UNIVERSITÀ DEGLI STUDI DI TORINO

DOCTORAL THESIS

UNIVERSITÀ DEGLI STUDI DI TORINO

Challenges of Hate Speech Detection on Social Media: The Role of Time, Topic and Demography

---

Author:
Komal FLORIO

Supervisor:
Prof. Viviana PATTI
Prof. Valerio BASILE

PhD Program Coordinator:
Prof. Marco GRANGETTO

A thesis submitted in fulfilment of the requirements
for the PhD in Computer Science (INF/01)
XXXIII Cycle

December 2021

To Barcelona.

Abstract

The availability of large annotated corpora from social media and the development of powerful classification approaches have contributed in an unprecedented way to tackle the challenge of monitoring users' opinions and sentiments in online social platforms across time. This thesis aims to explore in particular the phenomenon of hate speech messages on social media with a multiperspective approach. Abusive messages are often characterising online polarized debates in social media and are directed towards a multitude of different targets, e.g., immigrants, but the challenges in detecting such messages can be generalized. From the computational point of view, despite the increasing interest of the computational linguistics community on developing automatic systems abusive language detection and related tasks for different languages, the robustness of detection and monitoring systems emerges as a crucial factor to be addressed. One of the main limitations observed in the current proposals is to consider the NLP task of detecting abusive language in isolation, without taking into account the intersection with the contextual or social dimensions, that could contribute to a more holistic comprehension of the expression of abusive phenomena in language. Among the many factors playing a role in this challenge, in this work we focus on the temporal dimension of the phenomena, the debated topic and the demographic characteristics of the users, which brought us to intersect our NLP research with the field of Computational Social Science. At first we will tackle to challenge of temporal robustness of hate speech detection and monitoring systems by looking in depth at the interplay between the most commonly used classification algorithms and training data. More specifically we will focus on the trade off between the training dataset size and their temporal distance to the monitoring data in the context of different algorithms. We will then investigate the role of topic shift in social media public discourse and how this affects the performances of such algorithms. Intuitively we can say that quite often the public discourse on social media follows very closely breaking news from traditional and online media. We use as case study a dataset of tweets related to the COVID-19 induced lockdown in Italy (which was the first country in Europe to introduce such a dramatic restrictions) to measure how quickly the main topic on public discourse shifted in time, as this was the perfect example of government measures that deeply affected everyday life of citizens and hence had the potential to spark very heated debates online. At least we offer a demographic and socio-economics perspective on hate speech geographical distribution with a specific focus on the Italian case, given that Italy has been the first arrival Country for immigrants in the past decade and this makes for an interesting context for analysing the public discourse around immigration and integration of foreign people.

Acknowledgements

This thesis represents my best effort in describing the results obtained during the last tree years and that would have not been possible to achieve without the support, mentorship and guidance of many people.

First and foremost I have to thank my two supervisors, Prof. Viviana Patti and Prof. Valerio Basile, for having guided me through a new path in my professional life.

I owe a great debt to Prof. Judith Simon, for all the time dedicated to me and my research that she managed to carve out of her really busy schedule. Thank you for showing me how to be efficient at time managing.

I would like to send a big "thank you" to Dr. Mirko Lai, for helping me in my complicated relation with Python, with calm and infinite patience.

I am also feeling deeply grateful for the kind words written about this thesis by the two reviewers, Prof. Mariona Taulé and Prof. Nicole Novielli. After many months of writing in solitude, during a pandemic, and doubting the goodness of both writing and contents, your enthusiastic feedback brought a much needed ray of light.

I would also like to thank the PhD Committee (Prof. Inglesias, Prof. Novielli and Prof. Russo) that will attend the defence of this work in Turin in person. Being able to held this event in a hybrid mode signs a much hoped for return to some sort of new norm in academic life.

A PhD is also a journey with peers and among them I would like to thank Dr. Stefano Tedeschi for offering insights and support when I much needed them. Thanks also for sharing with me fun chats about sailing, baking and more. I really hope this will continue, even when our offices won't be just one thin wall away from one another.

Over the course of this PhD I was lucky enough to be surrounded by dear friends who brought into my life friendship, guidance, advice, physical and virtual hugs, wisdom, inspiration, encouragement, laughs and much needed information and support of different sorts, especially during the pandemic. In particular, and in no specific order, I would like to mention Dr. Yelena Mejova, Prof. Ciro Cattuto, Dr. Celine Sin, Ms. Wendie Shaffer & Prof. David Bree, Dr. Luca Chiarandini, Prof. Luca Aiello, Prof. Paolo Boldi, Luisa Lohmann & Hugo Rodriguez, little Emma B. and her parents.

I would not know where to begin if I was to list all the many ways in which Dr. Joseph Wakeling has been fundamental in helping me successfully achieve this milestone of my life, so I think I have to rely on a simple but deeply heartfelt "thank you".

Last but certainly not least, I would like to thank my parents, for their lifelong love and support.

Komal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the distinctive peculiarity of humans with respect to other species that inhabit our planet is the ability to develop rather complex technologies. Through the millennia humans have hence built an incredible variety of different technologies out of the need to solve collective problems or seek a better or easier life.

On the one had this evolution has allowed humanity to overcome issues and reach more ambitious goals both as a society and individual, but on the other hand it has arisen new unexpected challenges to face.

As an example, the advent of efficient and low cost efficient refrigerating technologies has saved many lives from food-related diseases, but in the medium to long term it has contributed to the worsening of the green house effect and the enlarging of the ozone hole due to the emission in the atmosphere of CFC gas. Another clear example is the widespread success of cars. Motorized vehicles revolutionized the transportation of people and goods, impacting deeply on multiple economic sectors such as industry, trades and tourism, just to mention a few. This new paradigm contextually ended the hygienic issues in highly populated cities where a sizeable number of horses were massively used for different purposes. As the adoption of this technology grew at a super fast rate it also started to show some unexpected collateral problem such as traffic, air pollution, pedestrian safety, just to name a few. For sure one the most prominent features of the year 2000s is the pervasive spread of Internet and the impact of it both on an individual and collective scale. The benefit of this revolution are multiple and diverse and it is definitely beyond the scope of this thesis to list and analyse them. In this work I will concentrate on hate speech on social media (from now on HS): a very specific issue that is arising as a multitude of free platform are now available for free for a multitude of people to express themselves, share and debate their opinion. The problem of negative messages in the public debate and their potential harm is of course not new, but the recent availability of online venues offer these sort of messages a wider audience and an easy access to echo chambers that tend to lead to a polarized debate [40]. This recent phenomenon created new challenges on multiple levels. The privately owned platforms have been place under scrutiny for their efforts to mitigate this behaviour of their users, as described for example here in Italian or here in English. Institutions and governments world wide have started working on measures to contain the problem and the potential harm when online conversations are turned into real life hate crimes.

For what regards Europe, in 2016 it was issued the The EU Code of Conduct, to prevent and counter the spread of illegal hate speech online.

Similarly, a couple of years later, in 2020, the UN issued the "United Nations Strategy and Plan of Action on Hate Speech" to address the issue on a national and global level.

Hate speech is a complex phenomenon and this reflects on the fact that over the time there were several possible definition proposed but so far there is no unified and universally accepted version.

To start with a general definition, we could refer to the one contained in the United Nations Strategy and Plan of Action on Hate Speech", that reads:

> *Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive.*

As in this thesis I will be focusing on social media, and on Twitter in particular, it is useful to review how hate speech is defined in their terms of service, that every user has to subscribe in order to be able to post:

> *Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national,sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease.*

This definition is characterized by two main elements: a target of hate speech and a call to action and these two principles constitute the foundations of what will be regarded as hate speech in this work [96]. The motivation for detecting hateful messages are manifold and a complete review of the negative impact of such behaviour is beyond the scope and purpose of this thesis. In general, it is useful to distinguish between online and offline harmful effects even tough the two are not completely unrelated. Among the major online harmful effects on online negative messages we can count the polarization and radicalization of opinions. When the online debate spills over in real life, then there is an increased risk that it translates into hate crimes that eventually impact the victims lives in a meaningful way both in terms of emotional response and perception of personal safety. Bridging the gap between online and offline is a hard task. Some works tried to investigate the impact of online racist messages and crimes, as in [9] and [28] but it is an open challenge.

The task of detection abusive message is a very challenging one and from multiple perspective. From the computational point of view, despite the increasing interest and effort of the community on developing automatic systems abusive language detection and related tasks [54, 96, 118] for different languages, the robustness of detection and monitoring systems emerges as a crucial factor to be addressed, where one of the main limitations observed is to consider the NLP task of detecting abusive language in isolation, without taking into account the intersection with the contextual or social dimensions, that could contribute to a more holistic comprehension of the abusive phenomena in language. Among the many factors playing a role in this challenge, in this work we focus on the temporal dimension of the phenomena, the debated topic and the demographic characteristics of the users, which brought us to intersect our NLP research with the field of Computational Social Science.

Specifically, in the work presented in this thesis we aim at addressing the following research questions, visually represented in 1.1:

1. **RQ 1:** How can we evaluate the temporal robustness of hate speech detection and monitoring systems for social media?

2. **RQ 2:** How can we investigate the rapid temporal shift of most debated topics on social media and leverage this information to gain more insight and eventually boost the temporal robustness of different hate speech prediction systems?

3. **RQ 3:** Hate Speech detection on social media is an online phenomenon that is rooted in offline real life, where socio-economic factors characterising geographical territories and people are key. How can we leverage information from traditional socio-demographics indexes about population of a country or region, and information on hate speech dynamics automatically extracted from social media to improve our understanding of interplay between economical and cultural factors and the expression of hate online?

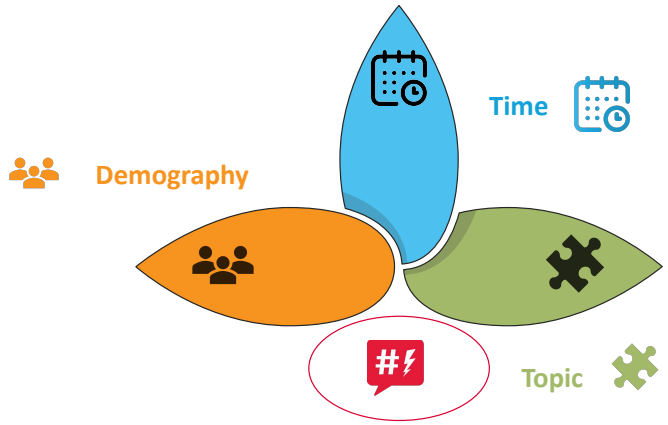


FIGURE 1.1: Thesis diagram.
Credits: Templated designed by PresentationGO.
Icon from Flaticon.com.

These research questions are addressed in the central sections, according to the following organisation.

RQ 1 In Section 3 I analyze how time plays a crucial role for hate speech detection systems and how it is possible to enhance the time robustness (and hence the accuracy) of hate speech prediction by carefully selecting the data during the training of the algorithms.

RQ 2 In Section 4 I explore the interplay of hate speech detection algorithms and the rapid topic shift in online debates. I argue that this shift in the most debated topic/news on social media is another critical factor for hate speech detection systems robustness because it causes changes in the language that prove to be difficult to be modelled and hence exploited to detect negative messages.

