* obtained by automatically expanding *SemLex* with our method;
- *NRC-HS* obtained by measuring the PMI between new words and a set of affectively charged hashtags Kiritchenko et al. (2014);

We also assessed the impact of calibrating the lexicons with a small amount of labeled data and compared with the performance of simple supervised text classifiers trained with the same amount of labeled data. Thus, we reserved a development set of 10% of each dataset and used the rest to test the models. The classification threshold was set to $t = 0$ and the lexicons were calibrated by filtering words with scores between the range $[s^{-}; s^{+}]$, so that only strongly negative words and strongly positive words were kept. To

[1] This can also be seen as *precision* versus *recall* trade-off.

that end, we performed a grid search over the intervals $s^- \in [-0.4, -0.1]$ and $s^+ \in [0.1, 0.4]$ and picked the values that maximize the models performance on the development set.

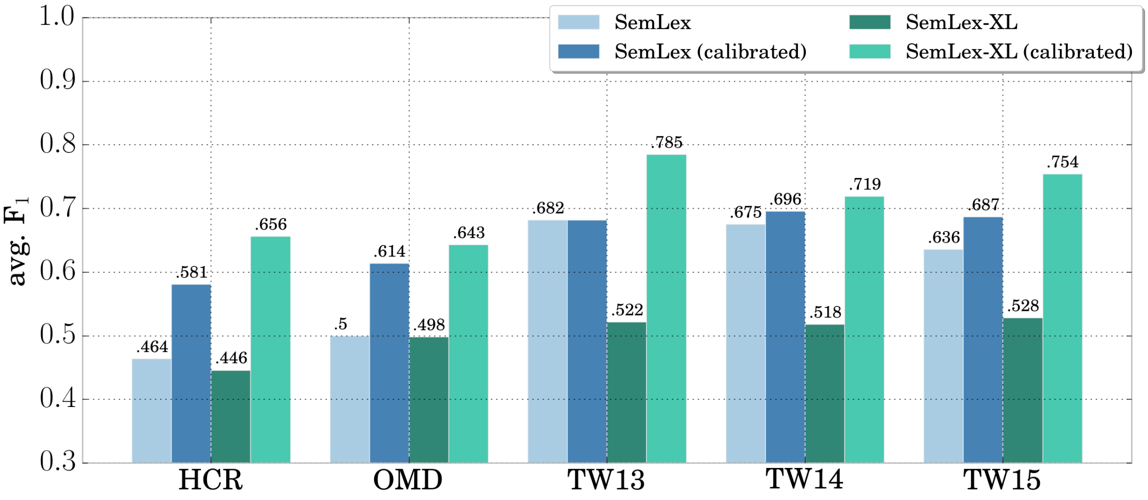


Figure 4.3: Improving the performance of lexicon-based classifiers

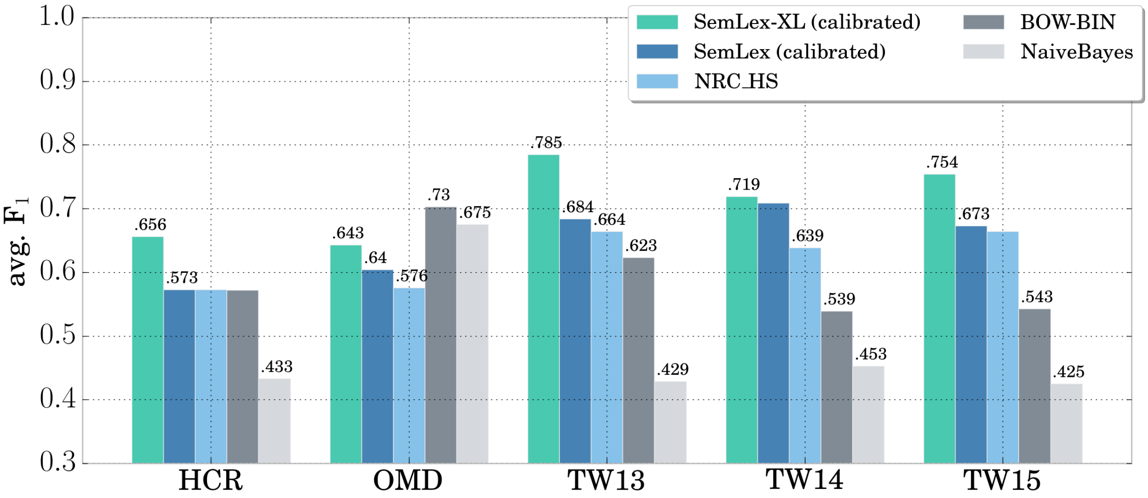


Figure 4.4: Performance of calibrated lexicon-based classifiers built on top of different lexicons compared with supervised models in binary classification over sentiment analysis and opinion mining datasets.

4.2.3 Results

Figure 4.3 compares the results obtained with classifiers induced with *SemLex* and *SemLex-XL*, with and without the aforementioned calibration procedure. We observe that calibrating the lexicons significantly improves models performance in nearly all the cases, particularly with the large-scale lexicon, yielding gains of up to 26% in absolute performance. Without this step, the large lexicon performs rather poorly, which was expected due to an increase in model variance. Overall, we can see that the expanded lexicons drastically improve model performance affording gains of up to 10% in absolute performance when compared to using just the original lexicon, provided the proper calibration. If we compare the performance of the large-scale calibrated lexicon against the small (uncalibrated) lexicon, then the gains go up to 19%. Figure 4.4 compares the performance of the lexicon-based models against bag-of-words (BOW) classifiers trained with the same datasets used to calibrate the lexicons. The plots show that *SemLex-XL* outperforms the other baselines in most cases, with the exception of the OMD dataset. Indeed, lexicon-based classifiers generally outperform supervised models trained with such limited amounts of labeled data (in these experiments we used 250 examples on average). These results demonstrate that lexicon-based classifiers can be a viable alternative to supervised models in the absence of large amounts of unlabeled data. Specially, by combining large scale lexicons, which can be obtained from a small number of examples, with small amounts of task-specific labeled data which can also be easily created in a few minutes.

4.3 Conclusions

This chapter demonstrated that the low-resource neural networks can be used to automatically adapt pre-existing lexicons to social media environments. Concretely, we can build specialized word-level models which can be trained from small lexicons to infer similar labels for unseen words. This approach was validated in a lexicon expansion task aiming to predict the ratings assigned by human judges to various subjective lexicons. The experimental results showed that our models can be trained to infer high-level subjective word properties and assign consistent labels (w.r.t. to those properties) to similar and related words. Furthermore, these models consistently and significantly outperform baselines based on generic embeddings, particularly on more nuanced properties (e.g. fine-grained emotions). We observed that generic embeddings can, to some extent, capture various subjective attributes but the tailored representations induced by the NLSE significantly improve performance and allows models to make better use of limited datasets.

Then, we assessed the impact of large-scale lexicons induce with our method in downstream SMA applications. To that end, we built and evaluated lexicon-based Twitter sentiment classifiers and compared the performance of small and expanded lexicons. We found that the more comprehensive lexicons can dramatically improve the performance of downstream models, but realizing those gains requires calibrating the lexicons to remove irrelevant or ambiguous words. This can be done by manually inspecting and filtering the lexicons or in an automatic fashion with a small amount of labeled data. Both cases require significantly less efforts than manually annotating the lexicons from scratch. Moreover, this also reduces the amount of work required from domain experts — even the manual calibration can, for the most part, be done with just common sense by discarding irrelevant and non-sensical words. We should note that in these experiments we focused exclusively on using lexicons to build predictive models; in practice, however, these resources can also be used to search and filter for relevant content but this would be more difficult to evaluate objectively.

Chapter 5

User-level and Contextualized Models

Social Media Analysis research and applications have been mostly focused on inference models operating over the contents of individual posts. Nevertheless, for many applications it is critically important to characterize the *users* involved in the communications. Furthermore, the ability to complement document-level models with user information can help alleviate the problem of lack of context caused by the brevity and spontaneity of social media. This is particularly relevant for more ambiguous and nuanced communications, for instance involving sarcasm and other types of figurative language, where contextual information plays a pivotal role.

The main contributions of this chapter are two-fold: first, we introduce *User2Vec*, a novel neural language model that leverages the prior posts of social media users to estimate representations that, similarly to word embeddings, capture latent aspects and encode similarities as vector distances (Section 5.1). We then exploit the embeddings to derive specialized user-level models with our low-resource neural network approach. These models were evaluated over a mental-health inference task, aiming to discriminate between Twitter users affected with depression or PTSD from matching controls, given their social media posting histories (Section 5.2).

Second, we leverage the user embeddings to derive *Content and User Embedding Neural Networks* (CUE-NN), a novel family of deep neural networks that combine representations of the content and of the author of a post, to support contextualized inferences over social media. The contents are represented with high-level features extracted with a set of convolutional filters, and the authors are represented with their respective user embeddings. The evaluation was conducted over a sarcasm detection task (Section 5.3). Finally, we examined the user embeddings in more detail. First, we investigated what kinds of information are being captured and encoded in these representations (e.g. political leanings); second, we systematically evaluated the embeddings ability to capture homophilic relations (Section 5.4).

5.1 Learning Neural User Embeddings

The general approach to learn unsupervised word embeddings entails associating words with parameter vectors, which are then optimized to predict other words that occur in the same contexts (Goldberg, 2016). Essentially, these methods rely on the distributional hypothesis to infer the latent semantics of words (Harris, 1954). For example, the widely popular *Skip-Gram* model operationalizes this approach by sliding a *window* of a pre-specified size across texts — at each step, the center word is used to predict the probability of one of the surrounding words, sampled proportionally to the distance to the center word (Mikolov et al., 2013a). Le and Mikolov (2014) later expanded this approach with *Paragraph2Vec*, introducing two methods to learn representations for paragraphs (or, more generally, sequences of words) — (i) *PV-DM*, which tries to predict the center word of the sliding window, given the surrounding words and the paragraph (i.e., their respective embeddings); and (ii) *PV-DBOW*, which tries to predict the words in a sliding window within a paragraph, conditioned only on the respective paragraph embedding. We hypothesize that user embeddings can also be learnt with a neural language modeling approach, i.e. associating *users* with parameter vectors, and optimizing these to accurately predict observable attributes or the words used in previous posts written by said user. We hope that, similarly to word embeddings, the resulting representations encode interesting latent personal aspects and capture a soft notion of *homophily*, implying that *similar* users are embedded in nearby regions of the vector space.

5.1.1 *User2Vec*

We now describe our approach to infer generic user embeddings from posting histories. The idea is to capture the relations between users and the written content they generate, by modeling the probability of observing a sequence of words conditioned on the respective author. More formally, let $\mathcal{U}$ be a set of users, $\mathcal{C}_j$ be a collection of posts authored by user $u_j \in \mathcal{U}$, and $d = \{w_1, \ldots, w_N\}$ be a post composed of words w_i from a vocabulary $\mathcal{V}$. The goal is to estimate the parameters of a user vector $\mathbf{u}_j$, that maximize the conditional probability

$$P(\mathcal{C}_j|u_j) \propto \sum_{d \in \mathcal{C}_j} \sum_{w_i \in d} \log P(w_i|\mathbf{u}_j) \tag{5.1}$$

However, directly estimating these quantities (e.g., with a log-linear model) would require calculating a normalizing constant over a potentially large number of words, a computationally expensive operation. Because we are only interested in the user vectors $\mathbf{u}_j$ and not the actual probabilities, we can approximate the term $P(w_i|\mathbf{u}_j)$ by minimizing the following Hinge-loss type objective

$$\mathcal{L}(w_i, u_j) = \sum_{\tilde{w}_k \in \mathcal{V}} \max(0, 1 - \mathbf{e}_i \cdot \mathbf{u}_j + \tilde{\mathbf{e}}_k \cdot \mathbf{u}_j) \tag{5.2}$$

where word $\tilde{w}_k$ (and associated embedding, $\tilde{\mathbf{e}}_k$) is a *negative sample*, i.e. a word not occurring in the post under consideration (authored by user u_j). In the aggregate, such words are less likely to be employed by user u_j than words observed in sentences he or she has authored. In other words, we approximate the objective function by learning to discriminate between observed *positive* examples (sampled from the true distribution) and *pseudo-negative* examples (sampled from a large space of predominantly negative instances), thus shifting probability mass to plausible observations[1]. Intuitively, minimizing this objective attempts to induce a representation that is discriminative with respect to word usage.

Notice that we represent both *words* and *users* via e-dimensional vectors — word vectors, $\mathbf{e}_i \in \mathbb{R}^e$ which are assumed to have been pre-trained via some neural language model; and user vectors $\mathbf{u}_j \in \mathbb{R}^e$ to be learned. We will refer to this approach as *User2Vec*[2]. Furthermore, we note that barring some minor operational differences, this model is equivalent to the PV-DBOW variant of *Paragraph2vec* (if users are viewed as *paragraphs*). The key differences are that: (i) *User2Vec* predicts **all** the words in a post, whereas PV-DBOW slides a window along the paragraph and only predicts one word per step; and (ii) *User2Vec* assumes that the word embeddings are pre-trained, whereas PV-DBOW aims to jointly learn the word and paragraph vectors.

[1] See Dyer (2014) for notes on Negative Sampling and Noise Contrastive Estimation

[2] This formulation is a simplification of Amir et al. (2016c) model. Specifically, we omitted a term in Eq.5.1, encoding the marginal probability of S; and we allow the negative samples to be drawn from all the words in $\mathcal{V}$. These simplifications dramatically reduce training time without significant loss of quality on the resulting embeddings.

5.2 User-level Models

In this section, we capitalize on the user embeddings to extend our low-resource neural network approach to support specialized user-level models.

Let $\mathcal{D} = \{(u^{(1)}, y^{(1)}), \ldots, (u^{(N)}, y^{(N)})\}$ be dataset associating users $u \in \mathcal{U}$ to labels $y \in \mathcal{Y}$, and $\mathbf{U} \in \mathbb{R}^{e \times |\mathcal{U}|}$ be an user embedding matrix representing a set of users $\mathcal{U}$, as e-dimensional vectors. We leverage the NLSE architecture to formulate a neural probabilistic classifier, that given user u represented as a *one-hot* vector $\mathbf{x} \in \{0,1\}^{|\mathcal{U}|}$ estimates a distribution over labels $\hat{\mathcal{Y}}$, as

$$P(\hat{\mathcal{Y}}|u;\theta) \propto \mathbf{W} \cdot \phi(\mathbf{x})$$

$$\phi(\mathbf{x}) = \sigma(\mathbf{H} \cdot \mathbf{U} \cdot \mathbf{x})$$

where $\sigma(\cdot)$ is a sigmoid non-linearity, $\mathbf{W} \in \mathbb{R}^{1 \times s}$ denote the model weights, and $\mathbf{H} \in \mathbb{R}^{s \times e}$ is an embedding subspace projection matrix that learns more nuanced and task-specific representations.

5.2.1 Evaluation: Depression and PTSD on Twitter

We evaluated our user-level classifiers over a mental-health prediction task, using a *self-reported* dataset created for CLPysch 2015, a workshop aiming to foster progress in NLP technologies related to mental health analysis over social media streams (Mitchell et al., 2015; Coppersmith et al., 2015b). The dataset comprises users that have publicly stated on Twitter to have been diagnosed with depression (327 users) or PTSD (246 users), and an equal number of randomly selected demographically-matched users as *controls*. For each user, the associated metadata and posting history was also collected — up to the 3000 most recent *tweets*, per limitations of the Twitter API[3]. The goal was then to automatically discriminate between users affected by depression or PTSD from matching controls, given their Twitter data. It should be noted that we did not participate in this shared task and thus could not obtain the official test data. For this reason, our results are not directly comparable to those reported by the participating teams. We did, nevertheless compare our approach with most proposed methods for this task, none of which was based on neural user representations.

We compared the performance of NLSE classifiers based on embeddings estimated with the *User2Vec* and *Paragraph2vec* models

[3] This data was collected according to the ethical protocol of Benton et al. (2017), and follows the recommendations spelled out in Mikal et al. (2016). For more details on the construction and validation of the data, see (Coppersmith et al., 2015b, 2014a,b).

- **U2V@NLSE**: NLSE model with user embeddings obtained with the *User2Vec* model
- **PV-DM@NLSE**: same as above but with embeddings obtained with *Paragraph2vec*'s *PV-DM* model
- **PV-DBOW@NLSE**: same as above but with embeddings obtained with *Paragraph2vec*'s *PV-DBOW* model

against linear classifiers using textual features based on

- **BOW**: bag-of-words vectors with binary weights, $\mathbf{d} \in \{0,1\}^{|\mathcal{V}|}$;
- **BOE**: bag-of-embeddings. Leveraging the *Skip-Gram* embeddings we built vectors, $\mathbf{d} = \sum_w \mathbf{E}_w$;
- **LDA**: bag-of-topics. We induced $t = 100$ topics using Latent Dirichlet Allocation (Blei et al., 2003), to build vectors $\mathbf{d} \in \{0,1\}^t$ indicating the topics present in user's posts;
- **BWC**: bag-of-word-clusters. We induced $k = 1000$ Brown et al. (1992) word clusters, to build vectors $\mathbf{d} \in \{0,1\}^k$ mapping words in a user's posts to their respective clusters;
- **U2V**: user embeddings obtained with the *user2vec* model
- **PV-DM**: user embeddings obtained with *Paragraph2vec*'s *PV-DM* model
- **PV-DBOW**: user embeddings obtained with *Paragraph2vec*'s *PV-DBOW* model

5.2.2 Experimental Setup

We first preprocessed all tweets with the conventional normalization steps and discarded users with fewer than 100 tweets. To learn *User2Vec* embeddings, we first pre-trained a set of *Skip-Gram* **word** vectors from a large unlabeled corpus comprising the task data and an additional set of 53 Million tweets. Next, for each user $u_j \in \mathcal{U}$, we sampled a held-out set $\mathcal{H}_j \subset \mathcal{C}_j$ with 10% of the posting history. The rest of the data was used to estimate an embedding $\mathbf{u}_j$, by minimizing Eq. 5.1 via SGD and using $P(\mathcal{H}_j|\mathbf{u}_j)$ as the early stopping criteria. *Skip-gram* and *Paragraph2Vec* vectors were estimated using the `Gensim` python package (Řehůřek and Sojka, 2010). To ensure a fair comparison, we kept the hyper-parameters consistent across all the models, which were set as follows: window size $w = 5$, negative sample size $s = 20$ and vector size $d = 400$.