RQ 3 In Section 5 I will investigate some perspective on this dynamics, where I analyse the geographical distribution of anti-immigrants tweets in Italy and explore

the correlation with socio-demographic characteristics of the areas with a highest density of such messages. This is one of the three perspective on hate speech that I investigate in this thesis.

As one of the fundamental pillow of this research are traditional NLP techniques: the following list recaps the definition of the fundamental concepts of such field for a more immediate comprehension of the terms and abbreviations used in the following chapters.

Basic Glossary

- **Natural Language Processing (or NLP or NLProc)**: subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data
- **Sentiment Analysis**: the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral
- **Corpus**: a collection of written or spoken natural language material, stored on computer, and used to find out how language is used
- **Annotation**: the manual process of adding meta data to a corpus to facilitate an algorithm to find relevant patterns and inferences
- **Gold Standard**: an annotated corpus which has been manually labelled by humans and it is used to evaluate different NLP tools that perform the same task with different algorithms
- **Training Set**: Subset of the corpus that is used to train supervised machine learning algorithms to perform a specific NLP task
- **Automatic hate speech classification or prediction**: the process of using an algorithm to predict (or classify) a text as containing hate speech (or not)

A more comprehensive review of relevant literature is presented in the next Section, starting from a general overview of hate speech detection as an NLP task, with a specific focus on the challenges of building labelled corpora. Then there is a review of works in the realm of HS detection and an entire section dedicated to BERT as it is the foundation of the findings presented in Section 3. To provide a deeper background to this results I also analysed the relevant studies on diachronic corpora. The final two sections are dedicated to the class of algorithm for topic modeling (as it is the foundation for the work presented in 4) and to multidisciplinary work that cross boundaries between NLP and statistics over traditional socio-economic data, which is the research topic of the very first paper I published and that is the basis for Section 5.

Ethics Statement All Twitter derived data were treated in compliance with both Twitter API terms of service and, being based in Europe, GDPR policy.

The crowdsourced annotations were carried out on well established platforms and followed best practice for guidelines and examples for the workers. The monetary compensation for the performed tasks were established following the pricing for similar task on crowdsourcing platforms.

The demographic data about Italy were derived from ISTAT (the National Statistics Institute) and are anonymized and aggregated at the source

Chapter 2

Related Works

The research work presented in this thesis is an attempt to create a bridge between different research communities, namely NLP and computational social science, in order to tackle the arduous problem of hate speech detection on social media. In the previous introductory section I presented at a high level the scope of the research described in the central sections of this thesis, while in this section I will review the most relevant and/or recent works published in the different research field which are fundamental as a background knowledge for the findings presented later on. The very first inspiration for this PhD research was represented by [85], a paper at the intersection of natural language processing, computational social science and demography where the authors use words frequencies in geo-tagged Twitter data combined with other sources of demographic data such as health and census data to create a happiness score for urban areas. The techniques presented, combined with data availability and expertise in our research group led to my first published paper [51] that explores the interplay of hateful messages on Twitter in Italy against immigrants and traditional socio-demographic indexes. The full details are described later on in Section 5. As mentioned, this thesis represents a bridge between NLP and computational social sciences and hence we need to account for the major challenges in both disciplines. First and foremost, when using social-media derived data to study social behaviours, it is fundamental to acknowledge that users of a specific social media do not necessary represent a good sample of the general population, as explored in [80]. Therefore it is mandatory to use caution when trying to infer prediction on general population from conclusions derived from such sources of online data. Secondly, NLP relies on data structures as corpora and made available for specific purposes. It is hence crucial to investigate in depth the assumptions and techniques behind the process of creating of such resources and how these affects the results obtained. In our case, as all the dataset mentioned in this thesis were generated using Twitter API, it is worth mentioning this work [86] that, even tough it is not too recent, it dives quite deeply into the specificity and mechanics of the API and how different sampling setting affect the generated datasets. In the next sections I will examine the most relevant papers with respect to the findings presented in this thesis. On a strictly methodological point of view, I took inspiration from [75], even though with a slightly less strict approach, given the multitudes of possible publication venues to be considered potentially. For this reasons I relied on the three major search engines available: Google Scholar, DBLP and ResearchGate where I conducted queries for survey papers or meta analysis with appropriate keywords. As a general rule of thumb I only deemed as relevant papers with a sizeable number of citations or, in case of meta analysis, published in recent years and hence capturing the latest research development.

2.1 Hate Speech Detection as a NLP task

The research findings presented in this thesis approach hate speech detection as a specific Natural Language Processing task (from now on NLP) task [71, 87, 47]: a very fine grained application of sentiment analysis to microblogging data. The field of NLP is a fast growing research area, with a sizeable number of new publications in multiple venues, hence it is beyond the purpose of this chapter to draw a complete map of all the latest findings about all the multiple perspective on hate speech detection as very specific an NLP task. In the following I will present surveys on relevant specific macro-topics and highlights specific papers which are either relevant for the work described in this thesis either in terms of inspiration or technical foundations for the analysis I conducted.

Hate speech detection is a relatively new topic of investigation, applying artificial intelligence technology to monitor extreme, potentially dangerous manifestation of hostility online. The field has been extensively surveyed in 2017 [110], where around 50 works on hate speech detection were analyzed, while in 2018 [54] classified and described 128 documents on the topic. The vast majority of the surveyed papers describe approaches to hate speech detection based on supervised learning, where the task is treated as a sentence or message-level binary text classification task. The two surveys however conclude that the different models and features presented in the literature are very difficult to compare effectively because the results are evaluated on individual datasets that are often not public, hence the surveys advocate for a wider availability of publicly available data. This evaluation gap is being bridged recently by evaluation campaigns for English, Spanish (SemEval [16]), German [113], and Italian (EVALITA [24]), whose shared tasks released annotated datasets for hate speech detection. The availability of benchmarks for system evaluation and datasets for hate speech detection in different languages made the challenge of investigating architectures, which are also stable and well-performing across different languages, an exciting issue to research [36, 91].

2.1.1 The challenges in building labelled corpora

The goodness of hate speech predictions for a specific task, performed using supervised methods, is heavily linked to the existence of a relevant and high-quality gold standard [122]. This entity is basically a set of manually annotated linguistic data, manually revised to provide the best possible correct labeling of texts for a specific task. It provides a gathering of shared knowledge on a very specific topic (e.g.: misogyny on Twitter, use of abusive language towards immigrants on Reddit, stance of British Twitter users on Brexit,...). The process of creating a gold standard is schematised in the following Figure 2.1.

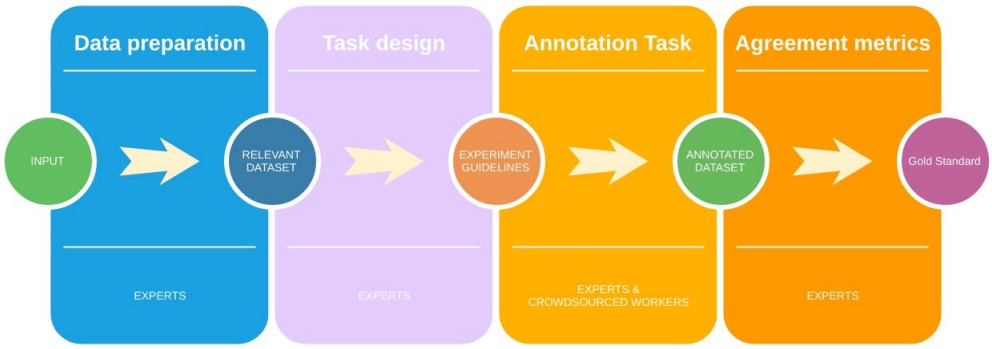


FIGURE 2.1: The process of creation of a gold standard.

Data preparation The first step consists in choosing the best possible data to describe the phenomenon of interest. This phase tends to be more crucial than what it is perceived intuitively, given the diversity of sources that became recently available. In here the focus will be on social media derived data, as are the base of the analysis described, but every field presents multiple challenges that need to be precisely addressed. First off, social media data are proprietary and hence their public accessibility is strictly regulated by data-owners rules. As a practical example, Twitter API allows free daily access to 1% of posted tweets, and allows a language and keywords based filtering. It is straightforward to see how this design can lead to important biases in the datasets and it is of foremost importance address them clearly before presenting any research findings. Some of this issues are extensively addressed in general in [86], while the specific challenges of sentiment analysis with such resources are described in [115].

Task design After having prepared the data in a suitable form to be annotated, specific decision on the annotation process need to be made by experts (for a general descriptions of the steps see for example [107, 64]. The first obvious one is based on the size of the dataset itself. If it is not too big, a small number of experts may be enough to carry out the task, on the contrary the support of crowdsourcing platforms may be instrumental to this purpose. The second factor affecting the task design is the nature of data itself: in case of sensitive information the process must be compliant to GDPR or similar privacy-related normative and hence either interfere with the use of crowdsourcing platforms or require an additional step of proper data anonimization. If this is the case, an evaluation of the potential impact of this requirements on the final results is clearly fundamental. A crucial step for a successful annotation experiment lies in the preparation of relevant, complete and clear annotation guidelines, both in cases of expert or crowdsourced workers, as it provides a consistent and well specified framework to deal with doubts about the task.

Annotation Task The ultimate goal of the annotation and the level of difficulty plays an important decisive role in determining which specific skills are essentials for the annotators. It is safe to say that the decision of relying just on experts, just on crowdsourced workers or a combination of both strictly depends on the context and has to be tailored every time on the specific task. The main pros and cons are summarized in the following Table 2.1

A set of experimental design to test this dichotomy is presented in details in [111], where the finding highlight and confirm the complexity of the issue. It is important

Experts	Crowdsourced workers
less prone to random annotation	more prone to random annotation and behaviour
allegedly higher consistency	relatively poor consistency
demographic control	poor demographic control
expertise in annotation methods	less prone to complex task
professional biases	less biases due to their work
not all experts are the same	potentially broader diversity in native language
no big dataset	good for massive datasets

TABLE 2.1: Pros and Cons of experts VS crowdsourced workers in annotation tasks.

to notice that the two settings are not at all mutually exclusive and in reality a mix if approach is often adopted, with different possible schemes. Aside from technical reasons, the use of crowdsourcing platforms comes with a set of criticalities that need to be evaluated case by case. The most relevant of them being:

- **Ethical Issues**: workers on such platforms are paid very little money and very often come from under-developed countries (see [53], [45] and more recently the 2021 ACL Code of Ethics[1])
- **Biases**: certain tasks may be affected by workers' biases, based on certain specific demographic characteristics (see [121])
- **Selection**: not all platforms allow a fine tuning of workers' selection based on their characteristics

Moreover the distribution of annotation loads across all the different annotators can follow different schemes (as described for example in [94]). The main ones include:

- **Joint annotation**: dataset is divided into non-overlapping subsets and are then annotated by different annotators. Every annotator annotates different data
- **Parallel annotation**: each worker annotates all the data
- **Mixed annotation**: a first joint annotation + agreement evaluation + parallel annotation

In conclusion the task design is a very complex and multi layered process that needs to be fine tuned for every specific NLP task and there is no rule of thumb that guarantees a satisfying outcome in each and every situation.