The classification experiments were conducted with 10-fold cross-validation methodology, keeping the same partitions across all the baselines. For each partition, the training split was divided into 80%/20% train/validation sub-splits to facilitate hyper-parameter selection and early-stopping. We performed grid-search to choose the best ℓ_2 regularization coefficient, over the range $C = \{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1, 10\}$, for the linear models; and the optimal subspace size $S = \{10, 15, 20, 25\}$ and learning rates $\lambda = \{0.01, 0.1, 0.5, 1\}$, for the NLSE model.

5.2.3 Results

The classifiers were mainly evaluated with respect to the macro average F_1. We also report results in terms of *binary* F_1, where we only average the scores for the `depression` and `ptsd` classes. This allows us to better ascertain the ability of the models to discriminate between mentally afflicted patients, which are less prevalent than the controls, but are the cases that we mostly care about. The main classification results are shown in Figure 5.1. The first thing to note is that the BOW is a very strong baseline, essentially outperforming all the other linear classifiers based on textual features and generic user embeddings. One reason is that users affected with mental illnesses, often talk about their conditions (whereas healthy users do not generally talk about such issues) and the BOW model can easily pick-up on such clues.

Regarding the user embeddings, we found that, despite being similar, the PV-DBOW performed much worse than the U2V, showing that better embeddings can be obtained by trying to predict **all** the words in users posts, and leveraging pre-trained word vectors. On

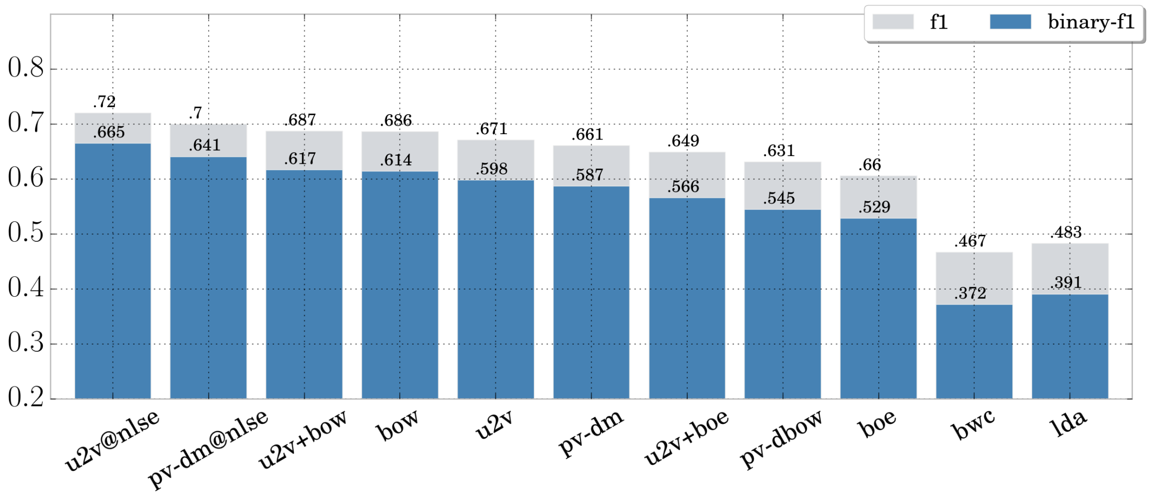


Figure 5.1: Predictive performance of different models at discriminating users with respect to mental condition, in terms of $F1$ and *binary* F_1

.

the other hand, the PV-DM model has a performance comparable to that of the U2V. Note also that these representations perform much better than the word embedding baselines, thus showing that the user embeddings capture more than the sum of their parts, i.e. the individual word (embeddings). Finally, we observe that the task-specific representations obtained via subspace projection outperform all the other baselines by a fair margin. Note also that the NLSE approach is particularly better at discriminating the minority classes, i.e., patients with `depression` and `ptsd`, as evidenced by the greater improvements in *binary* F_1, when compared to the other baselines. These results demonstrate that neural user embeddings induced from previous postings can indeed inform downstream predictive models and be leveraged to build specialized SMA applications in low-resource settings.

5.3 Contextualized Models

In this section, we leverage the user embeddings to derive a novel family of predictive models to support contextualized and personalized inferences over social media. The goal here is to capture both the relevant aspects of the *content* and the relevant *contextual* information about the author of a post. More formally, we are interested in models of the form $P(\mathcal{Y}|d,u) = f_\theta(g(d), z(u))$ that given a post d authored by user u, make predictions with basis on representations of the contents $g(d)$, and representations of the author $z(u)$.

5.3.1 Content and User Embedding Neural Networks

Assuming we have a user embedding matrix $\mathbf{U} \in \mathbb{R}^{e \times |\mathcal{U}|}$, where each column represents a user $u \in \mathcal{U}$ as a e-dimensional vector, we can simply map the author of a post to the corresponding embedding $z(u) = \mathbf{U}_u$. To represent the content, we use pre-trained word embeddings as the input to a convolutional layer that extracts high-level features. The input to this layer is a document matrix

$$\mathbf{D} = \begin{bmatrix} — & \mathbf{e}_1 & — \\ & \vdots & \\ — & \mathbf{e}_m & — \end{bmatrix} \tag{5.3}$$

obtained by selecting the embeddings corresponding to words in d from an embedding matrix $\mathbf{E} \in \mathbb{R}^{e \times |\mathcal{V}|}$. The convolutional layer is composed of a set of filters $\mathbf{F} \in \mathbb{R}^{e \times h}$, each of which *slides* across the input, extracting h-gram features to generate a feature map $\mathbf{m} \in \mathbb{R}^{|d|-h+1}$, where h is the filter *height*. Individual entries in this map are computed as

$$\mathbf{m}_i = \alpha(\mathbf{F} \cdot \mathbf{D}_{[i:i-h+1]} + b) \tag{5.4}$$

where $\mathbf{D}_{[i:j]}$ denotes a sub-matrix of $\mathbf{D}$ (from row i to row j), $b \in \mathbb{R}$ is an additive bias and $\alpha(\cdot)$ denotes an element-wise non-linearity, which we set to be the *Rectified Linear Unit* activation function (Nair and Hinton, 2010). The resulting feature map is then transformed into a scalar with a *max-pooling* operation, i.e., we extract the largest value in the map. We use 3 filters (with varying heights) each generating M feature maps which are then reduced to a single vector $\mathbf{f}^k = [max(\mathbf{m}^1) \oplus max(\mathbf{m}^2) \ldots \oplus\ max(\mathbf{m}^M)]$, where $\oplus$ denotes concatenation. Finally, the output of all the filters is combined to form the final representation $g(d) = [\mathbf{f}^1 \oplus \mathbf{f}^2 \oplus \mathbf{f}^3]$.

The model, which we will refer as *Content and User Embedding Convolutional Neural Network* (CUE-CNN) then induces a probability distribution over target labels, given a post d from user u as

$$\begin{aligned} P(\mathcal{Y}|d,u;\theta) &\propto \mathbf{W} \cdot \phi(\ g(d) \oplus z(u)\ ) + \mathbf{b} \\ \phi(\mathbf{x}) &= \alpha(\mathbf{H} \cdot \mathbf{x}) \end{aligned} \tag{5.5}$$

where $g(\cdot)$ denotes the activations of a hidden layer capturing non-linear interactions between the content and context representations, and $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{H}, \mathbf{h}, \mathbf{F}^1, \mathbf{F}^2, \mathbf{F}^3, \mathbf{E}, \mathbf{U}\}$ are parameters to be estimated during training. Here, $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times z}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{Y}|}$ are the weights and bias of the output layer; $\mathbf{H} \in \mathbb{R}^{z \times 3M+t}$ are the weights of the hidden layer. Similarly to the NLSE model, the hidden layer imposes a dimensionality reduction of the representations space which forces the model to learn more discriminative features. In this case, the projection also allows the model to capture the interactions between the content and user representations. See Figure 5.2 for an illustrative schematic depicting this approach.

5.3.2 Evaluation: Sarcasm Detection

We evaluated the CUE-CNN model over a sarcasm detection task with the experimental setup and (a subset of) the dataset described by Bamman and Smith (2015). The dataset consists of tweets with self-declarations of sarcasm, i.e., a tweet is considered sarcastic if it contains the hashtag `#sarcasm` or `#sarcastic` and deemed non-sarcastic otherwise.[4]

[4] Note that this is a form of noisy supervision, as not all sarcastic tweets will be explicitly flagged as such.

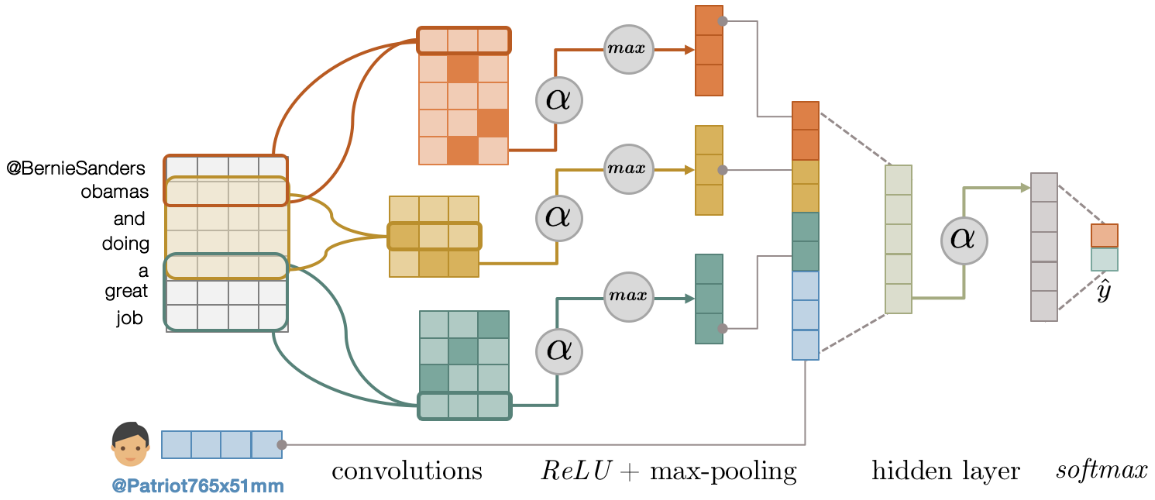


Figure 5.2: Illustration of the CUE-CNN model for contextualized inferences over social media.

To comply with Twitter terms of service, we were given only the tweet ids along with the corresponding labels and we also did not have access to the historical user tweets collected by Bamman and Smith, thus we had to collect the data ourselves. However, by the time we tried to retrieve the data some of the messages and users were no longer available, therefore we discarded messages for which no contextual information was available. Then, we scrapped up to the 1000 most recent *historical tweets* per user[5], resulting in a corpus of 11,541 tweets involving 12,500 unique users (authors and mentioned users).

Regarding the baselines, we evaluated the performance of a content based Convolutional Neural Network as described by Kim (2014) (**CNN**) as well as the **NLSE** and **BOE** baselines used in the previous experiments. We also reimplemented Bamman and Smith (2015)'s sarcasm detection model consisting of a logistic-regression classifier that exploits rich feature sets to achieve strong performance. These are detailed at length in the original paper, but we briefly summarize them here:

- **local-features**, encoding attributes of the target tweet text, including: uni- and bi-gram bag of words (BoW) features; Brown et al. (1992) word clusters indicators; unlabeled dependency bigrams (both BoW and with Brown cluster representations);

[5] The original study Bamman and Smith (2015) was done with at most 3,200 historical tweets. It should also be noted that in some cases our historical tweets were posted *after* the ones in the corpus used for the experiments.

part-of-speech, spelling and abbreviation features; inferred sentiment, at both the tweet and word level; and 'intensifier' indicators.

- **author-features**, aimed at encoding attributes of the author, including: historically 'salient' terms used by the author; the inferred distribution over topics[6] historically tweeted about by the user; inferred sentiment historically expressed by the user; and author profile information (e.g., profile BoW features).
- **audience-features**, capturing properties of the *addressee* of tweets, in those cases that a tweet is directed at someone (via the @ symbol). A subset of these, duplicate the aforementioned author features for the addressee. Additionally, author/audience interaction features are introduced, which capture similarity between the author and addressee, w.r.t. inferred topic distributions. Finally, this set includes a feature capturing the frequency of past communication between the author and addressee.
- **response-features**, for tweets written in response to another tweet. This set of features captures information relating the two, with BoW features of the original tweet and pairwise cluster indicator features, which the encode Brown clusters observed in both the original and response tweet.

We emphasize, however, that implementing these features took considerable time and effort, thus motivating our approach to effectively induce contextually-aware representations without manual feature engineering. Finally, we conducted ablation studies to ascertain the relative contributions of specific components of our model.

5.3.3 Experimental Setup

We first preprocessed the training data and induced word and user embeddings as described in Section 5.2.2. Similarly to other deep neural networks, our model can be operationalized with different architectures and hyperparameter configurations which strongly influence the models behavior (and performance). However, finding the optimal hyperparameter settings would require an extensive search over a very large configuration space. Therefore, following Zhang and Wallace (2015)'s recommendations, we focused our search over combinations of dropout rates $D = [0.0, 0.1, 0.3, 0.5]$, hidden layer sizes $Z = [25, 50, 75, 100]$, filter heights $H = [(1,3,5),(2,4,6),(3,5,7),(4,6,8),(5,7,9)]$, and number of feature maps $M = [100, 200, 400, 600]$.

[6] The topics were extracted from Latent Dirichlet Allocation Blei et al. (2003).

The experiments were conducted with a 10-fold cross-validation methodology, keeping the same partitions across all the baselines. For each partition, the training split was divided into 80%/20% train/validation sub-splits to facilitate hyper-parameter selection and early-stopping. We performed grid-search to choose the best ℓ_2 regularization coefficient, over the range $C = [1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1, 10]$ for the linear models; and randomly sampled (without replacement) $n = 50$ hyperparameter configurations for our model. The model parameters were estimated by minimizing the cross-entropy error between the predictions and true labels, the gradients w.r.t to the network parameters were computed with *backpropagation* (Rumelhart et al., 1988) and the model weights were updated with the AdaDelta rule (Zeiler, 2012).

5.3.4 Results

In Figure 5.3, we show the performance of linear classifiers with each of the manually engineered feature sets proposed by Bamman and Smith (2015), along with scores reported in the original paper. These results confirm Bamman and Smith (2015) findings about the need for contextual information for sarcasm detection and validate our reimplementation of their model. Despite the slight differences caused by the fact the we are using a different dataset, we observe the same general trends: namely, that including contextual features significantly improves the performance, and that the biggest gains are attributable to features encoding information about the authors of tweets.

The main classification results are shown in Figure 5.4. Once again, we find that modeling the context (in this case, solely the author) of a tweet yields significant gains in accuracy. The difference is that here the model jointly *learns* appropriate user representations, lexical feature extractors and, finally, the classification model. Interestingly, we observed that our proposed model not only outperforms all the other baselines, but also shows less variance over the cross-validation experiments which attests to its robustness. The ablation studies showed that pre-training the user embedding improves model performance, which was expected and in line with the literature. We further observed that it is beneficial to introduce a hidden layer capturing the *interactions* between the context (i.e., user vectors) and the content (lexical vectors). This is intuitively agreeable: the recognition of sarcasm is possible when we jointly consider the speaker and the utterance at hand. Finally, we note that Bamman and Smith also reported gains by including information about the audience of a post (albeit small), which suggests that further improvements can still be realized by this model.

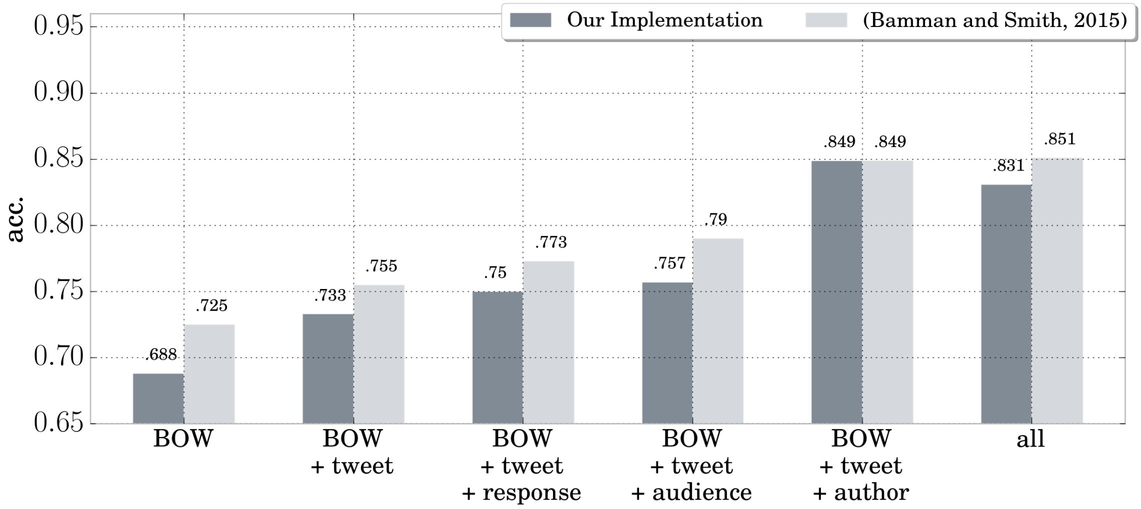


Figure 5.3: Predictive performance of linear classifiers with Bamman and Smith (2015) contextual features for sarcasm detection. We include the original results as a reference, and note that the discrepancies between their reported results and those we achieved with our re-implementation reflect the fact that their experiments were performed using a significantly larger training set and more historical tweets than we had access to.

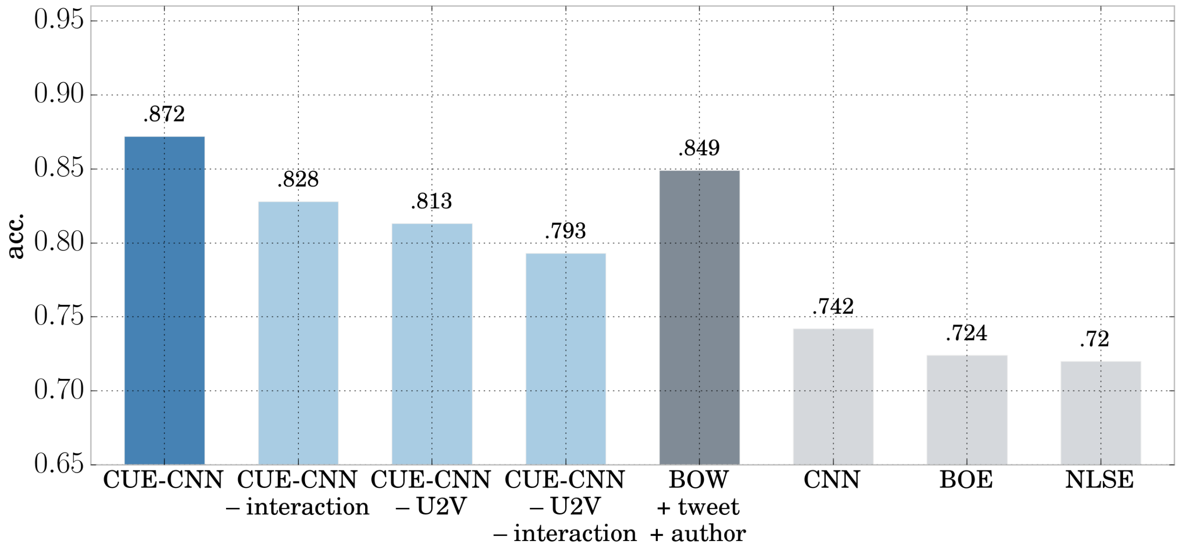


Figure 5.4: Predictive performance of neural models for sarcasm detection and ablation studies for the CUE-CNN model. We compared simple neural architectures that only consider the lexical content of a message with architectures that explicitly model the context (i.e. the author).

5.3.5 Personalized Predictions

To better understand the influence of contextual information on the predictions of the CUE-CNN model, we measured the responses to the same textual content with different hypothetical contexts (authors). We specifically selected two examples that were misclassified by a simple CNN operating only on the contents of a post and ran them through our CUE-CNN with three different user embeddings (that is, simulating scenarios where the posts were written by different users). In Figure 5.5, we show these examples along with the predicted probabilities of the post being sarcastic when: no user information is considered and when different authors are considered. We can see that the predictions drastically change when contextual information is available and that two of the authors trigger similar responses on both examples. This provides evidence that our model captures the intuition that the same utterance can be interpreted as sarcastic or not, depending on the speaker, and also hints at the fact that the model can exploit user similarities to inform predictions (which might help with the generalization of the model). This ability to tailor the responses to an input based on the author opens the door for *personalized models* that combine general knowledge about a phenomena with the peculiar characteristics of a specific individuals, e.g. to support applications delivering personalized/specialized health-care.

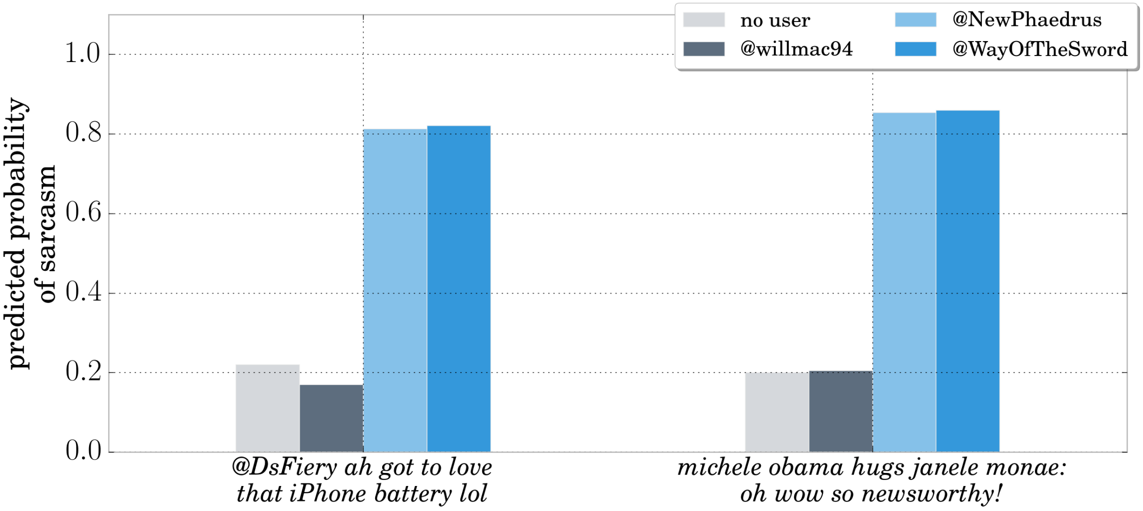


Figure 5.5: Two sarcastic examples that were misclassified by a simple CNN (`no user`). Using the CUE-CNN with contextual information drastically changes the model's predictions on the same examples.

5.4 User Embedding Analysis

Thus far, we have seen that neural user embeddings provide an effective approach to exploit unlabeled data and can directly inform user-level models or complement document-level models with contextual and personal information. In this section, we examine these representations in more detail. We know that word embeddings encode latent semantic attributes of words and express word similarities as vector distances — indeed, we already leveraged these properties explicitly to expand subjective lexicons. But what kinds of information are being captured by user embedding representations? It would be reasonable to expect that they capture latent *personal* attributes, in which case vector similarities could be interpreted as encoding *homophilic* relations between users, naturally implying that similar users should have similar representations. The interpretation of such relations largely depends on the specific domain or application but this mechanism in itself could enable interesting applications. For example, epidemiology systems could leverage user embeddings to recognize and characterize risk groups on social media by identifying individuals that 'look' like other patients known to be affected by some illness[7] (e.g. depression). This would in turn potentially enable scalable, real-time estimates concerning the prevalence of specific public health issues in particular sub-populations.

As we have seen before, one way to (indirectly) infer the latent properties encoded in embedding spaces is to project the respective the high-dimensional vectors into a two-dimensional space. Visualizing the structure of the resulting space (i.e. the relative positions of the vectors) provides clues about those properties, e.g. if the vectors of users affected with some mental illness are close, then they must have similar values along some dimensions, which implies that these dimensions carry information about said condition. While this approach offers a quick and intuitive way of discovering qualitative properties of those spaces, it also has several limitations: first, projecting high-dimensional objects into a lower-dimensional space always implies a loss of resolution, which might lose nuanced information; second, dimensionality reduction algorithms have specific assumptions and parameters which can introduce external biases or artifacts, thereby influencing the results; and third, this approach can not be used to quantify the relations between objects, nor objectively compare different embedding algorithms[8]. Therefore, we investigated the embeddings ability to capture homophily with a more

[7] In some cases this might be more informative than a hard classification

[8] These same limitations motived researchers to find better ways to evaluate unsupervised word embeddings, e.g. with word analogy tasks (Mikolov et al., 2013c)

systematic and quantitative approach based on explicit measures of vector similarity.

5.4.1 Latent Personal Aspects

Figure 5.6 shows a T-SNE projection (Van der Maaten and Hinton, 2008) of the embeddings induced in Section 5.2 for the mental-health experiments, colored according to the respective cohort. We observe that even generic embeddings induced without any label information can, to some extent, capture mental-health related information. Notice that at a local level, the user vectors tend to be surrounded by others corresponding to users in the same cohort. Moreover, we can see that the tailored capture more fine-grained information thus yielding much more discriminative representations with respect to the target labels (i.e. mental health condition).

In Figure 5.7 we show plot the user embeddings induced in Section 5.3 for the sarcasm detection experiments, colored according to their apparent political leaning and according to their interest in sports. These attributes were inferred using the Twitter accounts that a user follows, as a proxy. For the former, we considered users that follow at least one the following (democrats) accounts: *@BarackObama*, *@HillaryClinton* and *@BernieSanders*; or one of the following (republicans) accounts: *@marcorubio*, *@tedcruz* and *@realDonaldTrump*. For the latter, the 500 most popular accounts (according to the authors in our training data) were manually inspected and 100 sports related accounts were selected, e.g., *@SkySports*, *@NBA* and *@cristiano*. We can observe that users with similar political leanings (or that are similarly interested in sports) tend to have similar vectors. As we can see, these representations are able to capture a wide range of personal attributes from private mental states to personal preferences and affiliations.

5.4.2 Measuring Homophily

To investigate and quantify the extent to which learned user embeddings are able to capture homophily relations, we proceeded as follows. First, we considered each user in the corpus in turn as a ‘query’ with which to retrieve similar users; then, we calculated cosine similarities between the query user vector and the vectors associated to all other users, thereby inducing a similarity-based ranking. Intuitively, we would hope to see that individuals in the same mental health categories as the query user are comparatively similar to one another: e.g., users affected by depression should be most similar to other users also suffering from depression.

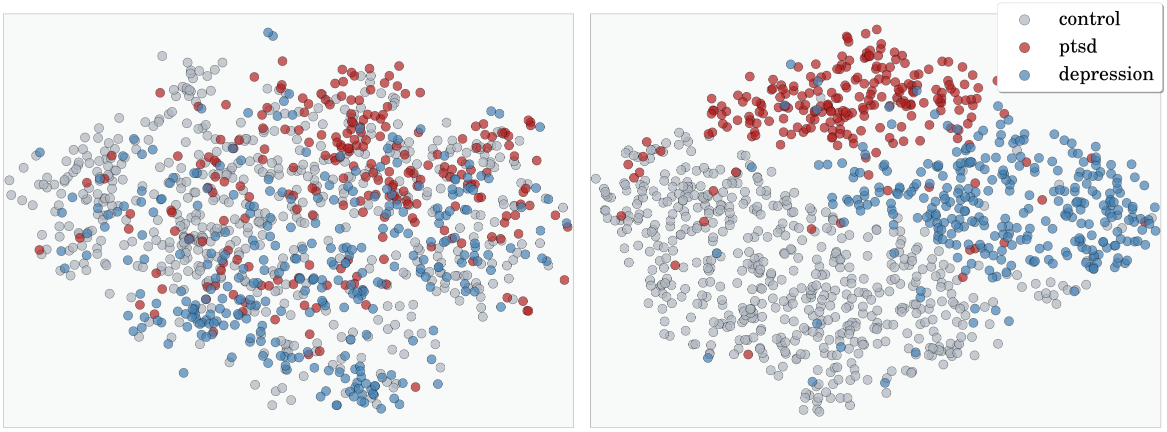


Figure 5.6: User embeddings projected into two dimensions, and colored according to mental-health status. The plot shows the U2V embeddings on the right-hand side and U2V_{sub} on the left-hand side.

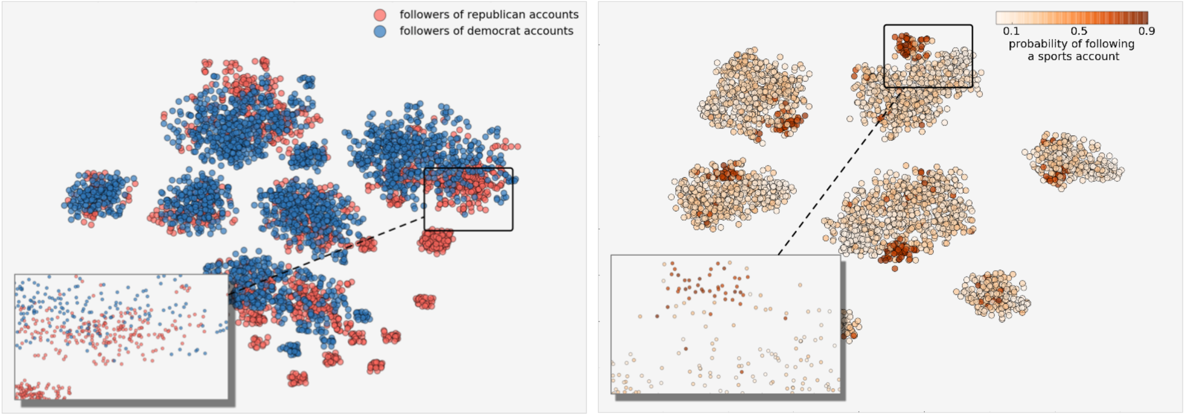
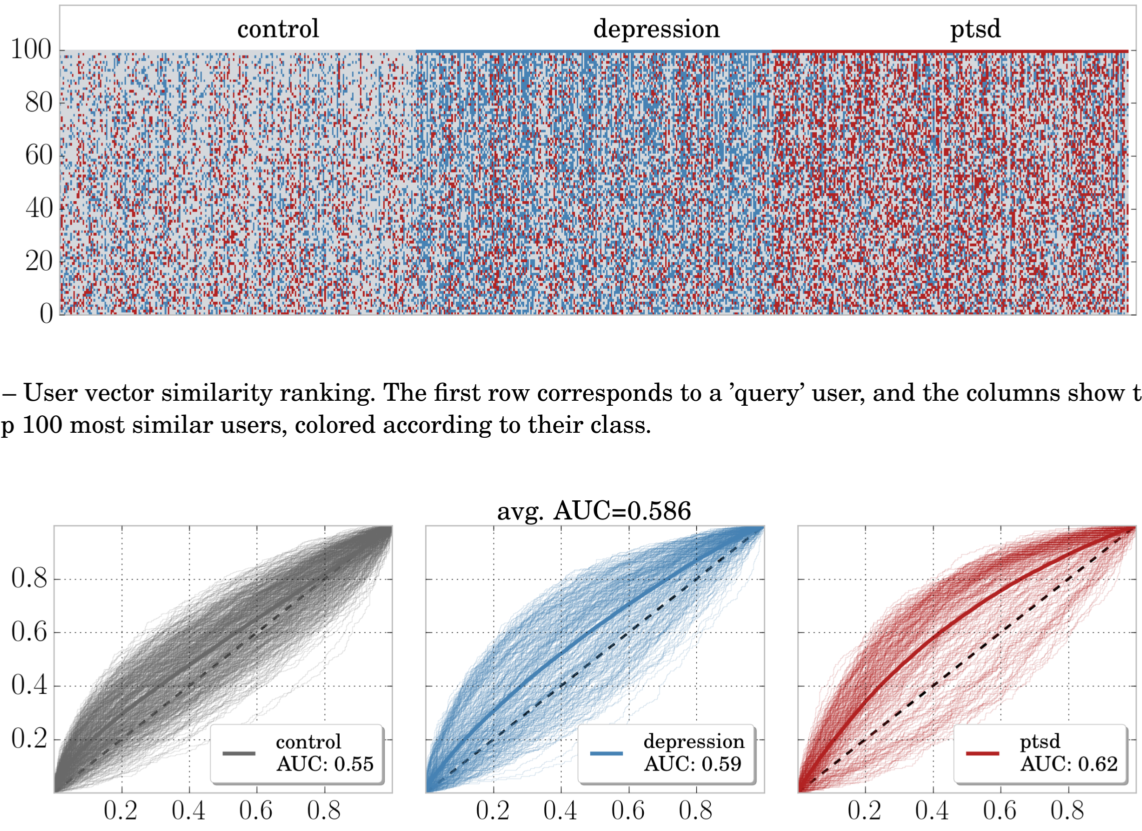


Figure 5.7: User embeddings projected into two dimension, and colored according to personal preferences. The left-hand side plot shows the vectors of users colored according to the politicians they follow on Twitter: the blue circles represent users that follow at least one account from a democratic politician; the red circles correspond to users that follow at least one account of a republican politician. Users that follow both parties were excluded. On the right-hand side, we show a plot of user vectors colored with respect to the likelihood of following a sports account. We discarded users for which the probabilities lied in the range between 0.3 – 0.7 to emphasize the extremes.

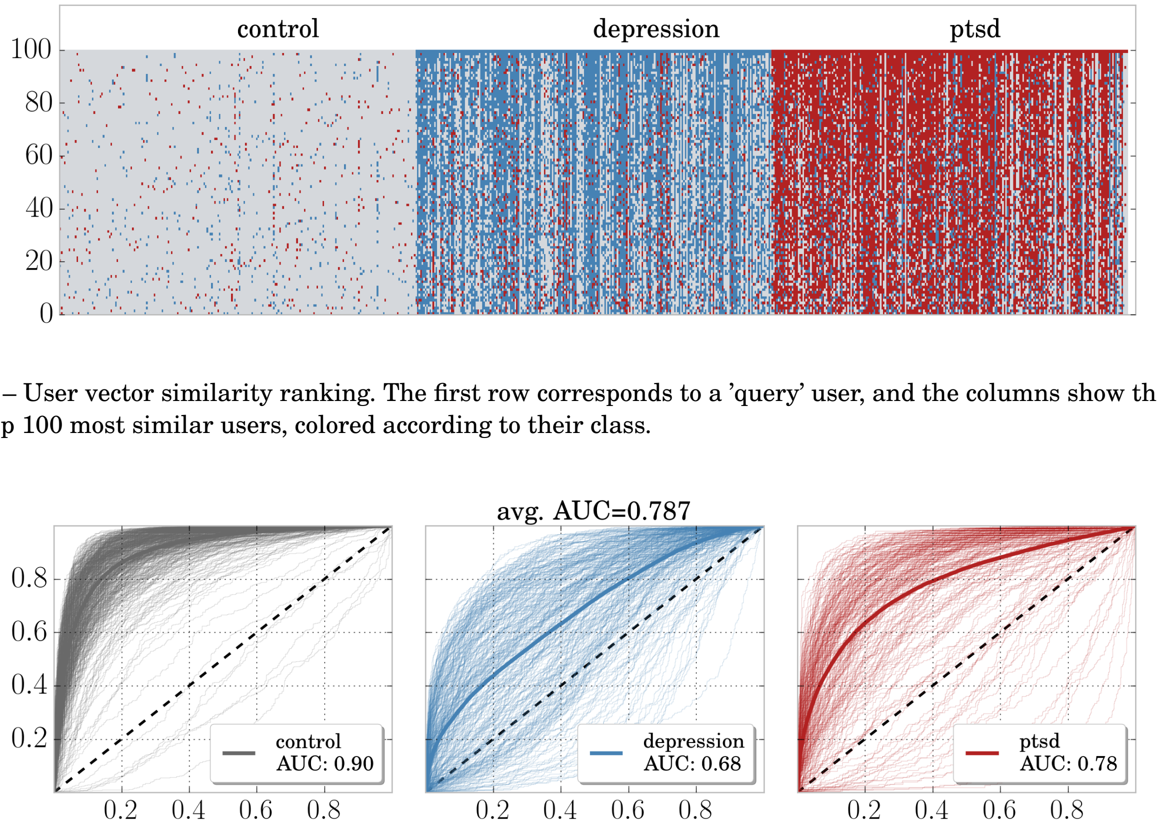
We induced these rankings for all the user embedding models used in Section 5.2.2 (i.e., PV-DBOW, PV-DM and U2V) and evaluated them with respect to the Receiver Operating Characteristic (ROC) curves and the average Area Under the Curve (AUC). We found that all the embeddings perform significantly better than chance; the rankings obtained average AUC scores of 0.56, 0.57 and 0.59, respectively. These results suggest that learned user embeddings do indeed capture signals relevant to mental health, if only weakly. Nonetheless, we believe this is somewhat surprising, given that these are unsupervised, generic representations that were in no way explicitly trained to capture attributes of mental health. Regarding the tailored embeddings, we found that they substantially improve the quality of the induced rankings bumping up the average AUC to 0.69, 0.68 and 0.79, respectively. These refined representations greatly improve performance overall, and particularly with respect to discriminating controls from unhealthy users, suggesting that the induced subspace captures more fine-grained signal related to mental statuses. Figure 5.9 and Figure 5.8 present the results obtained with the U2V vectors. The top plots show the induced similarity rankings, where each



a – User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

b – ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

Figure 5.8: Measuring homophilic relations with respect to mental-health status with vector distances over the (generic) user embedding space



a – User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

b – ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

Figure 5.9: Measuring homophilic relations with respect to mental-health status with vector distances over a domain-specific embedding subspace

column includes the top $k = 100$ most similar users to the query user (first row), colored according to the respective class. The bottom plots show the respective ROC curves under the induced rankings.