Agreement metrics The final step in creating the gold standard is based on the choice of the right agreement measure for the task, which is a very complex index that has to be able to capture three main aspects of the process (as explored for example in [88], [127], [64]):

1. **complexity** of the task
2. **performance** of the annotators

[1] https://2021.aclweb.org/calls/papers/#ethics-policy

3. **goodness** of guidelines and task design

Various indexes have been proposed over the time, from the simple percent Agreement which captures the simple percentage of cases where all the annotators agree in one label to more complex indexes such as Cohen's K [35], Fleiss's K [49],and Krippendorff's α [77].

In recent years the concept of gold standard has been questioned in depth especially in the context of high subjectivity annotation task [78, 3], where the disagreement between annotators may actually be a source of knowledge enrichment rather than just pure noise. In this sense, the paper published in 2015 by Aroyo and Welty [8] is a milestone in this new approach. The authors discuss all the foundation of annotation and propose a new paradigm based on two principles:

- **rejection** of "single truth" fallacy for semantic interpretation
- **vectorial** representation of sentence disagreement and workers' answers

Moreover multiple authors have been vocal in discussing the role and opportunity of annotation harmonization (for example [76], [13], [76]), where several alternative approaches are reviews, including a guidelines redesign and one or more rounds of annotation experiments until a consensus (i.e.: a total agreement) is reached and the release of an aggregated set for the gold standard, which completely fails to capture all the nuances that emerged during the process.

A more extensive exploration of the literature on this topic is beyond the purpose of this section, but a literature review was compiled in [64]. This section was prepared as a reduction of the materials presented as a Tutorial at IC^{2}S^2 2018, the 5th International Conference on Computational Social Science, which are publicly available at `http://www.di.unito.it/~florio/tutorial.html`. A further selection of relevant paper and a call to action to overcome the issues of traditional annotation method are available as part of "The Perspectivist Data Manifesto"[2].

2.2 Hate Speech Detection Methods

The number and variety of algorithms that have been exploited over time to detect hateful messages has changed over time, following both the evolving of computational capabilities of machines and the increasing amount of linguistic data available for training. Earlier methods include term frequency–inverse document frequency (TF-IDF vectors), Parts-of-Speech tags and linguistic features (as described in [39]). In the work presented in Section 3 and 5, texts are classified as hateful or not by means of a Support Vector Machine (SVM from now on). SVMs belongs to the family of supervised machine learning algorithms and is commonly used to classify data into two independent classes, which very often consists of text classification [72, 116]. In particular, the text, adequately encoded into its vectoral representation (e.g., TF-IDF [103], word-embedding [84]) is provided as training to the model in order to generalize the weight of the equation of a hyperplane which is able to divide the examples into the given classes. During the evaluation phase, when the text labels are unknown, the model applies the learned discrimination model for labeling the test examples. The SVM algorithm family is divided into two main classes: linear models, which represent the division of data into classes by means of a simple straight "line", and polynomial algorithms, which implement more sophisticated equation

[2] `https://pdai.info/`

to perform the same task in more complex scenarios. The strategy used for the construction of hyperplane is commonly known as the kernel function. A commonly used kernel is RBF (Radial Basis Function) [33] which in general shows good performance for many NLP tasks [58, 73].

A way more complex and powerful word embedding algorithm, word2vec [56], was introduced in 2013. It is a shallow neural network that has as input a large corpus of text and generates as an output a vector space, where each vector in said space is allocated to each unique word in the corpus. In simple words, a supervised neural network can be thought of as a black box with a learning and a predicting method. During the learning (or training) phase, the model receives both the inputs and the desired outputs and updates its "internal state" accordingly. During the "prediction" phase, the model takes input data and generates an output which as as close as possible to the desired one. The details have been extensively explored in literature, starting from [60, 6, 46], for example. Based on this, several other neural networks were trained for very specific tasks, including hate speech detection. A review of the finding based on these new models is presented in [10]. The following Figure presents the evaluation metrics of the most common hate speech detection metrics, finding that rule-base clustering (based on an incremental clustering) is the most effective one.

Table 16
Comparative analysis on hate speech detection methods.

Method	AUC	Accuracy	Precision	Recall	F1-score
Baseline Naive Bayes	0.71	0.87	0.84	0.87	0.85
Davidson et al. [28]	0.87	0.89	0.91	0.9	0.9
DL-Metadata only	0.75	0.61	0.80	0.61	0.66
DL-Text only	0.91	0.87	0.89	0.87	0.88
DL-Text & Metadata (Naïve Training)	0.90	0.87	0.89	0.87	0.88
DL-Text & Metadata (Tran. Lear.)	0.91	0.87	0.89	0.87	0.88
DL-Text & Metadata (Tran. Lear. FT)	0.90	0.87	0.89	0.87	0.88
DL-Text & Metadata (Interleaved)	0.92	0.90	0.89	0.89	0.89
Rule-based clustering	**0.97**	**0.95**	**0.93**	**0.92**	**0.93**

The results are obtained on the same computational platform.

FIGURE 2.2: Evaluation Metrics listed in [10].

A more recent meta analysis published in 2020 [7], after having briefly described some of the most popular datasets([119], [39] and [16]) offer a critical review of this task from a computational perspective. The authors question in details the validity of state of the art methodology presented in [11] and [2], focusing in particular on the risks posed by oversampling and class imbalance and proposing to alleviate the issues by a combined approach of re-sampling data, explore cross-lingual experiments and infuse the algorithms with features extracted from meta information.

2.3 BERT and AlBERTo

BERT is a groundbreaking task-independent language model [42] based on the idea of creating a deep learning architecture. The main aim of BERT is to tackle the limited availability of annotated training data for NLP tasks by means of pre-train a general-purpose language representation models directly on the unannotated text, because datasets of this kind tend to be much larger and more easily available. BERT relies on the latest development in pre-training contextual representations, such as

Semi-supervised Sequence Learning [37], ELMo [93], ULMFit [65], OpenAI Transformer [101] and Transformers [117] while implementing a deeply bidirectional architecture. More specifically it encompasses encoder and a decoder, so that the encoding level can be used in more than one NLP task while the decoding level contains weights which are then optimized for a specific task (fine-tuning). This system represents a meaningful improvement from previous techniques because it combines two crucial features: context awareness and bidirectionality. Context awareness means that the model creates a representation for each word in the dictionary based on the other words in the sentence, while bidirectionality indicates that the model predicts a word based on both what precedes and follows the term subject of prediction. The idea behind such models is that if a model can predict the next word that follows in a sentence, then it can generalize the syntactic and semantic rules of the language. Being computationally very expensive, researchers only recently succeeded in training BERT deep neural networks. BERT [42] was developed to work with a strategy very similar to GPT [102], hence the basic version is trained on a Transformer network with 12 encoding levels, 768 dimensional states, and 12 heads of attention for a total of 110M of parameters trained on BooksCorpus [128] and Wikipedia English for 1M steps. The main difference with GPT lies in the learning phase, which is performed by scanning the span of text in both directions, from left to right and from right to left. This strategy is however not entirely a novelty as it was previously implemented in BiLSTMs [66]. Moreover, BERT uses a "masked language model": during the training, random terms are masked to be predicted by the net. Jointly, the network is also designed to potentially learn the next span of text from the one given in input. These variations on the GPT model enable BERT to be the current state-of-the-art language-understanding model. Larger versions of BERT (BERT large) have been released and are scoring better results than the normal scale models, but they require far more computational power.

Considering the international focus on language models generated through deep neural networks and their lack for the Italian language, AlBERTo has been proposed as a valid resource to fill this gap, as it was developed starting from the BERT base model. AlBERTo has been trained on *TWITA* [15] a collection of domain-generic tweets in Italian extracted through API streams and free to use for research purposes. It was then fine tuned on a datasets that encompassed also the train and reference set from Haspeede [24], the first shared task on hate speech on Italian organized within EVALITA2018 evaluation campaign [3]. More details about AlBERTo are available in [97, 99, 98].

2.4 Diachronic studies of social media derived data

Computational approaches to the diachronic analysis of language [114] have been gaining momentum over the last decade. An interesting analysis of the dynamics of language changes has been provided by [55]. The authors describe what happens from the language analysis point of view on words that change their meaning during the time. Most of them show a *social contagion* where the meaning is changed by their common/wrong use on social media platforms. Clyne et al. in [34] discuss the changing of words meaning by the influence of immigrant languages, Lieberman et al. [79], instead, try to quantify these changes in the common language. These studies support our idea about the possible difficulties of an automatic machine learning approach to classify new sentences that have been collected in a time distant enough

[3] http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

from the one of training data. We suppose that the language of hate speech is very volatile and influenced by events, and it changes words meaning faster than usual: all these considerations have encouraged us to investigate the robustness of some machine learning models over time.

The recent availability of long-term and large-scale digital corpora and the effectiveness of methods for representing words over time played a crucial role in the recent advances in this field. However, only a few attempts focused on social media [43, 12], and their goal is to analyze linguistic aspects rather than understanding how lexical semantic change can affect performance in sentiment analysis or hate speech detection. From this perspective, our work represents a novelty: for the first time, we propose to tackle the issue of diachronic degradation of hate speech detection by exploring the temporal robustness of prediction models. The closest works found in recent literature are [68], where the authors explore the diachronic aspect in the context of user profiling, and [61], who provides a broader view on diachronicity in word embeddings and corpora. Nevertheless, this is the first work investigating the diachronic aspect in the specific context of hate speech detection, which is a crucial issue, especially in application settings devoted to monitoring the spread of the hate speech phenomenon over time.

2.5 Topic Modeling as a classification method

Our most powerful classification tool is by far represented by topic modeling, and the most comprehensive recent review is presented in [4]. This work reviews the most common methods for topic modeling (such as Latent semantic analysis, Latent Dirichlet allocation and others) and for topic evolution (among them the topic over time, the dynamic topic models and a few other significant ones). Other less recent survey on this classification technique include [38] and [19]. The basic assumption of a topic modeling algorithm is that documents consist of a combination of *topics*, defined as a collection of semantically related *words*. One of the fundamental techniques of topic modeling is the Latent Semantic Analysis (from now on, LSA): the core is a documents-terms matrix that is then decomposed into a document-topic matrix and a topic-term matrix. The most basic version relies on a simple raw word counts but this approach is very weak as it does not take into account the significance of each word. To correct this issue it has been proposed the use instead of a term frequency-inverse document frequency (tf-idf score) that introduces weights to account for words that occur frequently in a document but much less frequently across the whole corpus. It is very likely that the resulting matrix is a very sparse, noisy and redundant across many dimensions. For all these reasons a dimensionality reduction could prove useful: the model tackle the issue with a so called truncated singular value decomposition (SVD) that, following the rules of linear algebra, rewrites the matrix as a product of 3 other matrices, one of them is a diagonal one with the singular values of the original one.