5.5 Conclusions

This chapter extended our low-resource neural networks to user-level and contextualized models, with novel methods to automatically induce and exploit embedding representations for social media users. To that end, we first introduced *User2Vec*, a neural language model to estimate user embeddings from a user's collection of previous posts. Similarly to word representations, the user embeddings can also be used directly as features to inform predictive models, which allowed us to build specialized user-level models using the NLSE architecture. Unlike other approaches that explicitly exploit the structure of particular social media services, such as the forum where a message was posted or metadata about the users, learning user embeddings only requires their preceding messages — yet, the obtained vectors are able to capture relevant user attributes and a soft notion of homophily. This makes our model easier to deploy over different social media environments. The models were evaluated over a mental-health inference task aiming to discriminate between users diagnosed with depression or PTSD, from demographically matched controls. The results showed that our models outperform linear baselines based on lexical features and other user representations. The experiments also demonstrated that our user embeddings do capture mental health related signals and user similarities with respect to mental health condition, despite being trained without this information. This is in agreement with prior results from the field of psychology, establishing connections between word usage and mental status (Pennebaker et al., 2001). On the other hand, we found that the embeddings can be tailored, with a small amount of task-specific labeled data, to capture more granular information thus improving the quality of downstream models and applications.

Second, we leveraged the user embedding approach to derive a novel family of neural networks to support the deployment of contextualized models without manual feature engineering. The CUE-CNN models jointly learn and exploit representations for the content and the author of a post, thereby integrating information about what was said and who said it. The evaluation conducted over sarcasm detection — a task that crucially depends on contextual information — showed that our models outperform a recently proposed state-of-the-art sarcasm detection model based on an extensive set of hand-crafted features encoding user attributes and other external information. Furthermore, the ability to capitalize on user embeddings to support personalized inferences, opens the door to new deployment strategies for models that can be trained over (potentially large) external datasets and then fine-tuned to specific users by incorporating their personal embeddings. This can find applications even outside the realm of social media analysis, for example, similar methods could be used to adapt conversational agents or personal digital assistants to different users.

Chapter 6

Agile Social Media Analysis

This chapter concerns with the more practical aspects of deploying social media analysis systems in the real world. First, the methods developed in this thesis are transposed into an agile modeling framework to accelerate the development of specialized models for various applications (Section 6.1). Reducing the efforts of model development allows us to build more complex SMA applications. Thus, this framework is leveraged to operationalize a methodology to conduct demographically controlled analyses. To that end, a process is developed to automatically collect, process and organize social media data, such that it can subsequently be used to sample demographically representative *digital cohorts* of social media users (Section 6.2). Conducting studies over these cohorts allows analysts to take into account the biases of social media data, thereby improving the quality of these studies.

The tools and methodologies introduced in this thesis are then validated in a case-study of real-world demographically controlled agile social media analysis. The study consisted of the development of DEMOS, a Digital Epidemiology and Mental-Health Observation System, to track discussions around specific public health issues and monitor the prevalence of mental illnesses over Twitter data (Section 6.3). To evaluate the system, we created a digital cohort of US Twitter users and conducted a series of pilot studies regarding the discussion around public health related topics, such as contraception, pregnancy and heart disease. Then, we used DEMOS to estimate the prevalence of the depression and PTSD in the cohort and investigate how these mental-illnesses affect different demographic groups.

6.1 Agile Modeling Framework

As we discussed before, the major bottleneck of deploying SMA systems is the development of NLP models tailored for specific tasks. To address this problem, the previous chapters developed a methodology to induce low-resource neural networks to ease the induction of specialized NLP models for social media. The key idea is to separate model development in two steps: (i) unsupervised learning of generic neural embeddings for words and users, thereby complementing small labeled datasets with large amounts of unlabeled data; and (ii) inducing task-specific models by extracting tailored representations from the generic embeddings, thus eliminating the need for feature engineering. This allows us to reap the benefits of neural representation learning without paying the usual costs of these methods, thus providing a middle ground between feature engineering and deep learning approaches to build specialized models.

Decoupling the process of learning generic and specialized representations allows the same embeddings to be reused and adapted for different applications with a small amount of task-specific labeled data. It also helps to expedite the development process by allowing analysts to pre-train embeddings offline, potentially with massive datasets, and then leverage them to quickly build tailored models. Word embeddings are one of the key elements of this approach — these can be used directly as the model input but are also instrumental to learn user and document embeddings that serve as input for other models. However, this approach makes no assumptions about what the embeddings represent nor how they were computed, thus providing leeway for various types of models to be built, such as:

- **Document-level** models, which take word embeddings as input and operate over the contents of an entire document (as described in Chapter 3);

- **Word-level** models, which take word embeddings as input and infer attributes of specific words — these models can be used to automatically induce large scale lexicons tailored for social media (as described in Chapter 4);

- **User-level** models, which take user embeddings as input and make inferences with respect to individual users (as described in Chapter 5);

- **Contextualized** models, which combine document and user embeddings, thereby taking into account characteristics of the contents and of the context of a document, to make inferences over ambiguous documents (as described in Chapter 5).

This methodology can be transposed into an agile modeling framework, depicted in Figure 6.1, to accelerate the deployment of SMA applications by:

- addressing the fundamental challenges of processing subjective social media (i.e., the noise, brevity and lack of context);
- reducing manual data annotation and feature engineering efforts;
- reusing and adapting pre-existing linguistic tools and resources.

All the models and tools that were developed to operationalize this framework have been made publicly available[1]. Reducing the efforts of building NLP models for social media will allow analysts to quickly build models for unexpected and time-sensitive applications; or redirect those efforts to build more sophisticated systems for more complex and robust analyses. The rest of this chapter leverages this framework to improve the validity of indicators extracted from social media.

6.2 Demographically Controlled Social Media Analyses

Social media is a valuable data source but it is also inherently noisy and biased. On the one hand, the universe of users of any given social web platform is not a representative sample of the general population (e.g. urban young adults tend to be overrepresented). On the other hand, these platforms serve multiple purposes and host content created by billions of users, some of which correspond to individual citizens but others belong to organizations, brands, media outlets, celebrities and *social bots*, among others. Therefore, the trends or insights gleaned from these data might not be generalizable to the wider public. Yet, the data collection strategies of current SMA methodologies assume that all the data about a topic is equally relevant for the analysis. Moreover, the lack of contextual information about the data can obfuscate the complexities and nuances of the topic of investigation and distort the interpretations. Consequently, current SMA approaches can only capture general trends, but fail to provide insights as to *why* said trends are observed. Fortunately, this problem (known as *selection bias*) is hardly new and hence it can be addressed with well-established techniques that are routinely used in polling and survey-based research. Specifically, probability sampling strategies that

[1] `http://github.com/samiroid/ASMAT`

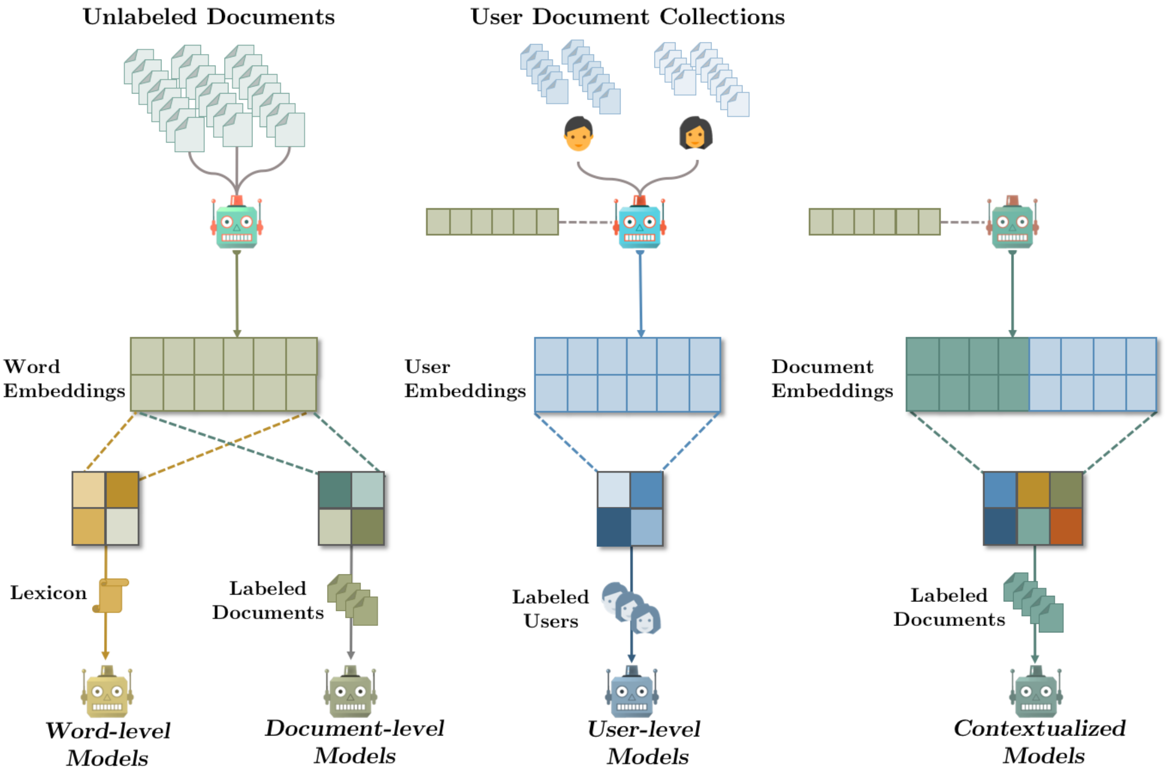


Figure 6.1: Schematic depiction of the proposed agile modeling framework based on low-resource neural networks. The framework exploits large amounts of unlabeled data to learn neural representations, which are then adapted to derive specialized models. These representations can be re-used and shared across various models, thus helping to expedite the deployment SMA applications.

serve as the basis for the validity of survey research, can also be used to improve the validity of social media based studies.

We can replicate universally accepted approaches in survey research by conducting studies over demographically representative samples of social media users, instead of analyzing all the data related to a given topic, as it is typically done. Leveraging representative cohorts sampled *before* conducting the analyses can help to ameliorate both the selection bias and confirmation bias problems. This in turn will allow social media analysts to produce externally valid results concerning national trends, from inherently biased samples and extrapolate findings to a broader population. Similar strategies have already proven to be effective for online surveys, which can have comparable validity to other survey modalities simply by controlling for basic demographic features such as the location, age, ethnicity and gender (Duffy et al., 2005). Furthermore, the demographic information will enable analysts to begin answering questions regarding not just the 'what' but also the 'who' of social media discussions, that is, what types of people discuss which topics, and how these vary by demographics.

6.2.1 Building a Digital Cohort

Herein, we propose a methodology to collect, process and organize social media data, such that it can subsequently be used to sample representative *digital cohorts* for social media studies. This methodology consists of the three following main steps:

1. **Data Collection**: The first step is to create a large database of social media accounts from which to draw candidate cohort members, while excluding accounts that are inactive or that belong to corporations, spammers and bots. Then, for each cohort candidate, collect and store their publicly available profile information and the respective post collection.

2. **Demographic Inference**: The second step entails enriching the candidates profile information with key demographic attributes. While information regarding such attributes on social media platforms might be inaccurate, outdated or simply unavailable, research has shown these can, to some extent, be inferred using statistical methods (Cesare et al., 2017). Therefore, this step requires collecting and annotating data to train models that infer demographic attributes of users, given their collection of posts, which can later be used to annotate the cohort candidates with respect to these attributes.

3. **Cohort Sampling**: The third step consists of using adaptive randomization techniques to select out a representative sample that matches a target demographic distribution, thus minimizing the differences between cohort demographics and the population of interest. These methods have been extensively studied and have been shown to efficiently generate statistically indistinguishable cohorts (Rosenberger et al., 2012).

This process can be easily automated to scale to arbitrarily large datasets and keep the cohorts up-to-date with minimal maintenance efforts. As a result, the makeup of a cohort can be easily changed at any given time to either improve representativeness or respond to cohort changes, e.g. detecting when an account is closed and replacing these members with others with similar demographic characteristics. Moreover, there are no additional costs in adding more accounts or including difficult to reach groups (e.g. young adults), meaning that the costly and crippling problems of maintenance and attrition in panel surveys can be dealt with trivially. In this respect, this approach is a major departure from the way surveillance systems are usually maintained, where tremendous resources are spent on locating, adding and maintaining study participation.

6.3 Case Study: Digital Epidemiology and Mental-Health Observation System

This thesis aimed to address the main limitations of current approaches to deploy social media analysis systems. The methods and techniques proposed thus far can now be leveraged to improve over traditional SMA methodologies (as described in Section 2.4). Thus, here we investigate the impact of conducting *demographically controlled analyses with specialized models*, as described in Table 6.1. To validate this approach, this section introduces a case-study of real-world social media analysis applied to public-health. The main goals of this study are two-fold: first, assess the impact of leveraging demographic information in social media studies; second, provide a proof concept of a SMA system built with the methods introduced in this thesis. These methods were operationalized in the development of DEMOS, a *Digital Epidemiology and Mental-Health Observation System*, designed to track discussions around specific public health issues, and monitor the prevalence of mental illnesses over Twitter data. The system provides a platform for the rapid deployment of custom inference models allowing social media analysts to conduct externally valid studies, vis-a-vis national trends or with focus on specific demographic groups (e.g. urban black females). To evaluate the system we conducted two pilot studies of demographically controlled public health surveillance over Twitter — the first, measuring how different demographic groups engage with different health related issues; and the second, investigates how mental illnesses affect different demographic groups.

6.3.1 System Architecture

DEMOS was operationalized as a highly scalable, distributed data collection and processing platform, implemented with state-of-the-art open-source packages for big data analytics and deployed with cloud computing solutions. The system consists of two main components: (i) a **Digital Cohort Sampler** module to collect and organize social media data in a manner that supports the creation of demographically representative cohorts;

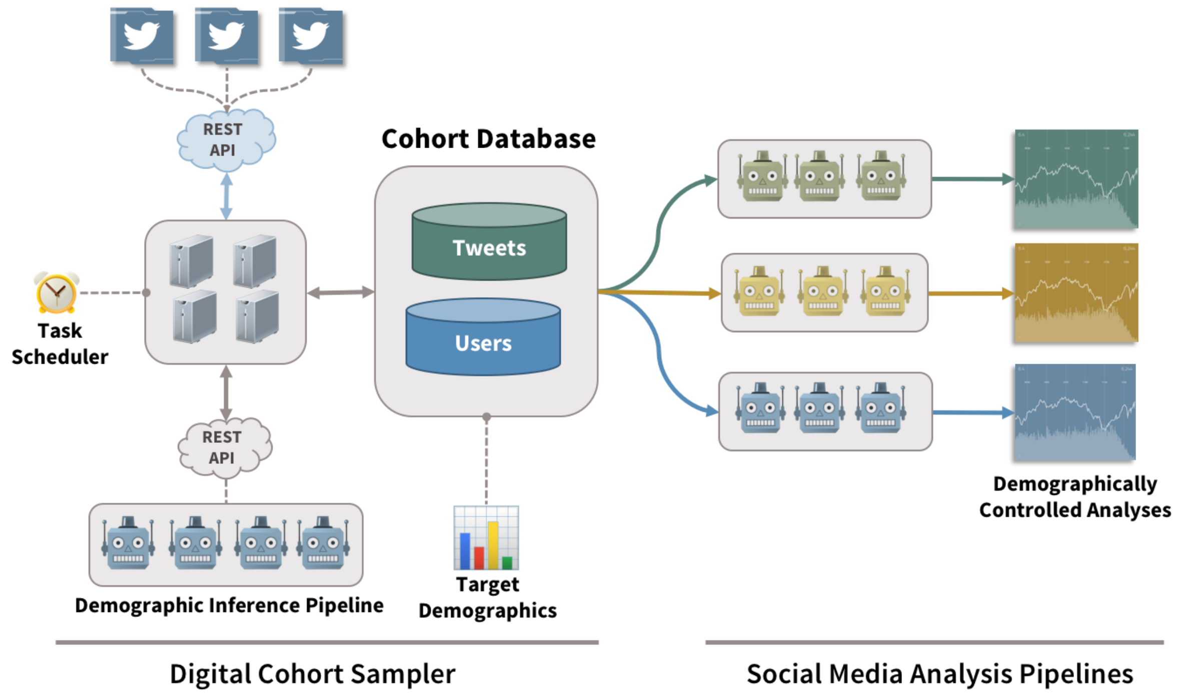


Figure 6.2: Schematic depiction of DEMOS

and (ii) a set of **Social Media Analysis Pipelines**, which leverage natural language processing methods to extract relevant properties of the data. Each pipeline consists of a set of specialized models, deployed as a standalone RESTful web applications. This allows the pipelines to be invoked asynchronously and the data to be processed in parallel. In the current implementation DEMOS includes two such pipelines:

- **Topic Detection Pipeline**, consisting of a set of lexicon-based models built with lexicons created by experts to identify posts discussing public health issues;

- **Mental-Health Inference Pipeline**, composed of a set of bespoke classifiers that given a collection of user posts, estimate the probability of the user being affected by a mental disease.

This modular architecture, depicted in Figure 6.5, enables analysts to easily tailor the system for different of applications simply by plugging in (or removing) specific inference models or pipelines, and allows the system to scale to large data volumes by adding computation nodes.

Data Collection	
Goals	Sample a demographically representative digital cohort of social media users using the process described in Section 6.2
Validation	The validation can be done automatically by specifying the target number of samples and demographic composition, and then monitoring the data collection process
Input	Target sample size and demographic composition for the cohort
Output	Demographically representative digital cohort of social media users

Model Development	
Goals	Build analytics pipelines to support demographic inference and task-specific inferences. The low resource neural networks described in the previous chapters can be used to quickly build specialized models for document-level (Chapter 3), word-level (Chapter 4) and user-level (Chapter 5) inferences
Validation	Measure model performance on a held-out dataset
Input	Training data for each model
Output	Analytics pipelines based on specialized models

Data Analysis	
Goals	Leverage demographic information to break down the results of the analyses
Validation	Manual inspection and interpretation of the trends gleaned from the analysis
Input	Raw signals inferred by the models
Output	Plots showing stratified aggregations of the inferred signals

Table 6.1: Methodology for demographically controlled social media analyses with specialized models

6.3.1.1 Digital Cohort Sampler

The Digital Cohort Sampler is responsible for implementing the cohort sampling methodology described in Section 6.2. This module consists of a database and a set of computation nodes coupled with a distributed task management system, which allows computations to be spread across a collection of servers. These nodes are responsible for executing tasks such as crawling, processing and storing the data. The system is managed by a task scheduler that invokes computation tasks periodically, which allows it to run continuously and keep the datasets up-to-date with little manual maintenance. This also allows the system to operate asynchronously and automatically detect and recover from data inconsistencies or data loss, without direct communication/synchronization between different components.

The data is collected by invoking the Twitter REST API, which returns tweets as JSON-encoded objects containing the text and other metadata, including the information about the author. The retrieved data is then stored using *MongoDB*, a free and open-source document-oriented database management system. MongoDB was chosen because it can naturally store JSON objects, scale to large datasets by distributing the data across multiple servers, and it is schema-free, meaning that arbitrary information can be added at any time (e.g. if new analytics are included or the existing models are updated). The data is organized in two separate MongoDB collections[2]: a **Tweets** collection, storing the raw tweet data; and a **Users** collection, storing the user information obtained from Twitter as well as the annotations produced by the demographic inference pipeline.

The demographic inference pipeline, leverages automated methods to estimate labels for the **location**, **gender**, **age** and **race/ethnicity**, of each candidate cohort member. The location was inferred using *Carmen*, an open-source library for geolocating tweets that uses a series of rules to lookup location strings in a location knowledge-base (Dredze et al., 2013). The other attributes were inferred with bespoke supervised classifiers induced with the agile modeling framework described in Section 6.1. The standard formulation of race and ethnicity is not well understood by the general public, so categorizing social media users along these two axes may not be reasonable. Instead, we used a single measure of multicultural expression that includes five categories: *White*, *Asian*, *Black*, *Latino*, and *Other*. Identifying age based on the content of a user can be challenging, and exact age often cannot be determined based on language use alone. Therefore, we used discrete categories that provide a more accurate estimate as to age: *Teenager* (below 19), *20s*, *30s*, *40s*, *50s* (50 years or older).

[2] MongoDB collections are equivalent relational database tables

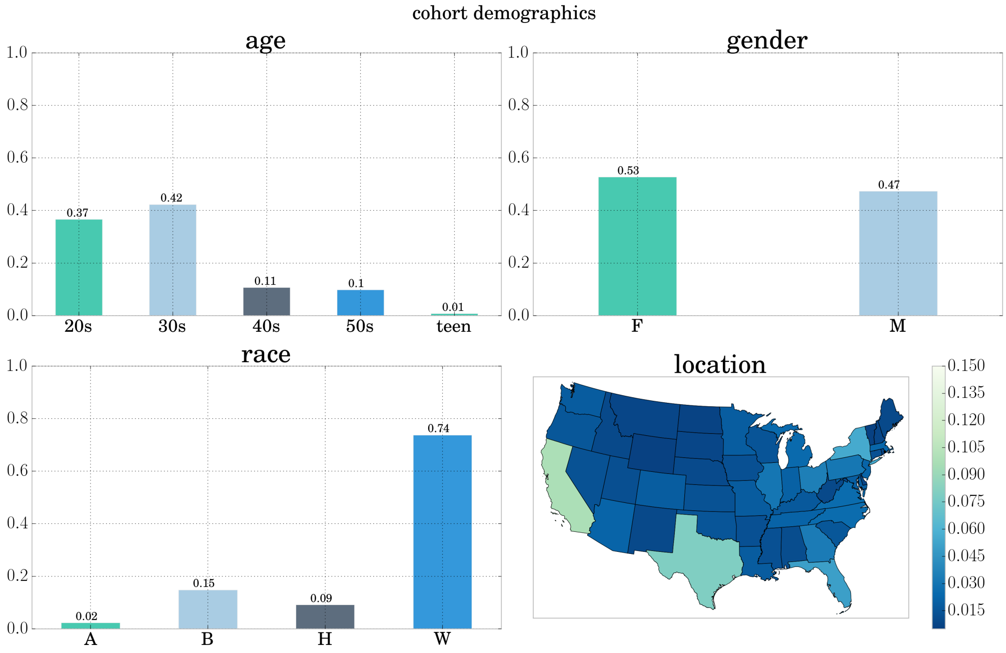


Figure 6.3: Cohort demographics

6.3.2 Data Collection

We leveraged DEMOS to conduct a series of pilot studies of demographically controlled public health surveillance over Twitter data. To that end, we first constructed a digital cohort by randomly sampling a set user accounts that are likely to be from the United States from a stream of geolocated tweets. This resulted in a set of 58K users, which were added to the system as cohort candidates. The data crawling process identified 48K active accounts belonging to users from the US and yielded a total of 120M tweets with an average of 2,068 tweets per user (median of $2,790$ and a standard deviation of $1,241$). The cohort members were then processed through the demographic inference pipeline, resulting in the composition shown in Figure 6.3. We can see that some demographic groups are overrepresented (e.g. white individuals) while others are grossly underrepresented (e.g. teenagers).

6.3.3 Model Development

The demographic inference and mental-health inference pipelines were operationalized with the specialized user-level models described in Chapter 5. For the latter, we trained independent classifiers for depression and PTSD using the mental-health analysis datasets described in the same chapter. Regarding the demographic inference, we used a semi-automated procedure to construct the training dataset. First, we downloaded a set of unique Twitter user profile pictures and respective posts (duplicate images were detected with a min-Hash algorithm). Second, we used a deep Convolutional Neural Network for face identification and classification to infer labels for the age, gender and race[3], which where mapped to the aforementioned categories. Users with images containing more than one face or with no face were discarded, and the rest were associated with their respective post collections. The advantage of this approach is that we can obtain a very large set of labeled data at minimal cost and effort, allowing us to build more accurate models. For this case-study, we collected a dataset of 5k users — Figure 6.4 shows the label distribution for each demographic attribute. We can see that the demographics are highly skewed towards white young adults, which further motivates the need to control for demographic factors when conducting social media based studies.

All the models were trained and evaluated with the usual methodology, using 80%/20% train/test splits computed with stratified sampling to keep the original class proportions. We compared the performance of our models to that of linear classifiers with features based on bag-of-words (BOW), bag-of-word-embeddings (BOE), and user embeddings (U2V). Figures 6.5 and 6.6 present the results in terms of the average F_1 metric for the demographic and mental-health inference models, respectively. We can see that the gender it is easier to predict than the age or race, partly due to a more skewed distribution on the training data and the higher cardinality of the response set. Nevertheless, we observe that our models significantly outperform the other baselines across all the datasets.

6.3.4 Data Analysis

For the first study, we processed all 120M tweets through the topic detection pipeline to identify posts discussing pregnancy and contraception — a post was considered relevant to a topic if it contained at least one word from the corresponding lexicon. We found that 0.06% of the tweets were related pregnancy and 0.01% were about contraception. Then, we measured how the authors of those tweets were distributed with respect to the demographic attributes. These measurements were performed at the tweet level, meaning that users with more than one relevant tweet were counted more than once. Figures 6.7a and 6.7b show how the discussions about pregnancy and contraception

[3] We used the implementation provided by *clarifai*[4], which makes their system available through an API

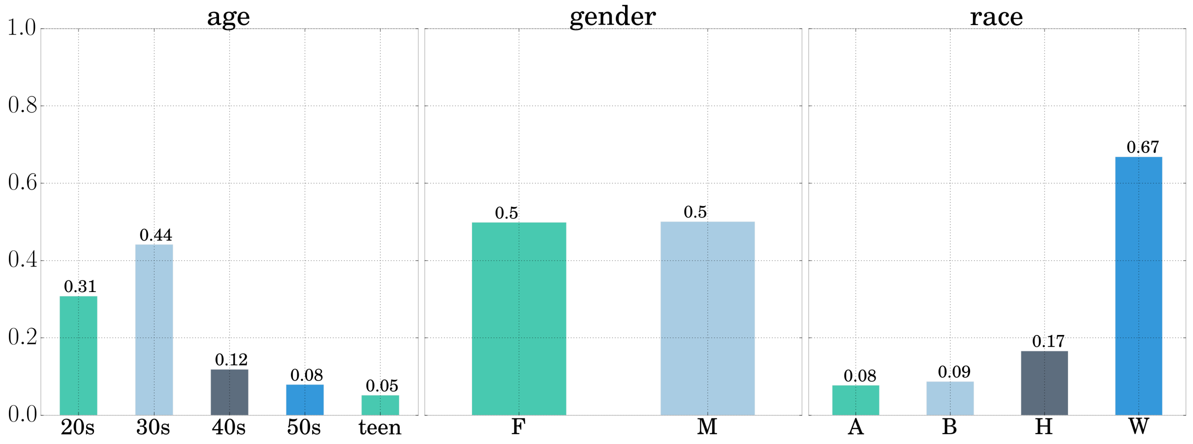


Figure 6.4: Label distribution for the demographic inference training data

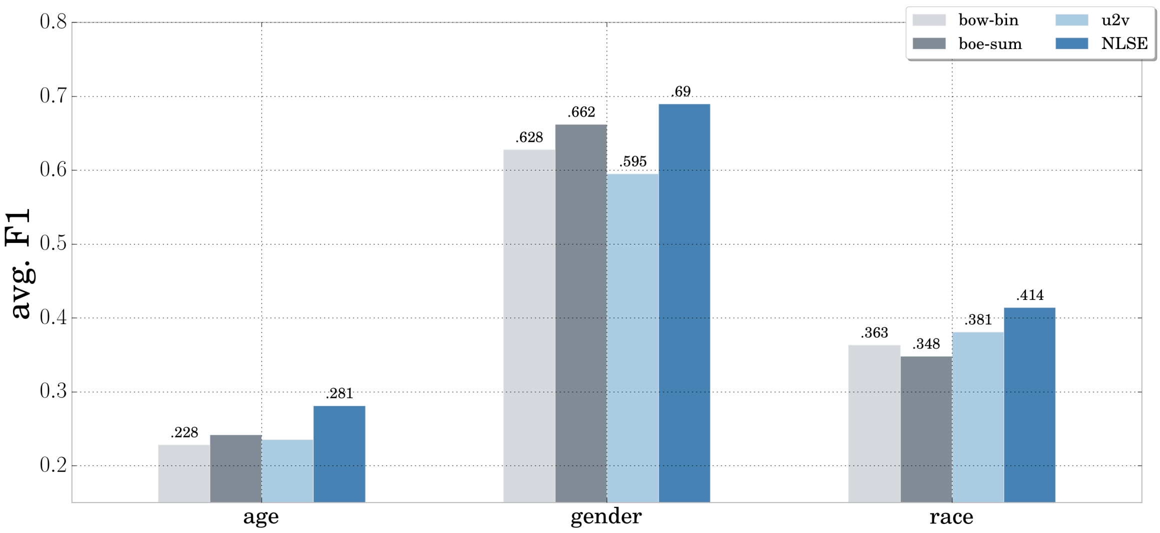


Figure 6.5: Classification performance of demographic inference models with respect to average F_1 metric

vary across different subpopulations. As expected, the distributions are very different as some topics are more pertinent than others to different segments of the population, e.g. tweets about pregnancy are much more likely to have been posted by a woman than man. Generally speaking, people seem to be more concerned with pregnancy at younger ages and more about contraception later in life. We can also see that Blacks talk more about pregnancy than contraception while the opposite is true for Hispanics. The most salient geographical pattern is with Florida, where contraception is much more discussed than pregnancy. These simple observations may not be particularly surprising to subject matter experts, but they do illustrate how demographic information can improve the

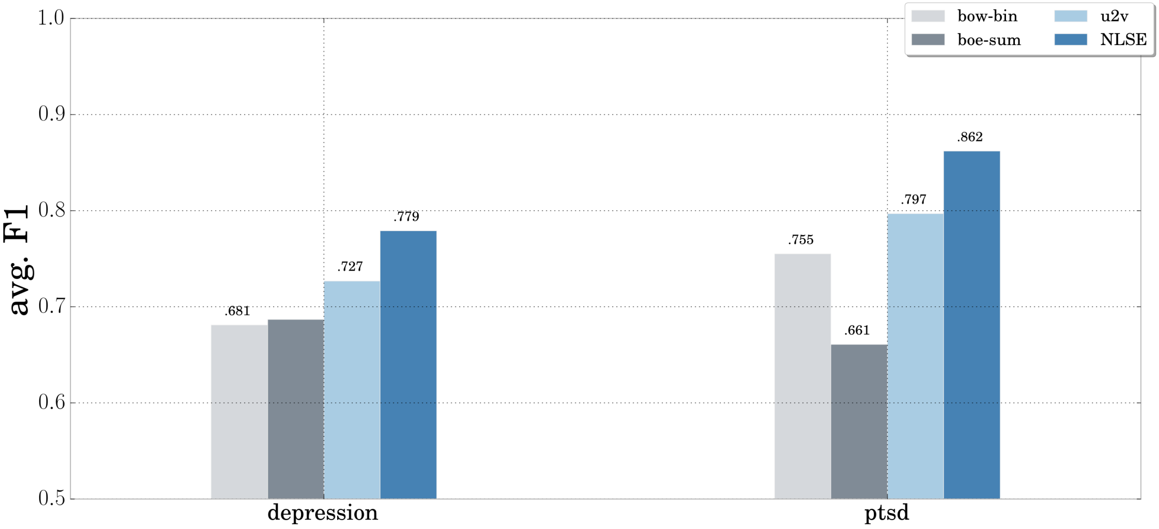


Figure 6.6: Classification performance of mental-health inference models with respect to average F_1 metric

interpretability of results obtained even with simple methods.

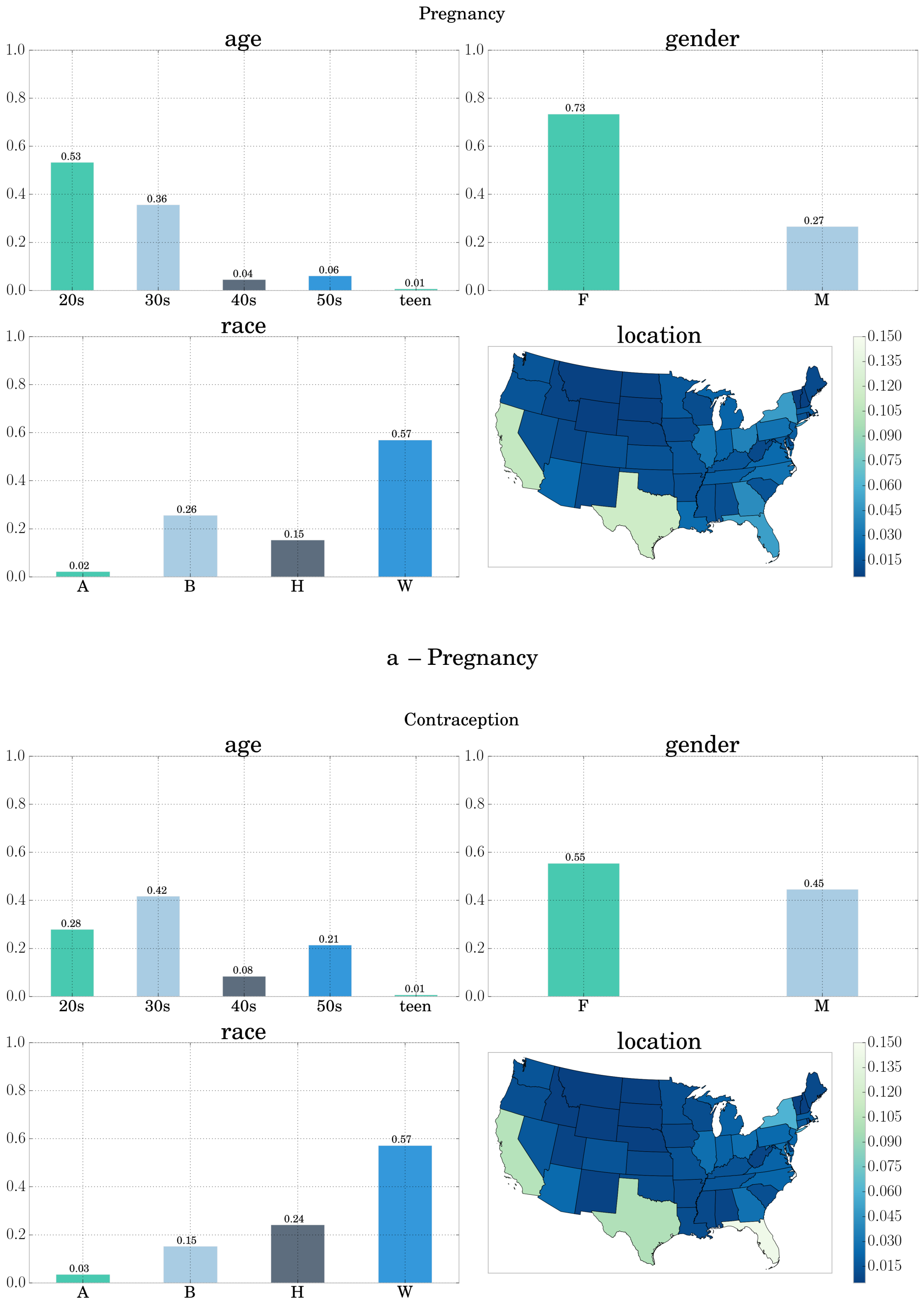
Estimating the Prevalence of Mental Disorders

For the second study, we processed the cohort through the aforementioned mental-health inference pipeline to estimate the prevalence of depression and PTSD, and examine how these illnesses manifest across the population. We found that 30.2% of the users are likely to suffer from depression, 30.8% from PTSD, and 20% from both. The first thing to note is that these estimates are much higher than the current statistics found on the NIH website, that report a prevalence of 6.7% for depression[5] and 3.6% for PTSD[6]. However, it should be noted that those statistics are outdated — the depression estimates are from 2015 and the PTSD estimates are from 2003 — and that this cohort is not representative of the US population. We also observe a significant overlap between people affected by depression **and** PTSD, which is not surprising given that the comorbidity between these disorders is well-known, with approximately half of people with PTSD also having a diagnosis of major depressive disorder (Flory and Yehuda, 2015).