The subsequent step in topic classification consists of a measurements of cosine similarity between documents vectors and term vectors to determine the similarity of such entities. This method is certainly quick and efficient to be implemented but it has some major setbacks: namely the embeddings are very difficult to interpret and accurate results need really large datasets, which are problematic to find in NLP. Some of these issues were tackled with the evolution of LSA into the Probabilistic Latent Semantic Analysis model (or pLSA) [63]. It introduced a probabilistic approach to topic modeling basic assumptions, considering the probability of seeing

a given document and a given word together, given a certain distribution of topics in each document. This new model is an improvement but again presents two main criticalities: there are no rules for assigning probabilities to new documents added to the corpus and the number of parameters, which increase with the number of documents, lead to the risk of over-fitting. All these mentioned criticalities are addressed by the so called Latent Dirichlet Allocation (or LDA) model, which could be described as a Bayesian version of pLSA. It was first introduced by Blei [22] in 2003. It essentially provides a method for creating samples of probability distribution for both the topics, and for the word distribution of each topic, both draw from a Dirichlet Distribution (hence the name). This model extracts topics in a human-interpretable form and is by construction more robust when deployed on unseen documents and therefore it has become very popular and applied in different contexts. A selection of the models derived from this one and various application are presented in [70]. In the work presented in Section 4 the focus was on temporal evolution of topics, and to achieve this goal the simple LDA was not enough. We hence decided to apply a Dynamic Topic Modeling to our data, which was first proposed by Blei [21] in 2006. This new model incorporates the idea of topics evolution over time, and specifically over custom-defined time intervals, giving the user the flexibility to adapt the model to the context in which it is adopted. The outcome is a sequence of LDA models where each topic is a sequence of related distributions over terms as the overall topic distribution and the term distribution for each topic differ sequentially depending on the time slice. This approach is indeed very powerful when the aim is the analyze for example the evolution of the debate over certain subjects over time, as it is our case. Last but not least consideration to be made on our case is that our data are made up of tweets, and being a specific type of short texts, they pose specific challenges which are investigated by Qiang et al. in [100].

2.6 Leveraging social media data and traditional indexes to study online reaction to immigration phenomena

The growing availability of massive user-generated datasets from social media and the fine tuning of techniques and computational power to collect, store and analyse them has paved the way to integrate these resources with more traditional demographic data (see [120] for a survey and [125] for an application to the topic of immigration). There are multiple sources of online data that can be combined to study migration phenomena and their effects on society with a new perspective: in [83] the role and benefit of Google+ data are being analyzed, while in [124], [44] and [112] authors leverage Facebook data to study migrations and migrant integration. Facebook has also been employed as source of data for hate detection in Italy [41]. A considerable number of works (including our own) are based on Twitter data, mainly due to their relative ease of access. In [48] Twitter is used to model the differences between migration and short-term mobility. In [123] the authors use Twitter to infer international and internal migration patterns. In [31] the focus is instead on the use of hashtags in Twitter in the effort to track racism online. Finally, in [81] Twitter is used to explore opinions and semantic orientations related to parenthood in Italy.

Abstracting away from the specific social network platform, the use of social media data presents a series of unprecedented challenges and specificity, raging from ethical concerns, to methodological issues and the crucial role of possible biases. The authors of [89] present a very detailed examination of the different limits of social dataset and methods and the connected ethical challenges. At first they highlight

that there is a lack of diversity of researchers in the field at various levels, from practitioners to peer reviewers and funding that reflects on to which research problems are more often tackled and/or funded. Then the authors argue that a precise quantification of data biases is and will remain a crucial issue. Nevertheless it is also argues that it could also be a desirable features in certain specific contexts, such as when the boundaries define the applicability of the solutions or may play a key role in design choices, for examples. The specific case of annotations for crisis scenario is the focus of [67] where Imran and coauthors showcase a Twitter corpora with two different sets of annotations, a topic-based categorization and a vocabulary-based tagging. Both the approach are validated by machine-learning classifiers and the authors argue that building this specific corpora is crucial for processing crisis-related messages and therefore extract information which can prove to be potentially very useful for humanitarian organizations. In a similar work [106], the reliability of annotations is measures in the specific context of the European refugee crisis. The author claim, as we stated previously, that specific guidelines play a key role in the process. This work also highlights the need of a multi-label classification to take into account users opinions as a complex message that does not fit a simple binary classification task. Another approach to improve annotation reliability is proposed in [5] where the authors argue how the use of text from suspended Twitter account could improve the outcome of the labelling process. Two random forest classifiers were trained on data from suspended and non suspended accounts and the first case led to a more accurate hate speech prediction. This result is certainly interesting but it intersects with the issue of data decay and more specifically the public availability of suspended account data, in light of the news API policy changes, and this shows the variety and complexity of challenges in such a multidisciplinary field of study. [4]

2.7 Summary

In this section I have presented an overview of recent literature which is fundamental as a background to the results that will be presented in the next three sections, which are the outcome of the research activities carried out during the last three years of this PhD. This section opens with mentioning the paper that served as an inspiration for the very first piece of work I published in 2019. I then moved to present a couple of works that highlight the most crucial challenges when deriving data from Twitter, as it is the source of the main resource of data used in this thesis. As I approached the detection of hate speech as an NLP task, I provided an overview of the pertinent literature and the recent development such as the latest shared task. At the core of every NLP task there is a dedicated corpora whose quality is crucial for the goodness of the findings. For hate speech detection it is fundamental the availability of a labelled corpora that serves as a gold standard: I hence relied on the material that I had prepared for a tutorial at IC^2S^2 to investigate all the phases and challenges related to the creation of such a resource. The second crucial aspect of hate speech detection is the algorithms used to predict which texts are hateful and which are not. I provided an overview of the most commonly used ones, with a specific focus on BERT, as it was a true groundbreaking paradigm when first introduced in 2019. The last three subsections provide a background for the three perspectives that we adopted to investigate this phenomenon: time, topics and demography. For what regards time, I reviewed the most relevant works

[4] https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api

that consider a diachronic study of social media. The subsequent section explores the most common topic modeling algorithm as a classification methods, while the last one is dedicated to integrate NLP measures with leveraging more traditional demographic data to create new indexes that allow us to study hate speech in the context of online debate about immigration and the presence of immigrants in Italy. We chose this specific perspective due the availability of annotated corpora and resourced specifically dedicated to this topic. The next Section explores the first the perspective we adopted: the role of time in hate speech detection and the impact on the performances on prediction and monitoring systems.

Chapter 3

The challenge of temporal robustness in hate speech detection and monitoring systems

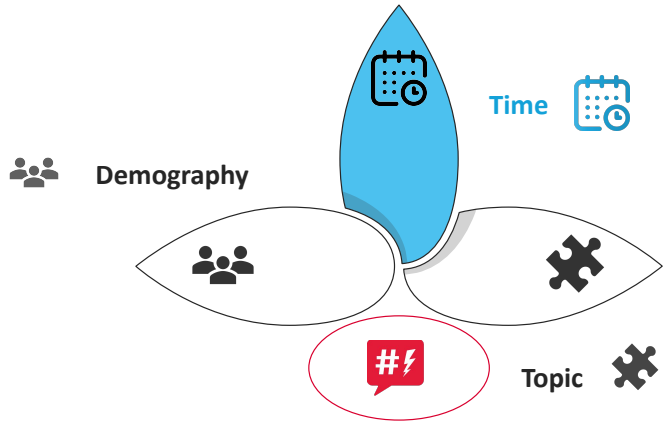


FIGURE 3.1: Thesis diagram.
Credits: Templated designed by PresentationGO.
Icon from Flaticon.com.

In the introduction of this thesis, in Section 1, we touched on how the recent abundance of linguistic data, especially from social media, allows the research community to tackle more in-depth long-standing questions such as understanding, measuring, and monitoring users' sentiment towards specific topics or events. However, this abundance of data has also contextually brought new challenges regarding the interplay of algorithms used to explore such a task (as described in Section 2) and the specific characteristics of these newly available data. In the specific case of social media, we can intuitively infer that most of the users tend to react to and discuss breaking news from different sources of media outlets very spontaneously, therefore language of the public discourse it is often characterized by significant variations over time in terms of topics and linguistic patterns. This issue will be tackled more in depth in Section 4.

These considerations suggest that there is a deep need to precisely measure the robustness of algorithms over time, and its interplay with training data. All the finding presented in this section were published in 2020 in [52].

3.1 The motivation for temporal robustness investigation

The motivation and the urgency for a diachronic study (i.e.: a focus on the evolution) arose at first from observing the difficulties that were encountered in the development of hate speech monitoring platforms such as "Contro l'odio" [27], an online platform launched in 2018 to predict and monitor hate speech messages against immigrants posted on the Italian Twitter. Monitoring and countering hate speech is a shared goal of several recent projects, which focused on different targets of hate, monitoring different countries and territories, and differ in the granularity of the detection, of the temporal spans considered, and regarding the visualization techniques provided to inspect the monitoring results. Let us mention the CREEP project on monitoring cyberbullying online [82], with an impact also on the Italian territory, HateMeter[1], with a special focus on Anti-Muslim hatred online in different European countries (Italy, France, and England), and the MANDOLA project [92] providing an infrastructure enabling the reporting of illegal hate-related speech.

The platform "Contro l'odio" [2] combines computational linguistics analysis with map-based visualization techniques. It offers a daily monitoring of hate speech against immigrants in Italy and its evolution over time and space to provide users with an interactive interface for exploring the dynamics of the discourse of hate against immigrants in Italian social media. Three typical targets of discrimination related to this topical focus are taken into account: migrants, people belonging to religious minorities, and Roma, since they exemplify discrimination based on nationality, religious beliefs, and ethnicity.

Since November 2018, the platform analyses daily Twitter posts and exploits temporal and geo-spatial information related to messages to ease the summarization of the hate detection outcome.

The automatic labelling of the tweets of the "Contro l'odio" platform is performed by a Support Vector Machine (SVM) classifier. It was trained on data from 2017 and then tested on messages streamed from October 2018 up to today. It is clear that for a service like this to be dependable and consistent over time, there is a need to explore in-depth the interplay of language and topic shifts in time and the robustness of the prediction system. We hence propose a novel approach to tackle the issue of diachronicity in hate speech prediction, by means of a transformer-based neural network classifier, AlBERTo, which is trained on Italian social media language data. AlBERTo provides a pre-trained language model of Italian, and it is fine-tuned on monthly samples from the "Contro l'odio" dataset to be able to classify instances of hate speech. In this section we introduce an evaluation of strategies to alleviate the diachronicity issue by tackling the following research questions:

RQ1 How can we evaluate the temporal robustness of different hate speech prediction systems, with respect to language and topic change over time?

RQ2 What is the impact of the size and temporal coverage of the training set on the temporal robustness of the prediction?

[1] http://hatemeter.eu/

[2] The platform is online and can be accessed at https://mappa.controlodio.it/

3.2 Methods and models

We designed a series of experiments to evaluate several strategies for hate speech detection in a diachronic setting. Individually, all the experiments follow the same structure, where a classifier is trained or fine-tuned on a training set and tested on a smaller test set. To test the robustness of prediction models against changes in language and topic over time, we trained our models in two different scenarios and then compared the performance. In the first case, we used training data from *one single month*, while in the second case, we *progressively increased the size of the training* set by injecting information about the history of the corpus and hence the evolution of language and topics over time. We compared the performance of different models in terms of precision, recall, and F1-score. We focused on these metrics relative to the *positive class* (the presence of hate speech), because the task at hand is a *detection* task, as opposed to a *classification* task. To smooth out any possible statistical anomaly, we ran every prediction for five times, each with a different seed for the random number generator, and then we averaged the metrics over all the runs. We employed a series of test sets drawn from the "Contro l'odio" dataset [27]: each one is a sample covering one month of Twitter messages.