But how do these conditions affect different parts of the population? To investigate this question, we looked at how the demographics of the unhealthy people diverge from

[5] `https://www.nimh.nih.gov/health/statistics/major-depression.shtml`

[6] `https://www.nimh.nih.gov/health/statistics/post-traumatic-stress-disorder-ptsd.shtml`



a – Pregnancy

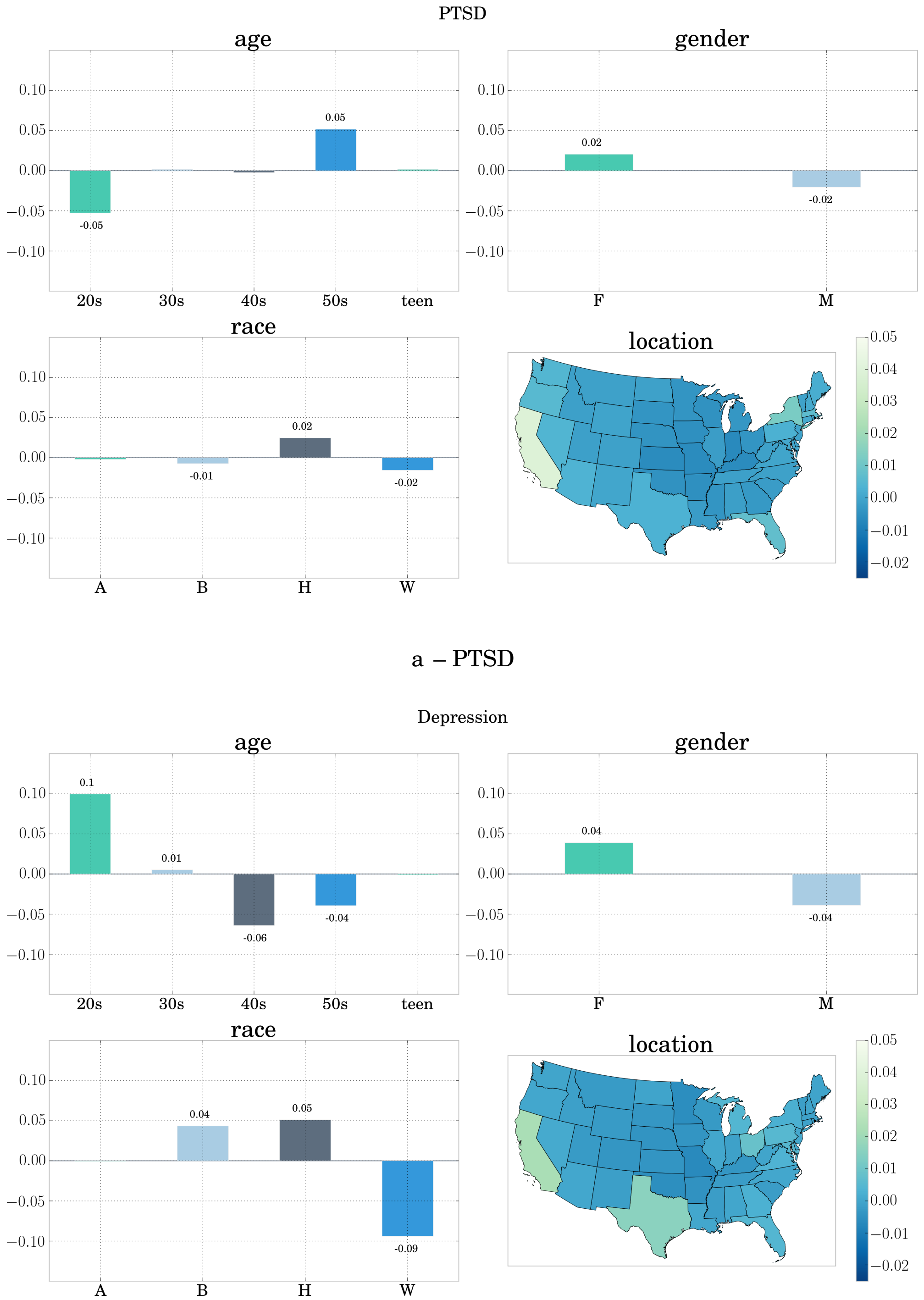
b – Contraception

Figure 6.7: Demographically controlled public health surveillance

those of the cohort. Figures 6.8b, 6.8a and 6.9 show the estimates for depression, PTSD and both, controlled for the cohort demographics. We can see very apparent generational differences — PTSD seems to be more prevalent amongst older people whereas depression affects predominantly younger people —, but we observe that, in both cases, women are more susceptible than men. Interestingly, these estimates correlate with the trends reported by the NIH, but when it comes to race, our findings disagree. We observe that Blacks and Hispanics are more likely to be affected by mental illnesses, whereas the NIH reports a higher prevalence amongst Whites. One possible reason for these disparities is that racial minorities are more likely to come from communities with lower education rates and socioeconomic status (SES), and to be in a position where they lack proper health coverage and mental-health care. Reports from the NIH and other US governmental agencies show that 46.3% of Whites suffering from a mental-illness were subjected to some form treatment, but this was case for only 29.8% of Blacks and 27.3% of Hispanics[7].

The high costs of treatments compounded with the stigmas and misconceptions about mental-health can discourage people from more vulnerable communities to seek help in these cases. While this has been the prevailing view on the racial disparities in health, recent studies show that the factors such as discrimination and perceived inequality have a stronger influence on mental-health than it was previously supposed, even when controlling for the SES (Budhwani et al., 2015). Others have found that acute and chronic discrimination causes racial disparities in health to be even more pronounced at the upper ends of the socioeconomic spectrum. One of the reasons being that for Whites, improvements in SES result in improved health and significantly less exposure to discrimination, whereas for Blacks and Hispanics upwards mobility significantly increases the likelihood of discrimination and unfair treatment, as they move into predominantly White neighborhoods and work environments (Colen et al., 2017). An in-depth analysis of this issue is out of the scope of this work, but these results suggest that it deserves further investigation. A follow-up study to investigate the role of discrimination in mental-health could be conducted by adding a model to identify users who reported that were victims of discrimination and compare the prevalence of mental-illness with a control group.

[7] https://www.integration.samhsa.gov/MHServicesUseAmongAdults.pdf



a – PTSD

b – Depression

Figure 6.8: Measuring the prevalence of mental illnesses

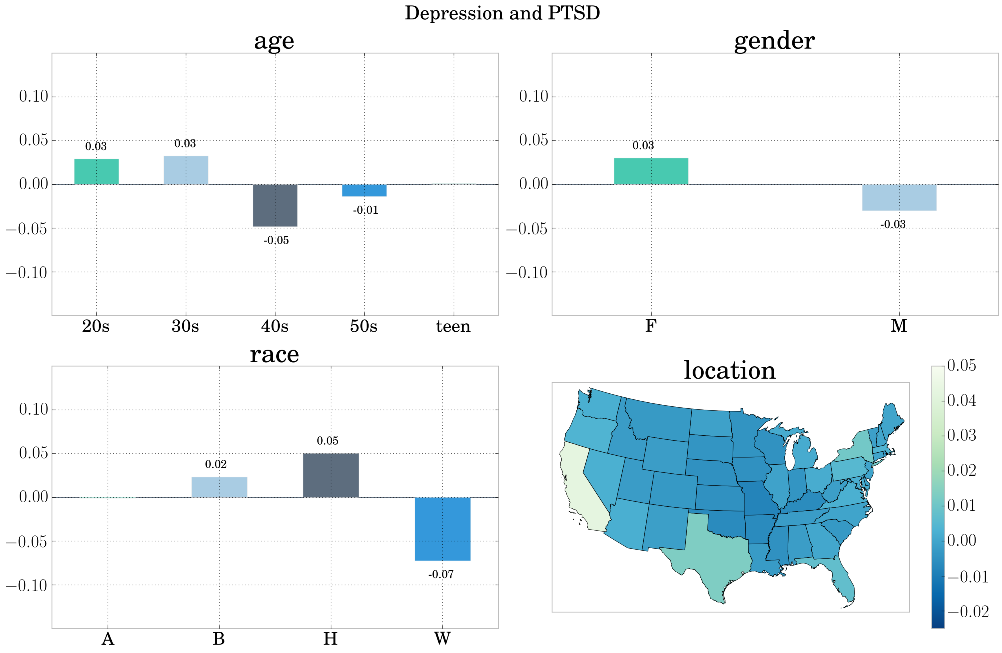


Figure 6.9: Depression and PTSD

6.3.5 Discussion

These pilot studies demonstrated how analyses conducted over arbitrary data samples, devoid from context, can hide the complexity and nuances of the phenomena being studied. In contrast, demographically controlled SMA can allow for more detailed studies and insightful analyses. In a real-word application, the insights gleaned by these pilots could prompt new questions or warrant more thorough investigations. This could involve deploying new models and pipelines (e.g. expanding the current lexicons or training new classifiers) or performing qualitative analyses, e.g. using a topic model to tease out particular "themes" relevant to an issue (e.g. specific contraceptive methods) or looking at the similarities/differences between healthy and unhealthy cohort members with the methodology described in Section 5.4.2. In practice, information about how different subpopulations perceive certain health issues, say contraception, could improve public health policies and inform intervention campaigns targeted for different demographics. This is still an experimental methodology and we should refrain from drawing strong conclusions from the analyses. The system has some noteworthy limitations, namely the fact that the training datasets for both the demographic and mental-health inference were produced in a semi-automated fashion and therefore may contain biases, e.g. self-selection biases from the self-reported mental-health datasets and biases coming from the CNN model that was used to generate the training data. These datasets are also very imbalanced meaning that the models may be better at predicting some categories than others. Nevertheless, the fact that some of our estimates correlate with statistics obtained through traditional methodologies suggests that this might be a promising approach to complement current epidemiology practices.

6.4 Conclusions

This chapter distilled the methods developed in this thesis into an agile modeling framework to reduce the efforts of deriving specialized models for SMA. Then, it demonstrated how reducing the efforts of model development allows for the deployment of more sophisticated systems. Concretely, the modeling framework was leveraged to operationalize a methodology to support the creation of digital cohorts of social media users. Conducting analyses over these cohorts allows analysts to ameliorate both sampling and confirmation biases. To validate this methodology, we conducted a case-study of demographically controlled social media analysis applied to the public health domain, specifically regarding: the discussion of heart-disease, pregnancy and contraception; the prevalence of depression and PTSD and how these illnesses affect different demographic groups. To that end, we developed and evaluated DEMOS, a Digital Epidemiology and Mental-Health Observation System, that provides a platform for rapid deployment of custom inference models allowing social media analysts to conduct externally valid studies, vis-a-vis national trends or with focus on specific demographic groups.

To operationalize the system, we developed bespoke specialized user-level models to predict key demographic traits of social media users (i.e. age, gender and race) and the probability of the users being affected by a mental disease — compared to the typical baselines, again we found that our models have a significantly superior performance across all the tasks, further demonstrating the adequacy of the proposed agile modeling framework. DEMOS was evaluated by conducting series of demographically controlled studies over a digital cohort of US based Twitter users. The pilot studies showed that the information about user demographics can indeed enable more insightful interpretation of the analyses by disentangling some of the factors that influence an observation. Moreover, this information helps to make the inherent biases of social media environments more explicit which allows analysts to control for those biases when interpreting the outcomes of a study.

This work presents a step forward from typical SMA systems that leverage arbitrary (and biased) data samples and assume that all instances contribute equally to measure a given phenomenon. Such approaches can only capture general trends and measure how these evolve over time. In contrast, DEMOS can be used to conduct demographically controlled social media analysis for various applications and has the potential to reach and account for vulnerable and underrepresented parts of the population. This opens the door to more responsive epidemiology practices, enabled by the ability to remotely identify at-risk groups and track the progress or the effects of targeted interventions, in near real-time. In turn, this can provide empirical support for increased resource allocation to programs dedicated to preventing and alleviating health related issues, tailored for specific demographics. Here, we focused our attention on public-health applications but this system can also be used for other purposes, e.g. demographically controlled analyses could improve political opinion mining applications and bring these methods closer to well established polling practices. These systems could even go beyond the traditional methods, e.g. pollsters often conduct measurements over samples that match the demographic characteristics of the (predicted) 'likely voters' — however, if the actual voters differ from what was presupposed, then the estimates will be inaccurate. DEMOS on the other hand, could be used to produce multiple estimates based on different assumptions about the effective electorate by performing analyses on different digital cohorts.

Chapter 7

Conclusions

Social media analysts seek to gain insights into current worldly events and circumstances by extracting and aggregating linguistic signals from large volumes of social media data. This new data source holds potential for numerous applications but the noise, brevity and ambiguity of this medium poses challenges to traditional natural language processing approaches. Therefore, analysts are often forced to rely on simplistic and sub-optimal methods or devote significant efforts on the development of specialized models. On the other hand, current SMA practices essentially ignore basic principles of data-driven research, such as acknowledging the inherent biases of the data and leveraging robust sampling strategies to account for those biases. This can distort the insights gleaned by these data and hamper the validity of the studies.

This thesis aimed to address these limitations with two main contributions:

1. A methodology to ease the development of specialized NLP models for noisy text in low-resource settings by: leveraging unlabeled data and pre-existing linguistic resources to reduce manual data annotation efforts; and relying on neural representation learning methods to reduce the efforts of task-specific feature engineering. This approach was then transposed into an agile modeling framework to accelerate the deployment of SMA applications.

2. A methodology to conduct demographically controlled social media analyses that helps to mitigate the biases of social media data. The agile modeling framework was used to derive demographic inference models to predict key user attributes,

such as age, gender and race. These model are used to automatically collect, process and organize social media data, such that it can subsequently be used to sample demographically representative *digital cohorts*. To that end, these cohorts can be leveraged to conduct demographically controlled analyses

7.1 Agile Modeling with Low-Resource Neural Networks

The proposed approach relies on a novel method to derive low-resource supervised neural networks by learning generic embeddings from large amounts of unlabeled data, and then using small labeled datasets to extract task-specific lower-dimensional representations. Chapter 3 introduced the NLSE, a simple neural architecture based on this approach, and showed that it can be used to derive specialized document-level models for noisy and subjective content. Chapter 6 applied the NLSE to derive word-level models, which can be used to improve lexicon-based SMA systems by automatically expanding small handcrafted lexicons. To extend this approach to user-level analyses, Chapter 5 introduced *User2Vec*, a neural language model to estimate user embeddings that, similarly to word embeddings, capture latent user aspects (e.g. political leanings) and similarities. In this case, however, the similarities can be interpreted as a soft notion of *homophily*. The NLSE could then be used to derive specialized user-level models by extracting tailored representations from the generic user embeddings. This approach was further extended by combining document representations and user representations to derive models to support contextualized inferences over highly ambiguous and subjective content, e.g. involving sarcasm. In these cases, the intended meaning of an utterance might also depend on the speaker. Therefore, the model extracts lexical representations of the contents of a document with a Convolutional layer which are concatenated with the authors user embedding. The resulting vector is then transformed through a hidden layer that induces tailored representations for the task, and captures the interactions between documents and authors representations. We refer to these models as *Content and User Embedding Neural Networks*.

The models were evaluated, first, over well-known academic datasets and international shared task competitions for subjective tasks such: as sentiment analysis, opinion mining, sarcasm detection, and mental-health inference; and second, over a case study of real-world social media analysis applied to the social sciences. The overall experimental results demonstrated that these models significantly and consistently outperform other off-the-shelf models — including simpler linear models, which are insufficient to cope with the noise and ambiguity of social media, and more sophisticated neural models, which require larger training datasets.

7.2 Demographically Controlled Analyses with Specialized Models

SMA systems are usually deployed as data processing pipelines to realize the following steps: (i) data collection; (ii) inference over the data with NLP methods; and (iii) data analysis. The methods proposed in thesis can be leveraged to improve over traditional SMA methodologies, as follows:

1. **Data Collection**: Instead of collecting arbitrary data samples as it is typically done, analysts are now able to conduct studies on demographically representative digital cohorts of social media users.

2. **Natural Language Processing**: Instead of relying on sub-optimal methods or manually designing task-specific models, analysts can leverage the agile modeling framework to quickly build specialized models for word-level, document-level and user-level inferences, even in low-resource settings.

3. **Data Analysis**: Instead of looking at a single measurement aggregating the raw signals inferred by the models, often taken out of context, analysts can break down the results by demographics, thus opening new ways of analyzing and interpreting the outcomes of the studies.

These methodological improvements were validated in a second case-study of real-world social media analysis, now focused on public health applications (Chapter 6). The study consisted of the development and evaluation of DEMOS, a Digital Epidemiology and Mental-Health Observation System, designed to track discussions around specific public health issues and monitor the prevalence of mental illnesses over Twitter data. The system provides a platform for the rapid deployment of custom inference models allowing social media analysts to conduct externally valid studies, vis-a-vis national trends or with focus on specific demographic groups. The evaluation was conducted over two pilot studies of demographically controlled public health surveillance over Twitter — the first, measuring how different demographic groups engage with different health related issues; and the second, investigates how mental illnesses affect different demographic groups. The studies have shown that this approach can capture and highlight differences in how specific populations engage with health topics, and how mental health issues affect these groups. For example, we found that Women and racial minorities are more susceptible to mental diseases such as depression and PTSD.