We focused on the use of two very different classification models: a Support Vector Machine (from now on: SVM) [57] and AlBERTo, the Italian BERT language model [97].

The core contribution relies on the exploration and evaluation of how the distance in time between a training and a test data impacts the performance of two models who display profound differences in how they were built and how they work when performing classification tasks.

SVMs belongs to the family of supervised machine learning algorithms and is commonly used to classify data into two independent classes, which very often consists of text classification [72, 116]. In particular, the text, adequately encoded into its vectoral representation (e.g., TF-IDF [103], word-embedding [84]) is provided as training to the model in order to generalize the weight of the equation of a hyperplane which is able to divide the examples into the given classes. During the evaluation phase, when the text labels are unknown, the model applies the learned discrimination model for labeling the test examples. SVMs are fast, and they perform well also with a limited amount of data. To better understand the way SVMs work, it can be possible to imagine elements of two classes plotted on a 2-d space. In such a scenario, the model can find a "line" that optimizes the separation of examples into the given classes. The SVM algorithm family is divided into two main classes: linear models, which represent the division of data into classes by means of a simple straight "line", and polynomial algorithms, which implement more sophisticated equation to perform the same task in more complex scenarios. The strategy used for the construction of hyperplane is commonly known as the kernel function. A commonly used kernel is RBF (Radial Basis Function) [33] which in general shows good performance for many NLP tasks [58, 73]. The SVM model has been implemented using the LibSVM java library[3] [29]. We used the simplest version available: a linear version of the kernel and a value of the parameter C equal to its standard value of 1. As already mentioned, our main goal was to observe the influence of the temporal distance among training and test data in the performance of supervised machine learning models. Consequently, we were not interested in obtaining state-of-the-art

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

results in the accuracy of classification. BERT [42], instead, is a groundbreaking task-independent language model which represents the current state-of-the-art among language-understanding models.

More specifically it encompasses an encoder and a decoder, so that the encoding level can be used in more than one NLP task while the decoding level contains weights which are then optimized for a specific task (fine-tuning). For this reason, a general-purpose encoder should be able to provide an efficient representation of the terms, their position in the sentence, context, the grammatical structure of the sentence, and semantics of the terms. The idea behind such models is that if a model can predict the next word that follows in a sentence, then it can generalize the syntactic and semantic rules of the language. Considering the international focus on language models generated through deep neural networks and their lack for the Italian language, AlBERTo has been proposed as a valid resource to fill this gap, as it was developed starting from the BERT base model. AlBERTo has been trained on *TWITA* [15] a collection of domain-generic tweets in Italian extracted through API streams and free to use for research purposes. More details about this model are available in [97, 99, 98].

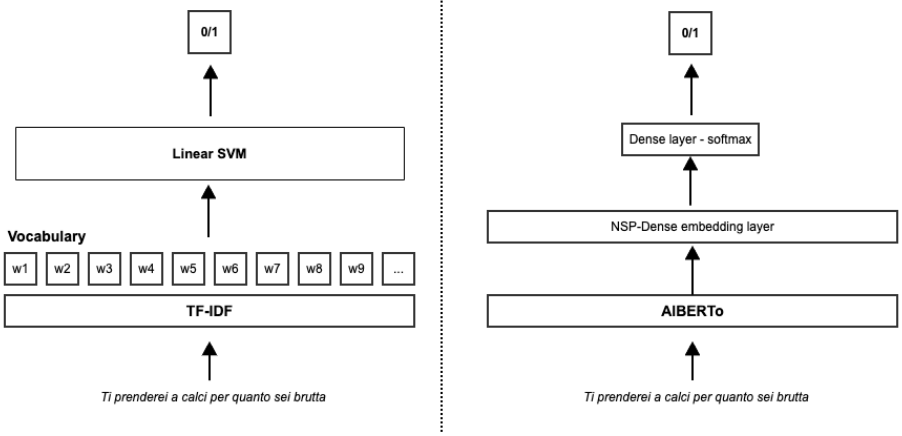


FIGURE 3.2: On the left it is showed the SVM model, on the right the one based on AlBERTo.

We implemented AlBERTo using mainly Tensorflow [1] and Keras[4], the famous deep learning library. The performance was evaluated with the metrics provided by scikit-learn[5]. The fine-tuning of the AlBERTo model for the specific classification task was performed predominantly on Google Colab using a TPU. The evaluation phase has required only a GPU on the same platform. Google Colab has been chosen as the running environment because, at the moment, it represented the most efficient and powerful cloud computing platform available for training deep learning models for free. During the fine-tuning phase of AlBERTo, we estimated the number of learning epochs as a result of an empirical evaluation carried out on a validation subset made of 200 sentences extracted from the same data distribution used in training and testing. Starting with 2 epochs, we increased the number by two at a time up to 10. From the results of this setting, we observed that the best performance equals to 0.518, considering the F1-score on the positive class, was obtained by setting the number of epochs to 8. This value was used as a fixed parameter for all the fine-tuning processes. The learning rate has been kept at its default value of 2e−5, while

[4] https://keras.io/
[5] https://scikit-learn.org/

the training and prediction batch size was set to 512 to improve the predictive accuracy of the model as much as possible. Since we mainly worked with short texts, we decided to leave the pre-defined maximum input size of 128 tokens. The fine-tuned version of AlBERTo has been used as part of a standard Keras classification model. In particular, as shown in Figure 3.2, we collected the embedding representing the input from the NSP-Dense layer of AlBERTo, i.e., the first dense layer after the CLS token embedding. Then we stacked on it a final dense layer with a SoftMax activation function in order to predict the probability that the sentence may be a hate speech (class equal to 1) or not (class equal to 0). The details on training and test datasets for the algorithms will be discussed in the next Section 3.3.

3.3 The datasets

As anticipated in the previous section, I will present in the following the training and test data used in our experiments. The most prominent characteristics is that both the training and the testing datasets are drawn from the same source: TWITA, a large-scale collection of Italian tweets started in 2012 and currently ongoing. It relies on the Twitter Streaming API to download a sample of messages in Italian posted each day. It is important to point out that, even though our sets are drawn from the same source, our training data originate from two different sets of annotated data, which from now on will be collectively referred to as Haspeede+.

The TWITA-based training subset Our data source is the TWITA collection, but, although it counts over half a billion tweets between 2012–2017, the subset targeted for our training purposes is much smaller, resulting from a topic-based selection. The complete TWITA dataset was hence filtered by means of a handcrafted list of keywords, extending the work from two previous works of our research group, described in [95, 109]. The baseline for this extension where a set of terms was compiled by assessing which minority groups are most likely to be targeted for HS in the online discourse about immigration. This choice was based on the results of the 2015 Eurobarometer Survey on discrimination in the EU[6]. As shown in the Table 3.1 the list consists of terms chosen to be as neutral as possible, in order to collect the largest possible number of tweets. At the same time, negative epithets were excluded in order to avoid biases towards negative comments. We further expanded this collection of terms using the Open Multilingual Wordnet (a large lexical database of words in different languages) [7] not only to collect a larger number of messages but also to capture a wider variety of expressions on this topic. For each of the these keywords we collected from OMW all the related hypernyms (i.e., a word with a broad meaning constituting a category into which words with more specific meanings fall; for instance, the term *foreigner* is a hypernym of *refugee*) and co-hypernyms (i.e., two different words that share the same hypernym; for instance, *refugee* and *migrant* are both hyponyms of *foreigner*). We then manually cleaned this new set of keywords to retain only the most relevant ones. The final list of our keywords is presented in Table 3.1.

When filtering out text data it is crucial to take into account the possibility of substring matching, e.g. “Rom" would match “Roma" (the capital city of Italy). We address this issue by implementing regular expressions in our filtering algorithm to

[6] http://ec.europa.eu/justice/fundamental-rights/files/factsheet_eurobarometer_fundamental_rights_2015.pdf

[7] http://compling.hss.ntu.edu.sg/omw/

Keywords in Italian	English translation
clandestin*	*illegal immigrant(s)*
corano	*Quran*
emigrant*	*emigrant(s)*
emigrat*	*emigrant(s)*
esul*	*exile(s)*
fondamentalist*	*fundamentalist(s)*
imam	*imam*
iman	*imam*
immigrant*	*immigrant(s)*
immigrat*	*immigrant(s)*
Islam	*Islam*
islamismo	*islamism*
islamit*	*Islamist(s)*
maomettan*	*Mohammedan(s)*
migrant*	*migrant(s)*
migrazione	*migration*
mussulman*	*muslim(s)*
mussulmanesimo	*Islam*
musulman*	*muslim(s)*
nomad*	*nomad(s)*
profug*	*refugee(s)*
sfollat*	*displaced*
stranier*	*foreigner(s)*

TABLE 3.1: List of keywords in Italian (and their English translation) used to compile the dataset.
Asterisks * stand for the different combination of word endings in Italian, e.g. clandestin* represents clandestina, clandestino, clandestine and clandestini.

match only entire keywords preceded and followed by white-spaces, punctuation or beginning/end of a sentence. Our approach is purely string-based, therefore we could collect tweets containing keywords occurring with a different meaning from expected, such as named entities (e.g., “Nomadi" is the Italian form for "nomads", but also the name of a popular band). However, upon manual inspection we noticed that these occurrences are extremely rare in our collection.

We then draw a selection of 3809 tweets from 2015 and 3200 from 2017 using a list of topic-based keywords and imposing the constrain of the tweet having a geotag in Italy, as we are interested in linguistic phenomena in Italian. For this reason, we need to select only the tweets that carry geolocation among their metadata. We extract the GPS coordinates encoded in the metadata of each tweet and apply a *reverse geocoding* procedure to obtain four codes respectively related to four administrative levels: country, region, province and city by means of territorial boundaries released in a public database[8]. The reverse geocoding works by matching the geographical point given by the GPS coordinates to one of the polygons defining the Italian municipalities. The same database is then used to link the municipality to its province and region.

[8] https://www4.istat.it/it/archivio/209722

The tweets were then annotated by three independent contributors on Appen (formerly Figure Eight), chosen to be Italian speakers and geolocated in Italy. The annotation guidelines and operational definition of hate speech are described extensively in [95]. The annotation process on the crowdsourced platform involved all the categories mentioned in this paper (hate speech, intensity of the hate speech, irony, stereotype, aggressiveness and offensiveness) but for the analysis presented in this section we only considered the classification for hate speech. As multiple annotators worked on our dataset, each tweet had a so-called confidence score that captures the level of agreement between multiple contributors and indicates the "confidence" in the validity of the labelling. This index, based on Krippendorff's α metric, also incorporates a weighted average over the annotators' trust scores that tracks the dependability and consistency of each annotator's labelling history over time on the platform.

The Haspeede training dataset The second set consists of both the training and reference dataset of the Haspeede (Hate Speech Detection) shared task, organized within EVALITA 2018 [24]: a total of 4000 tweets (3000 tweets in the training test and 1000 in the reference set) collected from 1 October 2016 to 25 April 2017. These tweets were annotated with a mixed procedure: a subset was manually annotated by five independent experts to create a gold standard while the rest of the data was crowdsourced on Appen[9] (formerly known as Figure Eight) following a set of shared guidelines that will be described next.

The Annotation Scheme The whole training and test dataset was annotated following exactly the same scheme and guidelines, both from the expert annotators and from the crowdsourcing service workers, to guarantee a uniformity in the results. The annotation scheme encompasses seven different categories described in Table 3.2. The rationale behind this choice is that hostile messages can vary in intentions, form and intensity and this scheme was elaborated to capture as many as possible of all the different nuances of the linguistic phenomena. The full annotation process (including information on the aforementioned inter-annotator agreement) is presented in [95, 109].

Category	Labels	Description
Targets	Ethnic group - Roma	the message is directed towards a specific minority group
Hate Speech	no - yes	the message contains a target and an action towards this specific target
Aggressiveness	no - weak - strong	aggressive or harmful message, or even instigation, in various forms, to violent acts against a given target
Offensiveness	no-weak-strong	potentially hurtful effect of the tweet content on a given target
Irony	no - yes	the message contains any of the following nuances:

[9] https://appen.com

Table 3.2 continued from previous page

Category	Labels	Description
		sarcasm, humour, satire
Stereotype	no - yes	implicit or explicit reference to (mostly untrue) beliefs about a given target.
Intensity	0 - 1 - 2 - 3 - 4	intensity of the hateful discourse, with 0 meaning absence of HS with 0 meaning absence of HS

TABLE 3.2: Annotation Categories.

3.3.1 The Test Dataset

Given the aim to investigate the temporal robustness of BERT in predicting hate speech on Twitter messages related to immigration phenomena in Italy we needed a source of data on this topic with an appropriate temporal structure. We were looking for a dataset that allowed us to capture variations in language and topics over time and then measure the hate speech detection systems performance using standard metrics, such as precision, recall, and F1-score. For this purpose, the data filtered as part of the "Contro l'odio" project, described in Section 3.1, were the perfect solution, both in terms of topic and temporal distribution. We used as test data random monthly samples of roughly 2000 tweets per set, from September 2018 to February 2019. It was also entirely annotated on Appen with the same strategy illustrated before. In Table 3.3, we list the detailed size and class balance of all our datasets. We notice that the percentage of HS tweets decreases with time, while tweets are being annotated by the same set of annotators. A possible explanation of this is that the data from 2019 may be significantly different from the examples given to the annotators as guidelines (which belonged to previous years), and this yields to an inconsistently in the quality of the annotation results and ultimately to this class imbalance. The average length of the tweets is 24 words in the training sets and 38 words in the test sets. However, the training sets were collected using the standard Twitter API truncating messages longer than 140 characters, while the test sets were collected with the updated API returning the full messages. In terms of language variability, the type-token ratios are expectedly low, ranging from 10% to 18% across data sets.

Dataset	Size	% non-HS	% HS
Haspeede Set	4000	67.6	32.4
Figure Eight Train Set 1 (data from 2015)	3809	85.5	14.5
Figure Eight Train Set 2 (data from 2017)	3200	82.7	17.3
test 2018_09	1991	67.5	32.5
test 2018_10	2000	82.9	17.1
test 2018_11	2000	84.2	15.8
test 2018_12	2000	84.1	15.9
test 2019_01	2000	90.2	9.8
test 2019_02	2000	91.4	8.6

TABLE 3.3: Dataset size and class balance.

3.4 Experimental evaluation

We devised a set of experiments that allowed us to track precisely how different models performed when trained and fine-tuned on different combinations of datasets, covering different temporal ranges.

3.4.1 Experimental Design

For what regards the prediction systems, we decided to compare AlBERTo against a traditional SVM, as it is the one in use in the "Contro l'odio" project. We then trained each system in two different scenarios: a *sliding window* model and an *incremental* model. In the first case we trained the system on month t_i and then tested it on the following month t_{i+1}.

In the second case instead we progressively incremented the size of the training set: we tested the models on month t_i but trained them on data from all the previous months, from t_0 to t_{i-1}. The rationale behind this choice was to evaluate how the system performance vary while injecting information on language and discourse about the recent past. To explore the interplay between the size of the training set and the temporal gap with the test data, we performed a second set of experiments with a fixed test set but adding Haspeede+ to each of the two training schemes. The reason for this was to evaluate how the systems performed when trained on a larger but older dataset, injected with information on language and topics far away in the past. For comparison, we also tested both systems after having trained them only on Haspeede+, as a baseline for comparison with the other settings.

To smooth out any possible random effect, we repeated every single experiment for all the possible setups five times, each of which had a different random seed. We then computed the arithmetical average of the standard metrics.

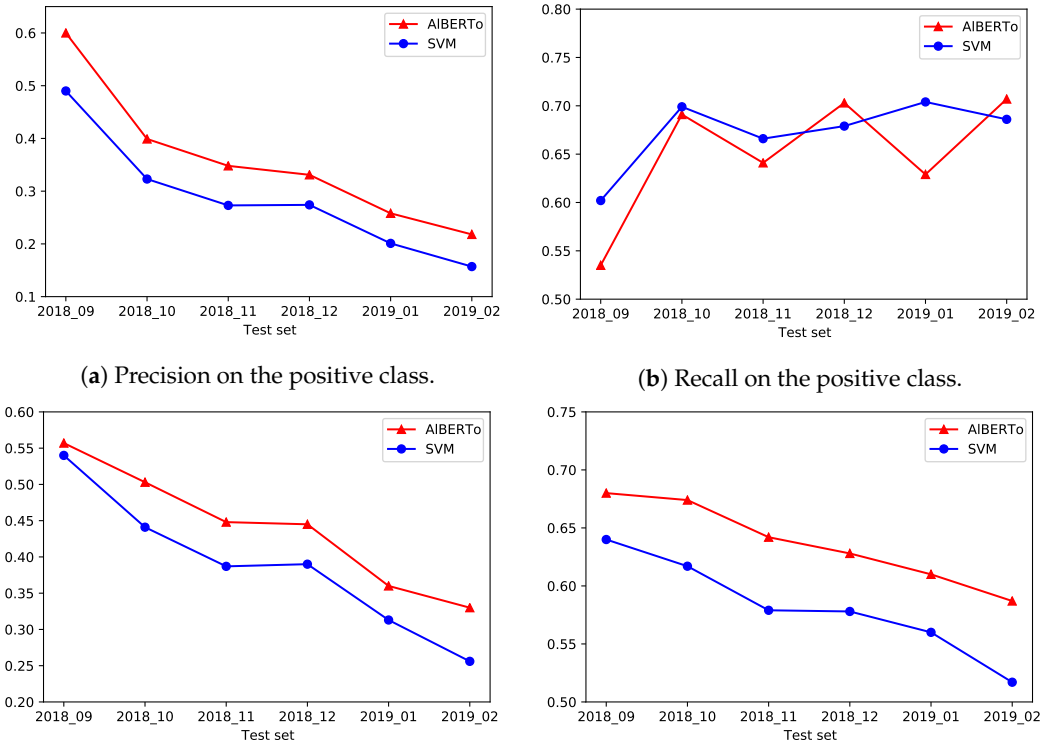
Moving from the consideration that we are performing a hate speech detection task and not a more general classification task, we decided to focus on precision, recall, and F1 for the positive class (the presence of hate speech) and macro F1-score on both classes. The reason behind this choice is that we believe that the key point in our experiment was to measure the effectiveness of the algorithm when correctly detecting hate speech messages, rather than correctly labeling non-hate speech messages. As an example, we believed that for us, the model needed to be able to correctly classify the hate speech sentence "You are ugly, kill yourself" more than classifying the sentence "Today is a good day" as not hate speech. We also computed the macro F1-score that averages on both classes because, as seen in Table 3.3, the distribution

of the label in our training datasets is unbalanced, and this last metrics provides a better insight on the system performance in this specific case.

3.4.2 Results

We trained the model on Haspeede+, a fixed set of data from 2015 and 2017, which are a few years older than our test set. This experiment represents a sort of baseline to evaluate the performance in the other setups. All the metrics from this setting are presented in Table 3.4.

We notice in Fig. 3.3 that, as we expected, both the precision and the F1-score display a generically decreasing trend over time in both cases, and AlBERTo does not outperform the SVM significantly in this case. The two models display in general a similar trend over time for what regards the recall. These considerations are supported by the last chart presented in Fig. 3.3, as the macro F1-score is built as an average of the F1 over the two classes. Specifically, in six months, the model based on AlBERTo has lost 0.227 points of F1 while SVM 0.284. However, this value is influenced by the recall that instead tends to increase with time, as the model increasingly struggles to output accurate results.



(**a**) Precision on the positive class.

(**b**) Recall on the positive class.

(**c**) F1-score on the positive class.

(**d**) Macro-averaged F1-score.

FIGURE 3.3: Evaluation of the model trained on Haspeede+ (fixed training set). (**a**) Precision on the positive class. (**b**) Recall on the positive class. (**c**) F1-score on the positive class. (**d**) Macro-averaged F1-score.

This trend becomes clear if we observe the chart of the precision of the positive class: this metric is crucial as our goal is the minimization of false positives. In such

graph, the two models have an equivalent downward trend for each month, showing that the diversification of the language used in sentences strongly influences the classification performance.

Test Set	SVM				AlBERTo			
	Prec.	**Rec.**	**F1**	**F1 macro**	**Prec.**	**Rec.**	**F1**	**F1 macro**
2018_9	0.490	0.602	0.540	0.640	0.600	0.535	0.557	0.680
2018_10	0.323	0.699	0.441	0.617	0.399	0.691	0.503	0.674
2018_11	0.273	0.666	0.387	0.579	0.348	0.641	0.448	0.642
2018_12	0.274	0.679	0.390	0.578	0.331	0.703	0.445	0.628
2019_01	0.201	0.704	0.313	0.560	0.258	0.629	0.360	0.610
2019_02	0.157	0.686	0.256	0.517	0.218	0.707	0.330	0.587

TABLE 3.4: Numerical results of the evaluation of the model trained on Haspeede+ (fixed training set).

Test Set	SVM				AlBERTo			
	Prec.	**Rec.**	**F1**	**F1 macro**	**Prec.**	**Rec.**	**F1**	**F1 macro**
2018_10	0.350	0.500	0.412	0.629	0.406	0.641	0.497	0.679
2018_11	0.475	0.319	0.375	0.639	0.454	0.513	0.448	0.686
2018_12	0.427	0.230	0.299	0.600	0.491	0.447	0.445	0.694
2019_01	0.331	0.214	0.260	0.598	0.425	0.367	0.360	0.661
2019_02	0.382	0.169	0.234	0.592	0.421	0.342	0.330	0.673

TABLE 3.5: Numerical results of the evaluation of the model trained on Sliding Window (no Haspeede+) dataset.

Test Set	SVM				AlBERTo			
	Prec.	**Rec.**	**F1**	**F1 macro**	**Prec.**	**Rec.**	**F1**	**F1 macro**
2018_10	0.350	0.500	0.412	0.629	0.406	0.641	0.497	0.679
2018_11	0.343	0.435	0.384	0.624	0.415	0.694	0.519	0.694
2018_12	0.389	0.387	0.388	0.636	0.464	0.627	0.533	0.704
2019_01	0.273	0.362	0.311	0.611	0.436	0.434	0.435	0.687
2019_02	0.266	0.448	0.334	0.624	0.356	0.539	0.429	0.679

TABLE 3.6: Numerical results of the evaluation of the model trained on Incremental (no Haspeede+) dataset.