These innovations contribute to improve the practice of social media analysis for applications concerned with human activities that lack enough data to afford data-driven methods, such as the social sciences and public health. Experts in these fields, will be able to use social media to ask more pertinent and nuanced questions about the world and obtain answers much faster than before. In turn, this has the potential to drastically improve the responsiveness of public policies and programs, and our general understanding of ourselves and our societies.

7.3 Main Findings

The main findings of this research can be summarized as follows:

1. Supervised Neural Networks can be trained with scarce labeled data

In recent years supervised neural networks, particularly with deep architectures, have dominated the landscape of machine learning research and applications. One of the main drawbacks of these methods is the need for large training datasets to be properly optimized. This research demonstrated that it is possible to learn sophisticated NLP models with small amounts of labeled data, by pre-training generic neural embeddings with large amounts of unlabeled data and constraining the capacity of the architectures. The generic embeddings can be used to extract rich representations for specific tasks, thus allowing practitioners to reap the benefits of neural representation learning without paying the usual costs of these methods. This provides a middle ground between feature engineering and deep learning approaches to build specialized models. These results dispel the myth that neural representation learning is only feasible with very large labeled datasets and demonstrate that these methods are particularly beneficial for subjective tasks.

2. Learning to predict contexts captures latent semantics

Neural language models are able to induce rich word representations by learning statistical correlations between words and the contexts where they occur, which allows them to capture latent semantics (in the distributional sense, as hypothesized by Harris

(1954)). From a machine learning standpoint, we have that word vectors that are used to predict the same words (i.e. that occur in the same contexts) will converge to similar solutions. As a result, similar words end up having similar representations which also helps to improve model generalization. The methods proposed in this work rely heavily on neural language models since they provide a simple and generic mechanism to incorporate unlabeled data into the models and reduce feature engineering efforts.

This thesis then extends this idea to learn representations for users. Assuming that the topics that people choose to discuss and engage with reflect their interests and state of mind, the distributional principle can be also applied to characterize individuals. By learning to predict the words that are employed by a user, the model captures the latent characteristics of users that utter those words. Similarly to the word vectors, if the vectors of two users are optimized to predict the same words, they will gravitate towards similar solutions. Having similar users associated to similar representations offers not only better model generalization, but also provides a mechanism for analysis, e.g. epidemiologists can identify and characterize at-risk groups and gain insights into commonalities and differences in populations.

3. Demographic information significantly improves the value of social media analyses

Social media datasets are very noisy and ambiguous and thus analyses based on naive methods can often paint a misleading picture of the subject of study. This is ever more pertinent given the trend of increased polarization on controversial issues that has been observed in many contemporary societies. For example, a single measure of approval or disapproval on political issues is not very informative as it can mask strong opposing views held by different segments of the public. Moreover, the universe of social media users is not representative of the broader population, thus trends gleaned by the analysis of this data might not be generalizable.

This thesis has shown that including demographic information into social media analyses can help make some of these biases explicit, and allow analysts to break down results per demographic strata. This can help to ameliorate the negative effects of both selection bias and confirmation bias, thereby improving the validity of analyses concerning national trends. Demographic information will allow analysts to tease out possible explanations for the observed outcomes, thus fostering deeper insights into the subjects of study. Moreover, conducting social media analyses over demographically representative cohorts will allow analysts to ask more fine-grained questions, regarding not only the topics that are discussed on social media, but also how these vary across the population. Ultimately, such analyses can provide stronger evidence to inform decision processes or to support public policies.

4. Social media is a valuable data source but its analysis requires expertise

This research has shown that social media data can indeed be a valuable data source to get insights into current trends, events and states-of-affairs. However, effectively mining this data might not be so easy and straightforward as initially suggested by some researchers. Social media analyses come with many caveats, and are based on predictive models that also come with caveats. SMA systems are composed of several models and components, each of which can introduce biases and errors that might not be easy to identify without deep knowledge of the subject matter and of the various parts of the system. Suppose, we use social media to investigate support for a new policy and the results show something unexpected, e.g. much higher support than it was anticipated. Can those results be trusted? There are many ways in which those results could be biased, from poor models to data biases and deliberate manipulation. On the other hand, analyses that only confirm what we already know have limited value. Nowadays, it is possible to find data on the internet corroborating any view (even extreme) on any issue. The question then becomes how to sift through the noise. Therefore, now more than ever, it is incumbent on experts to make sense of the data, digest the findings of the analyses, put them in context, and determine their implications.

As NLP technologies evolve we can expect advances on what kinds of knowledge can be derived from social media. However, one should always be careful about drawing conclusions from these methods and acknowledge their biases and limitations. This is particularly important whenever the studies involve human subjects since these methods can reflect the biases of the analysts (e.g. the choice of the training data). One of the main goals of this thesis has been to reduce the manual efforts and domain knowledge that are required to build SMA pipelines. However, the goal is not to replace nor reduce the contribution of the experts, but to empower them. Alleviating the efforts on repetitive and time-consuming tasks will allow experts to spend more time designing robust studies and carefully analyzing their results. Lowering the effective costs of model development will allow analysts to conduct studies based on multiple such models thus enriching the value of their analyses.

7.4 Limitations

7.4.1 Low Resource Learning

It is well known that in many cases simpler machine learning models trained with more training data outperform more sophisticated models trained with less data (Banko and Brill, 2001). Specially if the assumptions under which the models operate are reflected in the task and data. A key assumption of supervised machine learners is that the training and test data are samples from the same distribution and that the training samples are independent and identically distributed. A small training set might not be enough to fully characterize the data distribution. In other words, the model does not see enough examples to generalize. The problem is that the data we use to *test* the models might be similar to the training set (since they were all collected at the same time and with the same criteria) but not representative of all the data that will be encountered at inference time. Consequently, the metrics reported in prediction tasks frequently overestimate the performance the models. This issue is greatly exacerbated when dealing with really small datasets.

This thesis aimed to improve the quality of the models that can be learned in the absence of large amounts of labeled data. Low-resource learning methods can allow for quick analyses into unexpected events (e.g. breaking news, emergency response) and facilitate exploratory studies. For example, to rapidly create bespoke models and pipelines to test hypotheses, and identify, evaluate and refine research questions. However, for analyses involving ambiguous and complex aspects, especially if they involve human subjects, efforts should be made to ensure that the models are powerful enough to capture the relevant signals and are trained with enough data to properly generalize. This is particularly relevant if the studies aim to produce actionable information or have broader implications.

7.4.2 Static User Embeddings

A limitation of the user embeddings induced by the *User2Vec* model, is that each user was represented as a static embedding that encapsulates all the information about a user. This can be problematic for applications that require modeling phenomena that evolves and fluctuates over time, for example the expression of some mental diseases such as chronic stress, anxiety and depression. This can be addressed by including the temporal structure into the embedding learning process. One way to do this would be to split users data into temporal bins and induce a vector for each bin, providing a ‘snapshot’ of the user at different points in time. Another would be to adopt recurrent neural network models. These models are able to process inputs sequentially and produce intermediate representations at each time step, conditioned on the previous steps. As such, we could process a user’s posts collection sequentially and use the internal representations at time step t to characterize the user at time t. Applications for these *user embeddings over time* include better modeling the onset of specific illnesses and their symptoms, building non-intrusive systems for early detection, or monitoring the effects of health interventions on a population.

7.4.3 Digital Epidemiology

Despite the promising results of the digital epidemiology system presented in this thesis, this work still leaves many open research questions. On the technical side, the research on methods to infer demographic attributes and mental-health signals from social media is still in its infancy and advances in these areas will directly improve the quality and reliability of this system. Moreover, the current implementation still assumes that all the data comes from legitimate users, and thus it is vulnerable to intentional manipulation attempts, e.g. by *sock puppet* social bots (see Ferrara (2017)). This problem is exacerbated by the fact that the current NLP models are still very easy to fool, particularly simple lexicon-based methods. Nevertheless, the positive results obtained by our user-level models suggest that this approach could also be used to detect social bots, or at least provide a reasonable baseline for improved methods. On the methodological side, it is still not clear the impact of using self-reported data to build models for the general public, but more importantly we still do not know how reliable are the signals extracted by SMA systems with regards to public-health. This leaves many open questions, such as: how trustworthy are the outcomes of the analyses, if they are to inform actionable decisions? what kinds of decisions can be informed by these methods?

7.4.4 Ethics, Data Privacy and Consent

The majority of social media analysis systems try to extract signals from pools of publicly shared posts. If the data is randomly sampled it is unlikely that the sample will contain more than a few messages from the same user. Moreover, in this case there is no need to record any personal information about the user. However, as we start moving towards analyses at the user-level, we are collecting and storing complete records of social media users communications. Even though this information is publicly available, people might not be consciously aware of the implications of sharing all their data and certainly have not given explicit consent for their data to be analyzed in aggregate. This is even more pertinent for analyses involving sensitive information about people, such as mental-health statuses. As it has been demonstrated by the recent incidents involving companies sharing personal data of their users, there is a serious danger of abuse and exploitation for systems that collect and store large amounts of personal data.

Even though this is in large part an ethical question, there are technical solutions that can be used to partially address this issue. One is to use anonymization techniques to obfuscate any details that allow third parties (even analysts) to identify the individuals that are involved in the study. Another is to store only the user embeddings and discard the actual profile content. The representations could then be update when new data became available. This way it would still be possible to run user-level analytics over those users, but not document-level ones. In regards to consent, there are initiatives to support voluntary data donation for research purposes, e.g. the *Our Data Helps* program[1].

7.5 Future Work

This work represents the initial steps towards more accurate inferences over social media data repositories and thus leaves open many doors for future work. In regards to agile modeling, this work relies heavily on neural language models and the representations they produce. The research in this field is still very active and this work will directly benefit from any developments in this area. For example, Felbo et al. (2017) showed that *emotion-aware* embeddings, that yield state-of-the-art results for subjective tasks, can be induced by training deep neural networks to predict the presence of emoji. Others are moving towards hierarchical models, that similarly to deep models for computer vision, capture representations at different levels of abstraction (Peters et al., 2018). Another related venue worth pursuing is the use of cross-lingual and multi-lingual embeddings, i.e. learned vectors that represent words in different languages with similar representations which allow practitioners to train models with labeled data in one language and perform inference on a different language (Klementiev et al., 2012; Hermann and Blunsom, 2014). This would provide a cheap mechanism to extend SMA to languages that lack computational linguistics resources. Finally, regarding user embeddings, future work could investigate how to compose user embeddings to form representations of groups of

[1] `https://ourdatahelps.org/`

people. This could have applications for epidemiology by allowing analysts to characterize groups and investigate the factors that contribute to the spreading of specific diseases.

Regarding the work on digital epidemiology, future developments include expanding the analyses to other public health issues and mental diseases. Further work must also be done to understand how reliable are the indicators induced by these systems, and how can they be leveraged to inform decisions. The DEMOS case-study was meant as a proof-of-concept for a large-scale deployment of a digital epidemiology system. However, these systems can also be deployed for analyses at a local level, for example, measuring the well-being of students in a school; or implementing 'neighborhood watch' programs where community leaders could directly measure the well-being of their members and inform first responders in crisis situations. This also raises concerns regarding privacy and consent, as discussed above. Finally, the ability to use social media to track mental-health and well-being could also be used for self-analysis and personal development. Indicators extracted by the models could be complemented with Ecological Momentary Assessment surveys (Moskowitz and Young, 2006) to periodically assess user well-being or other factors that contribute to well-being (e.g. exercising, getting enough sleep). In this regard, I already developed a fully functional rule-based conversational agent for Slack[2] environments to elicit this information from users, in a non-intrusive manner[3]. This information could then be aligned with the model indicators to uncover correlations between the use of language and self-reported measurements of physical and mental wellbeing. Taking it one step further, we could then use the self-reported measurements to train personal models to infer these measurements from language use.

[2] Slack is a team chat application — `https://slack.com/`

[3] `https://github.com/samiroid/sleek`

Bibliography

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011).
Sentiment analysis of Twitter data.
In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amir, S., Almeida, M. B., Martins, B., Filgueiras, J. a., and Silva, M. J. (2014).
TUGAS: Exploiting unlabelled data for Twitter sentiment analysis.
In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 673–677, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Amir, S., Astudillo, R., Ling, W., Carvalho, P. C., and Silva, M. J. (2016a).
Expanding subjective lexicons for social media mining with embedding subspaces.
arXiv preprint arXiv:1701.00145.

Amir, S., Astudillo, R., Ling, W., Silva, M. J., and Trancoso, I. (2016b).
Inesc-id at semeval-2016 task 4-a: Reducing the problem of out-of-embedding words.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 238–242. Association for Computational Linguistics.

Amir, S., Coppersmith, G., Carvalho, P., Silva, M. J., and Wallace, B. C. (2017).
Quantifying mental health from social media with neural user embeddings.
In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321, Boston, Massachusetts. PMLR.

Amir, S., Ling, W., Astudillo, R., Martins, B., Silva, M. J., and Trancoso, I. (2015).
INESC-ID: A regression model for large scale Twitter sentiment lexicon induction.
In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618, Denver, Colorado. Association for Computational Linguistics.

Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (2016c).
Modelling context with user embeddings for sarcasm detection in social media.
In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, CONLL '16, pages 167–177. Association for Computational Linguistics.

Astudillo, R., Amir, S., Ling, W., Martins, B., Silva, M. J., and Trancoso, I. (2015a).
INESC-ID: Sentiment analysis without hand-coded features or linguistic resources using embedding subspaces.
In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 652–656, Denver, Colorado. Association for Computational Linguistics.

Astudillo, R., Amir, S., Ling, W., Silva, M., and Trancoso, I. (2015b).
Learning word representations from scarce and noisy data with embedding subspaces.
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)*, ACL '15, pages 1074–1084, Beijing, China. Association for Computational Linguistics.

Asur, S., Huberman, B., et al. (2010).
Predicting the future with social media.
In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.

Ayers, J. W., Leas, E. C., Allem, J.-P., Benton, A., Dredze, M., Althouse, B. M., Cruz, T. B., and Unger, J. B. (2017).
Why do people use electronic nicotine delivery systems (electronic cigarettes)? a content analysis of Twitter, 2012-2015.
PLOS ONE, 12(3):e0170702.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010).
SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.
In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*.

Bamman, D. and Smith, N. A. (2015).
Contextualized sarcasm detection on Twitter.
In *Proceedings of the 9th International Conference on Web and Social Media*, pages 574–77. AAAI Menlo Park, CA.

Banko, M. and Brill, E. (2001).
Scaling to very very large corpora for natural language disambiguation.
In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL, pages 26–33. Association for Computational Linguistics.

Barbieri, F. and Saggion, H. (2014).
Modelling irony in Twitter.
In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 56–64.

Barbosa, L. and Feng, J. (2010).
Robust sentiment detection on twitter from biased and noisy data.
In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bengio, Y., Courville, A., and Vincent, P. (2013).
Representation learning: A review and new perspectives.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(8):1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
The Journal of Machine Learning Research, 3:1137–1155.

Benton, A., Coppersmith, G., and Dredze, M. (2017).
Ethical research protocols for social media health research.
In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102. Association for Computational Linguistics.

Bermingham, A. and Smeaton, A. (2011).
On using Twitter to monitor political sentiment and predict election results.
In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10.

Bessi, A. and Ferrara, E. (2016).
Social bots distort the 2016 US presidential election online discussion.
First Monday, 21(11).

Bestgen, Y. and Vincze, N. (2012).

Checking and bootstrapping lexical norms by means of word similarity indexes.
Behavior Research Methods, 44(4).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
the Journal of Machine Learning research, 3:993–1022.

Bollen, J., Mao, H., and Pepe, A. (2011a).
Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11.

Bollen, J., Mao, H., and Zeng, X. (2011b).
Twitter mood predicts the stock market.
Journal of computational science, 2(1):1–8.

Bošnjak, M., Oliveira, E., Martins, J., Mendes Rodrigues, E., and Sarmento, L. (2012).
TwitterEcho: A distributed focused crawler to support open research with Twitter data.
In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1233–1240, New York, NY, USA. ACM.

Bradley, M. M. and Lang, P. J. (1999).
Affective norms for english words (ANEW): Instruction manual and affective ratings.
Technical report, The Center for Research in Psychophysiology, University of Florida.

Broniatowski, D. A., Paul, M. J., and Dredze, M. (2013).
National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic.
PLOS ONE, 8(12):e83672.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992).
Class-based n-gram models of natural language.
Computational Linguistics, 18(4):467–479.

Budhwani, H., Hearld, K. R., and Chavez-Yenter, D. (2015).
Depression in racial and ethnic minorities: the impact of nativity and discrimination.
Journal of racial and ethnic health disparities, 2(1):34–42.

Carvalho, P., Sarmento, L., Silva, M. J., and De Oliveira, E. (2009).
Clues for detecting irony in user-generated contents: oh...!! it's so easy;-).
In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Carvalho, P., Sarmento, L., Teixeira, J., and Silva, M. J. (2011).
Liars and saviors in a sentiment annotated corpus of comments to political debates.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, ACL '11, pages 564–568. Association for Computational Linguistics.

Castillo, C., Mendoza, M., and Poblete, B. (2011).
Information credibility on twitter.
In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA. ACM.

Cesare, N., Grant, C., and Nsoesie, E. O. (2017).
Detection of user demographics on social media: A review of methods and recommendations for best practices.
arXiv preprint arXiv:1702.01807.

Chung, J. and Mustafaraj, E. (2011).
Can collective sentiment expressed on twitter predict political elections?
In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI'11, pages 1770–1771. AAAI Press.

Colen, C. G., Ramey, D. M., Cooksey, E. C., and Williams, D. R. (2017).
Racial disparities in health among nonpoor african americans and hispanics: the role of acute and chronic discrimination.
Social Science & Medicine.

Colleoni, E., Rozza, A., and Arvidsson, A. (2014).
Echo chamber or public sphere? predicting political orientation and measuring political homophily in Twitter using big data.
Journal of Communication, 64(2):317–332.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
The Journal of Machine Learning Research, 12:2493–2537.

Colston, H. and Gibbs, R. (2007).
A brief history of irony.
Irony in language and thought: A cognitive science reader, pages 3–21.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., and Flammini, A. (2011).
Political polarization on twitter.
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11.