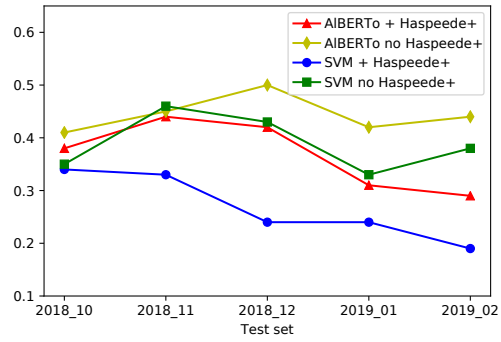
Test Set	SVM				AlBERTo			
	Prec.	Rec.	F1	F1 macro	Prec.	Rec.	F1	F1 macro
2018_10	0.343	0.649	0.449	0.634	0.387	0.764	0.514	0.677
2018_11	0.329	0.571	0.418	0.628	0.445	0.479	0.461	0.684
2018_12	0.347	0.569	0.431	0.640	0.415	0.507	0.456	0.665
2019_01	0.239	0.551	0.334	0.603	0.315	0.525	0.394	0.649
2019_02	0.195	0.494	0.280	0.575	0.283	0.514	0.365	0.636

TABLE 3.7: Numerical results of the evaluation of the model trained on Sliding Window and Haspeede+ dataset.

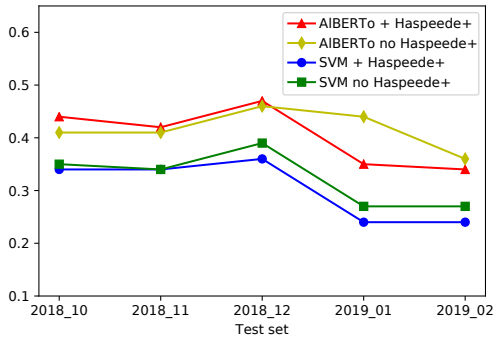
Test Set	SVM				AlBERTo			
	Prec.	Rec.	F1	F1 macro	Prec.	Rec.	F1	F1 macro
2018_10	0.343	0.649	0.449	0.634	0.439	0.672	0.524	0.694
2018_11	0.335	0.587	0.427	0.633	0.415	0.616	0.493	0.684
2018_12	0.360	0.535	0.430	0.645	0.471	0.516	0.470	0.680
2019_01	0.243	0.464	0.319	0.603	0.352	0.505	0.412	0.666
2019_02	0.239	0.529	0.329	0.611	0.339	0.523	0.407	0.667

TABLE 3.8: Numerical results of the evaluation of the model trained on Incremental and Haspeede+ dataset.

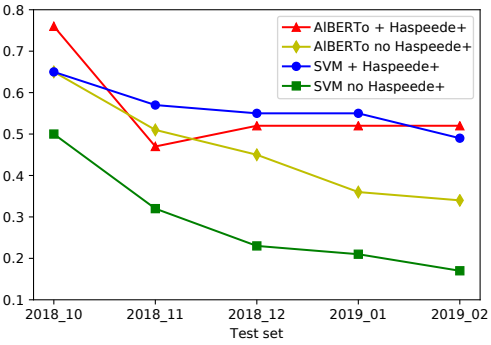
In Fig. 3.4 we present the compared results of the experiments with the two different training set scenarios: the sliding window (on the left side) and the incremental (on the right side). In each graph, we plotted the results with and without the injection of the Haspeede+ set. For the sake of clarity and completeness we present all the metrics for the experiments without the Haspeede+ dataset in Table 3.5 and Table 3.6. The results of the experiments with the Haspeede+ dataset are instead listed in Table 3.7 and Table 3.8.



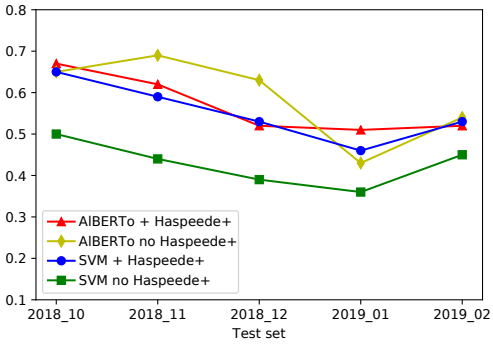
(**a**) Precision on the positive class.



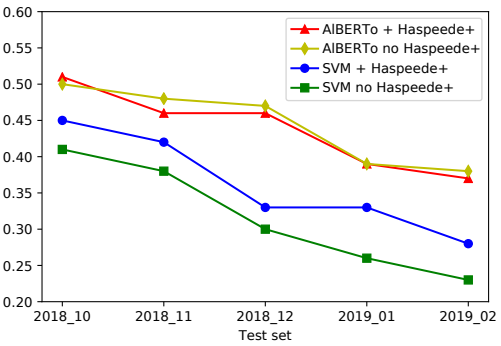
(**b**) Precision on the positive class.



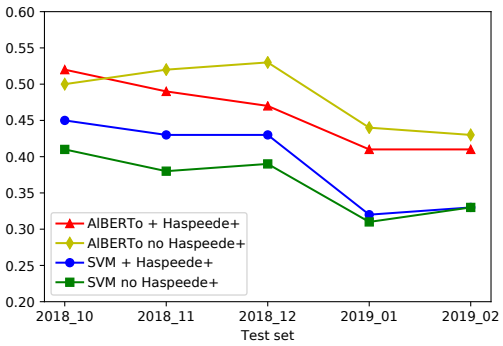
(**c**) Recall on the positive class.



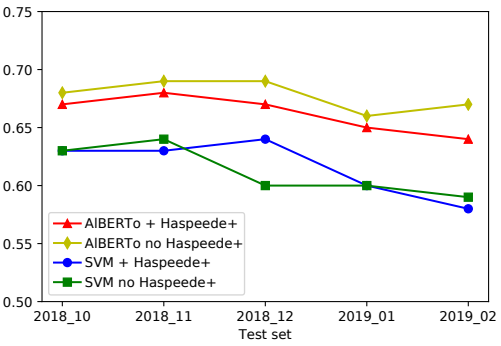
(**d**) Recall on the positive class.



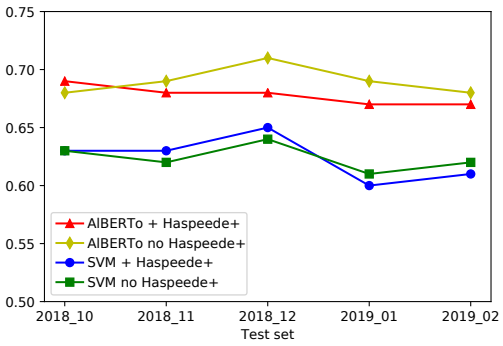
(**e**) F1-score on the positive class.



(**f**) F1-score on the positive class.



(**g**) Macro-averaged F1-score.



(**h**) Macro-averaged F1-score.

FIGURE 3.4: Evaluation of the models trained on a Sliding Windows (left columns) and Incremental dataset (right column). (**a**,**b**) Precision on the positive class. (**c**,**d**) Recall on the positive class. (**e**,**f**) F1-score on the positive class. (**g**,**h**) Macro-averaged F1-score.

Our most meaningful results are presented in Fig. 3.4f: AlBERTo overcomes the performance obtained using a fixed training set (Fig. 3.3c). It can successfully mitigate the decay of the performance with the passage of time as shown in Fig. 3.4g–h. Moreover, AlBERTo trained on an incremental dataset performs better than the same model trained on an incremental scheme built using only on the more recent data, and better than SVM as well.

We can observe that using both a sliding-window and incremental training strategy, the models' performance tends to reduce over time. Nevertheless, the drop in performance, in both the approaches, is smaller compared to the one obtained using a fixed training dataset. This observation demonstrates the importance of the diachronic training. This behavior is especially evident if we look at the F1 of the positive class, apart from small irregularities. As an example, the trend of both the F1-score shows an inversion around December in the incremental setup. An additional factor impacting the performance of all classifiers on the last two test sets is likely the lower relative rate of HS messages (see Table 3.3). However, other reasons concur in the specificity of these monthly samples, in particular lexical and topical features, as explained in the next section.

The strategy based on incremental training set generally works better than the one based on sliding window as a consequence of the largest amount of recent data available for training. The key to a successful fine-tuning of AlBERTo is the use of data that are not too distant in time from the test set: we estimate a max value of six months. As proof of our claim, when adding Haspeede+, model performance tends to decrease. This behavior is a consequence of its internal algorithm that uses fine-tuning to focus the model on many specific and timely aspects. Consequently, older data addition can introduce noise that does not help the model to converge better. The SVM strategy has a similar behavior of AlBERTo when comparing the two strategies of training. The main difference is that SVM is more sensitive to the quantity of data than AlBERTo, and consequently, it performs better if Haspeede+ is included in the training set. As a general claim, we can then affirm that the best strategy to train models for hate speech detection is to use a large amount of data as updated as possible because both these aspects influence machine learning models.

Table 3.9 shows the results of the fixed and incremental windows experiments in comparison. In order to understand the significance of our results, we performed a paired Wilcoxon non-parametric test. This analysis shows statistical confidence for the results of the two different experiments for $p < 0.01$.

To support our hypothesis about the importance of updated data for reducing the negative influence of time factor on machine learning models, we decided to train both models on a new dataset injected with Haspeede+ data, which are temporally very distant from the test set data. We can observe that the performance of both models in this condition tend to decrease over times, which is a proof of our claim: an injection of timely distant data introduce a degree of noise that ultimately leads to a decrease of the model performance, in both cases similarly. All the results of this experiment and their statistical significance (tested as before with a Wilcoxon test) are listed in Table 3.10.

Test Set	SVM no Haspeede+				AlBERTo no Haspeede+			
	Fixed	Incremental	Δ	*p*-Value	Fixed	Incremental	Δ	*p*-Value
2018_10	0.617	0.629	+ 0.012	4.1×10^{-18}	0.674	0.679	+ 0.005	7.1×10^{-7}
2018_11	0.579	0.624	+ 0.045	3.6×10^{-38}	0.642	.694	+ 0.052	3.4×10^{-14}
2018_12	0.578	0.636	+ 0.058	8.8×10^{-63}	0.628	0.704	+ 0.076	1.2×10^{-8}
2019_01	0.560	0.611	+ 0.051	2.3×10^{-56}	0.610	0.687	+ 0.077	8.8×10^{-11}
2019_02	0.517	0.624	+ 0.107	4.8×10^{-62}	0.587	0.679	+ 0.092	3.3×10^{-14}

TABLE 3.9: Comparison of the macro F1 scores between the fixed and incremental windows experiments.