Coppersmith, G., Dredze, M., and Harman, C. (2014a).
Quantifying mental health signals in twitter.
In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. Association for Computational Linguistics.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015a).
From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses.
In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10. Association for Computational Linguistics.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015b).
Clpsych 2015 shared task: Depression and ptsd on twitter.
In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39. Association for Computational Linguistics.

Coppersmith, G., Harman, C., and Dredze, M. (2014b).
Measuring post traumatic stress disorder in Twitter.
In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM14.

Coppersmith, G., Hilland, C., Frieder, O., and Leary, R. (2017).
Scalable mental health analysis in the clinical whitespace via natural language processing.
In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 393–396. IEEE.

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

Cover, T. M. (1965).
Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.
IEEE transactions on electronic computers, pages 326–334.

Culotta, A. (2010).
Towards detecting influenza epidemics by analyzing Twitter messages.
In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Davidov, D., Tsur, O., and Rappoport, A. (2010a).
Enhanced sentiment learning using twitter hashtags and smileys.
In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Davidov, D., Tsur, O., and Rappoport, A. (2010b).
Semi-supervised recognition of sarcasm in Twitter and Amazon.
In *Proceedings of the 14th Conference on Computational Natural Language Learning*, CONLL '10, pages 107–116. Association for Computational Linguistics.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. (2016).
Discovering shifts to suicidal ideation from mental health content in social media.
In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA. ACM.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990).
Indexing by latent semantic analysis.
JAsIs, 41(6):391–407.

Dews, S., Kaplan, J., and Winner, E. (1995).
Why not say it directly? the social functions of irony.
Discourse processes, 19(3):347–367.

Diakopoulos, N. A. and Shamma, D. A. (2010).
Characterizing debate performance via aggregated Twitter sentiment.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA. ACM.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011).
Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter.
PLOS ONE, 6(12):e26752.

Dredze, M. (2012).
How social media will change public health.
IEEE Intelligent Systems, 27(4):81–84.

Dredze, M., Broniatowski, D. A., and Hilyard, K. M. (2016a).
Zika vaccine misconceptions: A social media analysis.
Vaccine, 34(30):3441.

Dredze, M., Broniatowski, D. A., Smith, M. C., and Hilyard, K. M. (2016b).
Understanding vaccine refusal: why we need social media now.
American journal of preventive medicine, 50(4):550–552.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013).
Carmen: A Twitter geolocation system with applications to public health.
In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005).
Comparing data from online and face-to-face surveys.
International Journal of Market Research, 47(6):615.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988).
Using latent semantic analysis to improve access to textual information.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, pages 281–285, New York, NY, USA. ACM.

Dyer, C. (2014).
Notes on noise contrastive estimation and negative sampling.
arXiv preprint arXiv:1410.8251.

Eisenstein, J. (2017).
Unsupervised learning for lexicon-based classification.
In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3188–3194.

Eldan, R. and Shamir, O. (2016).
The power of depth for feedforward neural networks.
In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, Columbia University, New York, New York, USA. PMLR.

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

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017).
Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 1615–1625. Association for Computational Linguistics.

Ferrara, E. (2017).
Disinformation and social bot operations in the run up to the 2017 french presidential election.
First Monday, 22(8).

Filgueiras, J. and Amir, S. (2013).
POPSTAR at RepLab 2013: Polarity for reputation classification.
In *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*.

Firth, J. R. (1961).
Papers in Linguistics 1934-1951.
Oxford University Press.

Flory, J. D. and Yehuda, R. (2015).
Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and treatment considerations.
Dialogues in clinical neuroscience, 17(2):141.

Gayo-Avello, D. (2013).
A meta-analysis of state-of-the-art electoral prediction from Twitter data.
Social Science Computer Review, 31(6):649–679.

Gayo Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011).
Limits of electoral predictions using Twitter.
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11. Association for the Advancement of Artificial Intelligence.

Ghahramani, Z. (2004).
Unsupervised learning.
In *Advanced lectures on machine learning*, pages 72–112. Springer.

Go, A., Bhayani, R., and Huang, L. (2009).
Twitter sentiment classification using distant supervision.
CS224N Project Report, Stanford, pages 1–12.

Goldberg, Y. (2016).
A primer on neural network models for natural language processing.
Journal of Artificial Intelligence Research, 57:345–420.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011).
Identifying sarcasm in twitter: A closer look.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL '11, pages 581–586. Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep Learning.
MIT Press.
`http://www.deeplearningbook.org`.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016).
Inducing domain-specific sentiment lexicons from unlabeled corpora.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 595–605. Association for Computational Linguistics.

Han, B., Cook, P., and Baldwin, T. (2013).
Lexical normalization for social media text.
ACM Transactions on Intelligent Systems and Technology (TIST), 4(1):5.

Hao, B., Li, L., Li, A., and Zhu, T. (2013).
Predicting mental health status on social media.
In *Cross-Cultural Design. Cultural Differences in Everyday Life.CCD 2013*, Lecture Notes in Computer Science, pages 101–110. Springer, Berlin, Heidelberg.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Helbing, D. and Balietti, S. (2011).
From social data mining to forecasting socio-economic crises.
The European Physical Journal Special Topics, 195(1):3.

Hermann, K. M. and Blunsom, P. (2014).
Multilingual models for compositional distributed semantics.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68. Association for Computational Linguistics.

Hinton, G. E. (1986).
Learning distributed representations of concepts.
In *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, volume 1, page 12. Amherst, MA.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006).
A fast learning algorithm for deep belief nets.
Neural computation, 18(7):1527–1554.

Hinton, G. E. and Salakhutdinov, R. R. (2006).
Reducing the dimensionality of data with neural networks.
Science, 313(5786):504–507.

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

Hu, M. and Liu, B. (2004).
Mining and summarizing customer reviews.
In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Huang, X., Smith, M., Paul, M., Ryzhkov, D., Quinn, S., Broniatowski, D., and Dredze, M. (2017).
Examining patterns of influenza vaccination in social media.
In *AAAI Workshops*.

J. Paul, M. and Dredze, M. (2017).
Social Monitoring for Public Health.
Morgan & Claypool Publishers.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014).
Synthetic data and artificial neural networks for natural scene text recognition.
arXiv preprint arXiv:1406.2227.

Jungherr, A., Jürgens, P., and Schoen, H. (2012).
Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with Twitter: What 140 characters reveal about political sentiment".
Social science computer review, 30(2):229–234.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014).
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. Association for Computational Linguistics.

Khattri, A., Joshi, A., Bhattacharyya, P., and Carman, M. (2015).
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. Association for Computational Linguistics.

Kim, S.-M. and Hovy, E. (2006).
Identifying and analyzing judgment opinions.

In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 200–207, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, Y. (2014).
Convolutional neural networks for sentence classification.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016).
Character-aware neural language models.
In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2741–2749.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014).
Sentiment analysis of short informal texts.
Journal of Artificial Intelligence Research, pages 723–762.

Klementiev, A., Titov, I., and Bhattarai, B. (2012).
Inducing crosslingual distributed representations of words.
In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING 2012, pages 1459–1474. The COLING 2012 Organizing Committee.

Kober, T. and Weir, D. (2015).
Optimising agile social media analysis.
In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–40, Lisboa, Portugal. Association for Computational Linguistics.

Kouloumpis, E., Wilson, T., and Moore, J. (2011).
Twitter sentiment analysis: The good the bad and the omg!
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11, pages 538–541.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *Advances in neural information processing systems*, pages 1097–1105.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009).
Life in the network: the coming age of computational social science.
Science (New York, NY), 323(5915):721.

Le, Q. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998).
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2324.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009).
Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.
In *Proceedings of the 26th International Conference on Machine Learning*, ICML-09, pages 609–616. ACM.

Li, Z., Xiong, Z., Zhang, Y., Liu, C., and Li, K. (2011).
Fast text categorization using concise semantic analysis.
Pattern Recognition Letters, 32(3):441–448.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015a).
Two/too simple adaptations of word2vec for syntax problems.
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '15, pages 1299–1304. Association for Computational Linguistics.

Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., and Luis, T. (2015b).
Finding function in form: Compositional character models for open vocabulary word representation.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 1520–1530. Association for Computational Linguistics.

Liu, B. (2012).
Sentiment analysis and opinion mining.

Synthesis Lectures on Human Language Technologies, 5(1):1–167.

Liu, B. (2015).
Sentiment analysis: Mining opinions, sentiments, and emotions.
Cambridge University Press.

Lukin, S. and Walker, M. (2013).
Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue.
In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40. Association for Computational Linguistics.

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., and Rodrigue, J. (2012).
A demographic analysis of online sentiment during hurricane irene.
In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 27–36. Association for Computational Linguistics.

Manning, C. D., Manning, C. D., and Schütze, H. (1999).
Foundations of statistical natural language processing.
MIT press.

Marchetti-Bowick, M. and Chambers, N. (2012).
Learning for microblogs with distant supervision: Political forecasting with Twitter.
In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612. Association for Computational Linguistics.

McCaughey, D., Baumgardner, C., Gaudes, A., LaRochelle, D., Wu, K. J., and Raichura, T. (2014).
Best practices in social media: Utilizing a value matrix to assess social media's impact on health care.
Social Science Computer Review, 32(5):575–589.

Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. (2011).
How (not) to predict elections.
In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 165–171.

Mikal, J., Hurst, S., and Conway, M. (2016).

Ethical issues in using Twitter for population-level depression monitoring: a qualitative study.
BMC medical ethics, 17(1):1.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a).
Distributed representations of words and phrases and their compositionality.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b).
Linguistic regularities in continuous space word representations.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '13, pages 746–751. Association for Computational Linguistics.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c).
Linguistic regularities in continuous space word representations.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '13, pages 746–751.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007).
Measurement and analysis of online social networks.
In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42, New York, NY, USA. ACM.

Mitchell, L., Harris, K. D., Frank, M. R., Dodds, P. S., and Danforth, C. M. (2013).
The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place.
PLOS ONE, 8(5).

Mitchell, M., Coppersmith, G., and Hollingshead, K., editors (2015).
Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.
North American Association for Computational Linguistics, Denver, Colorado, USA.

Miura, Y., Sakaki, S., Hattori, K., and Ohkuma, T. (2014).

Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data.
In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland. Association for Computational Linguistics.

Mnih, A. and Hinton, G. E. (2009).
A scalable hierarchical distributed language model.
In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1081–1088. Curran Associates, Inc.

Mohammad, S. M. and Turney, P. D. (2013).
Crowdsourcing a word–emotion association lexicon.
Computational Intelligence, 29(3).

Moreira, S., Batista, D., Carvalho, P., Couto, F. M., and Silva, M. J. (2011).
POWER-Politics Ontology for Web Entity Retrieval.
In *Advanced Information Systems Engineering Workshops*, pages 489–500. Springer.

Morin, F. and Bengio, Y. (2005).
Hierarchical probabilistic neural network language model.
In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics.

Moskowitz, D. S. and Young, S. N. (2006).
Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology.
Journal of Psychiatry and Neuroscience, 31(1):13.

Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. T. (2011).
Vocal minority versus silent majority: Discovering the opionions of the long tail.
In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 103–110.

Nagy, A. and Stamberger, J. (2012).
Crowd sentiment detection during disasters and crises.
In *Proceedings of the 9th International ISCRAM Conference*, pages 1–9.

Nair, V. and Hinton, G. E. (2010).
Rectified linear units improve restricted boltzmann machines.

In *Proceedings of the 27th International Conference on Machine Learning*, ICML-10, pages 807–814.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013).
Semeval-2013 task 2: Sentiment analysis in twitter.
In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. Association for Computational Linguistics.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010).
From tweets to polls: Linking text sentiment to public opinion time series.
In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM10.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013).
Improved part-of-speech tagging for online conversational text with word clusters.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '13, pages 380–390. Association for Computational Linguistics.

Pak, A. and Paroubek, P. (2010).
Twitter as a corpus for sentiment analysis and opinion mining.
In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Paltoglou, G. and Thelwall, M. (2010).
A study of information retrieval weighting schemes for sentiment analysis.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395. Association for Computational Linguistics.

Pan, S. J. and Yang, Q. (2010).
A survey on transfer learning.
Knowledge and Data Engineering, IEEE Transactions on, 22(10):1345–1359.

Pang, B. and Lee, L. (2008).
Opinion mining and sentiment analysis.
Foundations and trends in information retrieval, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002).

Thumbs up? sentiment classification using machine learning techniques.
In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02. Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013).
On the difficulty of training recurrent neural networks.
In *Proceedings of the 30th International Conference on Machine Learning*, ICML-13, pages 1310–1318.

Paul, M. J. and Dredze, M. (2011).
You are what you tweet: Analyzing Twitter for public health.
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11. Association for the Advancement of Artificial Intelligence.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001).
Linguistic inquiry and word count: LIWC 2001.
Mahway: Lawrence Erlbaum Associates, 71:2001.

Pennington, J., Socher, R., and Manning, C. (2014).
Glove: Global vectors for word representation.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL '18, pages 2227–2237. Association for Computational Linguistics.

Plutchik, R. (1980).
A general psychoevolutionary theory of emotion, pages 3–33.
Academic press.

Purver, M. and Battersby, S. (2012).
Experimenting with distant supervision for emotion classification.
In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491. Association for Computational Linguistics.

Rajadesingan, A., Zafarani, R., and Liu, H. (2015).
Sarcasm detection on Twitter: A behavioral modeling approach.
In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 97–106, New York, NY, USA. ACM.

Rao, D. and Ravichandran, D. (2009).
Semi-supervised polarity lexicon induction.
In *Proceedings of the 12th Conference of the European Chapter of the ACL*, EACL '09, pages 675–682. Association for Computational Linguistics.

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011).
Detecting and tracking political abuse in social media.
In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11. Association for the Advancement of Artificial Intelligence.

Řehůřek, R. and Sojka, P. (2010).
Software Framework for Topic Modelling with Large Corpora.
In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Reyes, A., Rosso, P., and Veale, T. (2013).
A multidimensional approach for detecting irony in Twitter.
Language Resources and Evaluation (LREC), 47(1):239–268.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013).
Sarcasm as contrast between a positive sentiment and negative situation.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 704–714. Association for Computational Linguistics.

Ritterman, J., Osborne, M., and Klein, E. (2009).
Using prediction markets and Twitter to predict a swine flu pandemic.
In *Proceedings of the First International Workshop on Mining Social Media*, volume 9, pages 9–17.

Rosenberger, W. F., Sverdlov, O., and Hu, F. (2012).
Adaptive randomization for clinical trials.
Journal of biopharmaceutical Statistics, 22(4):719–736.

Rosenblatt, F. (1961).
Principles of neurodynamics. perceptrons and the theory of brain mechanisms.
Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015).
Semeval-2015 task 10: Sentiment analysis in Twitter.
In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463. Association for Computational Linguistics.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988).
Learning representations by back-propagating errors.
Cognitive modeling, 5:3.

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., et al. (2012).
Digital epidemiology.
PLoS computational biology, 8(7):e1002616.

Saleiro, P., Amir, S., Silva, M., and Soares, C. (2015).
POPmine: Tracking political opinion on the web.
In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 1521–1526.

Saleiro, P., Rei, L., Pasquali, A., Soares, C., Teixeira, J., Pinto, F., Nozari, M., Félix, C., and Strecht, P. (2013).
POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter.
In *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. (2014).
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. Association for Computational Linguistics.

Settles, B. (2011).

Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances.
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1467–1478. Association for Computational Linguistics.

Shamma, D. A., Kennedy, L., and Churchill, E. F. (2009).
Tweet the debates: Understanding community annotation of uncollected sources.
In *Proceedings of the First SIGMM Workshop on Social Media*, WSM '09, pages 3–10, New York, NY, USA. ACM.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012).
Building a sentiment lexicon for social judgement mining.
In *International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017).
Mastering the game of go without human knowledge.
Nature, 550(7676):354.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011).
Twitter polarity classification with label propagation over lexical links and the follower graph.
In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014).
Dropout: a simple way to prevent neural networks from overfitting.
The Journal of Machine Learning Research, 15(1):1929–1958.

Tang, D., Wei, F., Qin, B., Zhou, M., and Liu, T. (2014).
Building large-scale Twitter-specific sentiment lexicon : A representation learning approach.
In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, COLING 2014, pages 172–182. Dublin City University and Association for Computational Linguistics.

Tjong Kim Sang, E. and Bos, J. (2012).
Predicting the 2011 dutch senate election results with twitter.

In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60. Association for Computational Linguistics.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010).
Predicting elections with Twitter: What 140 characters reveal about political sentiment.
In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM10.

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010).
Word representations: A simple and general method for semi-supervised learning.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Turney, P. D. and Littman, M. L. (2003).
Measuring praise and criticism: Inference of semantic orientation from association.
ACM Transactions on Information Systems, 21(4).

Van der Maaten, L. and Hinton, G. (2008).
Visualizing data using t-SNE.
Journal of Machine Learning Research, 9(2579-2605):85.

Vapnik, V. N. (2000).
The nature of statistical learning theory. statistics for engineering and information science.
Springer-Verlag, New York.

Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. (2010).
The viability of web-derived polarity lexicons.
In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '10, pages 777–785. Association for Computational Linguistics.

Wallace, B. C., Choe, D. K., and Charniak, E. (2015).
Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment.
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), ACL '15, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Wallace, B. C., Choe, D. K., Kertz, L., and Charniak, E. (2014).
Humans require context to infer ironic intent (so computers probably do, too).
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '14, pages 512–516. Association for Computational Linguistics.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013).
Norms of valence, arousal, and dominance for 13,915 english lemmas.
Behavior research methods, 45(4):1191–1207.

Wibberley, S., Weir, D., and Reffin, J. (2013).
Language technology for agile social media science.
In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 36–42. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005).
Recognizing contextual polarity in phrase-level sentiment analysis.
In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Zeiler, M. D. (2012).
Adadelta: an adaptive learning rate method.
arXiv preprint arXiv:1212.5701.

Zhang, X., Fuehres, H., and Gloor, P. A. (2011).
Predicting stock market indicators through Twitter "i hope it is not as bad as i fear".
Procedia-Social and Behavioral Sciences, 26:55–62.

Zhang, Y. and Wallace, B. (2015).
A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification.
arXiv preprint arXiv:1510.03820.

Zhu, X. (2005).
Semi-supervised learning literature survey.
Technical Report Computer Sciences TR 1530, University of Wisconsin–Madison.