Sliding Window			
Months: Training->Test	**Macro-F score Linear SVM**	**Macro-F score BERT**	***p*-Value**
9->10	0.629	0.679	7.6×10^{-02}
10->11	0.639	0.686	7.5×10^{-10}
11->12	0.600	0.694	2.4×10^{-8}
12->1	0.598	0.661	1.8×10^{-2}
1->2	0.592	0.673	1.1×10^{-2}
Sliding Window + Haspeede+			
Months: Training->Test	**Macro-F score Linear SVM**	**Macro-F score BERT**	***p*-Value**
9->10	0.634	0.677	1.8×10^{-1}
10->11	0.628	0.684	3.7×10^{-16}
11->12	0.640	0.665	2.2×10^{-6}
12->1	0.603	0.649	7.3×10^{-7}
1->2	0.575	0.636	3.1×10^{-7}
Incremental Window + Haspeede+			
Months: Training->Test	**Macro-F score Linear SVM**	**Macro-F score BERT**	***p*-Value**
9->10	0.634	0.694	7.6×10^{-2}
9+10->11	0.633	0.684	8.4×10^{-8}
9+10+11->12	0.645	0.680	2.3×10^{-6}
9+10+11+12->1	0.603	0.666	1.3×10^{-6}
9+10+11+12+1->2	0.611	0.667	3.0×10^{-12}

TABLE 3.10: Wilcoxon Test *p*-values.

Consequently, we can affirm that machine learning techniques are affected in performance by a bias consequent of the change of language over time in new text analyzed, especially in a domain of hate speech. The issue is strongly related to the amount of data provided at the model for the training phase, and consequently, the use of data updated and large enough is the best option for preserving good performance of an automatic hate speech detection model. In the event, it is difficult to obtain frequently enough updated data, a possible strategy to use for mitigating the issue is to use an incremental training set that merges old and new data in order to guarantee the model enough data for generalizing correctly and some new examples that include the updated vocabulary.

3.5 Lexical analysis

To gain a more in-depth insight into the phenomena causing the prediction performance described in the previous section, we performed an additional set of experiments aiming at understanding the topics of discussion emerging from the data,

and their diachronic properties. Our main statistical tool is the *weirdness index* [74], an automatic metric to retrieve words characteristic of a *special language* with respect to their typical usage.

In practice, given a *specialist* text corpus and a *general* text corpus, the weirdness index of a word is the ratio of its relative frequencies in the respective corpora. Calling w_s the frequency of the word w in the specialist language corpus, w_g the frequency of the word w in the general language corpus, and t_s and t_g the total count of words the specialist and general language corpora respectively, the weirdness index of w is computed as:

$$Weirdness(w) = \frac{w_s/t_s}{w_g/t_g}$$

When applied to an annotated corpus of hate speech, we expect that the words with high WI will reflect the most characteristic concepts in that corpus, those who distinguish it most from generic language. By analyzing the words with the highest weirdness index in each test set (treated as specialized corpora) against the training set Haspeede+ (treated as the general corpus), we aim at discovering patterns among the emerging topics that are novel with respect to the original training set. Table 3.11 shows the top ranked words by Weirdness Index from each of our test sets. Please note that words occurring only once in the data set were filtered out before the computation of the index. Indeed, at the top of each ranked list of words by weirdness, words appear that refer to specific events. For instance, the test set from January 2019 is dominated by the topic of the Sea Watch NGO ship and the refusal of the Italian government to let it enter their ports[10]. In almost all cases, the topics emerging from the weirdness analysis are different from one month to the following. In rare occurrences, the echo of an event on social media spans two months, as is the case of the political discussion around the Global Compact for Migration pact[11], observed among the top ranked words in November as well as December 2018.

September 2018	October 2018	November 2018	December 2018	January 2019	February 2019
dalai	cialtronaggine	credito	strasburgo	sea	sea
lama	@giovanniproto67	global	global	47	#salvininonmollare
l'escamotage	all'opposizione	carte	@lavaligiadianna	#salvininonmollare	47
applicare	eurotassa	moavero	giuseppe	siracusa	recessione
slavi	#leu	#baobab	sea	#portichiusi	emirati
#deluca	l'illegalità	▪	:/	#restiamoumani	@danilotoninelli
ℹ	@gbongiorno66	assegni	@europarl_it	battisti	@openarms_it
@time	incompetente	ruspe	open	49	processare
magazine	#unhcr	@lavaligiadianna	versato	#giornatadellamemoria	#portichiusi
costituirsi	aste	flessibilità	@openarms_fund	valdese	@medhope_fcei
abramo	@tgrsicilia	polonia	venuto	olandese	2019
luisa	867	peschereccio	#bergoglio	totalmente	#bergoglio
ranieri	#voisapete	unhcr	antonio	#fakenews	tav
sfavore	avessimo	firmare	babbo	magistris	febbraio
gyatso	paladino	@baobabexp	emendamento	#cesarebattisti	@rescuemed
xiaomi	#iostoconmimmolucano	eletta	international	palermitani	#catania
profetessa	organizzava	#pakistan	praticano	disumane	@openarms_fund
giudea	riacesi	dell'onu	#manovra	tedesche	fazio
busto	donano	meningite	natale	claudio	#martina
asselborn	combinato	hiv	presepe	chiedendo	laureato

TABLE 3.11: Top 20 words by Weirdness Index in each test set.

[10] https://en.wikipedia.org/wiki/Sea-Watch

[11] https://en.wikipedia.org/wiki/Global_Compact_for_Migration

September 2018	October 2018	November 2018	December 2018	January 2019	February 2019
delinquere	parassiti	zingari	feccia	#primagliitaliani	incompatibile
zingari	stupratori	stupri	parassiti	😈	rotto
barconi	stuprare	parassiti	assassini	invasori	fanculo
auto	pamela	stuprano	negri	stupri	stupratori
biglietto	violentata	bambine	civiltà	#rai	vergognatevi
#stopinvasione	ns	💩	moderato	infami	esistono
hotel	strade	uccidono	cacciati	auto	ladri
calci	dell'islam	intanto	stupratori	pamela	etnie
clandestino	feccia	🤢	venire	visti	bus
famiglie	nomadi	cesso	infedeli	autoctoni	siriani
studenti	merde	ladri	💩	paghiamo	pensionati
modello	cani	#pakistan	#primagliitaliani	film	nullafacenti
assistenza	dobbiamo	etc	cancro	recessione	negri
#movimentonesti	farci	buonismo	onesti	spacciatori	l'invasione
ladri	abusivi	tramite	assassino	chiese	forze
feccia	assassini	strade	ospiti	invasione	nonni
subito	campi	moderato	rispetta	merde	maledetti
rapine	dovete	spaccio	#allah	ospite	bestie
pagano	mantenerli	diventata	#corano	tale	90
cinesi	rimpatriare	stupro	#stopinvasione	stuprata	pago

TABLE 3.12: Top 20 words by Polarized Weirdness Index in each test set.

We then apply the weirdness index to the same sets in a different way, to gauge the topics most associated with the hateful language in the labeled dataset. The mechanism is straightforward: instead of comparing the relative frequencies of a word in a special language corpus (the test set, in the previous experiment) against a general language corpus (the training set), we compare the relative frequencies of a word as it occurs in the subset of the labeled datasets identified by one value of the label against its complement. We refer to such variant as *Polarized Weirdness Index* (PWI). Formally, consider a labeled corpus $C = \{(e_1, l_1), (e_2, l_2), ...\}$ where $e_i = \{w_1, w_2, ...\}$ is an instance of text, and l_i is the label associated with the text where e_i occurs, belonging to a fixed set L (e.g., $\{HS, not - HS\}$). The *polarized weirdness* of w with respect to the label $l*$ is the ratio of the relative frequency of w in the subset $\{e_i \in C : l_i = l*\}$ over the relative frequency of w in the subset $\{e_i \in C : l_i \neq l*\}$ The outcome of the calculation of the Polarized Weirdness index is again a ranking over the words contained in the subset of each test set identified by the hateful label. Words occurring only once in each test set were again filtered out before computing the index. High-PWI words from a class will give a strong indication of the most characteristic words to distinguish that class (e.g., hate speech) from its complement (e.g., not hate speech).

Following this analysis, whose results are shown in Table 3.12, we found many action verbs among the top-ranking words in all the test sets. Such verbs refer to negative, in particular criminal, actions such as *killing* or *robbing*, indicating a strong link between the topics emerging in the messages labeled as hateful and events in the news. However, the main verbs are different from month to month. For instance, verbs related to *drug dealing* are prevalent in November 2018, while verbs related to

rape are relevant from October 2018 to January 2019 with a peak in December 2018, and verbs related to killing are mostly concentrated in December 2018.

Finally, our considerations are confirmed by the chart in Figure 3.5, showing a simpler frequency-based analysis provided by Sketch Engine[12]. Here, the vertical axis shows the frequency of the items in each test set relative to the average of all six sets. This score is higher than 100% when an item occurs more often than the average in a month, e.g., "compact" occurs almost four times the average in December 2018). The lemmas related to criminal activity, "rubare" (*to steal*), "stuprare" (*to rape*), and "uccidere" (*to kill*) show different patterns, likely linked to events in the news. The effect is even more prominent with more topical words, such as "porti" (*harbors*, referring to the Sea Watch event) and "compact" (from the aforementioned Global Compact), showing clear peaks in specific months.

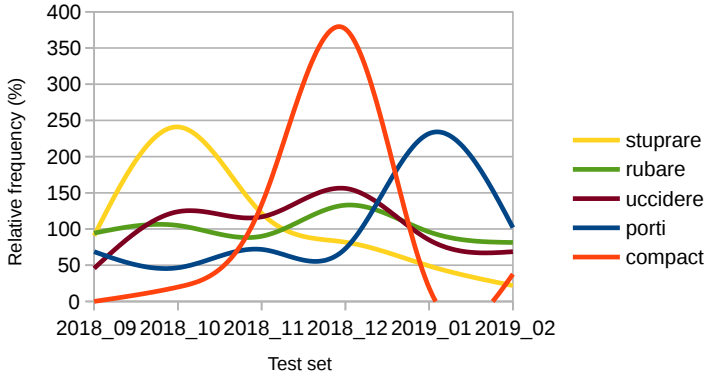


FIGURE 3.5: Relative frequencies of topical words and lemmas over time.

3.6 Conclusions and final remarks

The focus of the experiments described in this chapter was the evaluation of the temporal robustness of different hate speech prediction systems, with respect to language and topic change over time, as stated in our first research question. We designed two different experiments: in the first case, we trained the models on data from a single month and tested it on the following month. In the second case, we injected information on the recent past (thus increasing the size of the training set) by using data from all the months preceding the one from which we draw the test sample[13]. Unsurprisingly, injecting training data temporally closer to the test set sharply improves the prediction performance of AlBERTo compared with the SVM (partly answering our second research question), since the training data are very similar to the test data form a linguistic and topic perspective. On the contrary, our experiments show that increasing the size of the training set does not necessarily lead to equally improved performance. To provide a more complete answer to RQ2, we also repeated the experiments adding a larger training set from a distant time span. Our results show how this setting has a beneficial effect on the SVM, but a negative effect on the performance of the transformer model. To gain a better understanding of the linguistic differences between our monthly samples, we also ran a statistical

[12] https://www.sketchengine.eu/

[13] All codes and data are publicly available to the community here: https://github.com/komal83/timeofyourhatepaper

analysis of the topics from a temporal perspective. The analysis confirms that there is a relatively fast shift in topics in the online discourse, and this constitutes the main challenge to overcome in order to improve the robustness over time of the predicting systems for hate speech detection.

We applied our methodology to a real Italian case study. However, the experimental design is agnostic with respect to the language. Therefore, the approach can be expanded from a multilingual perspective, provided the development of suitable diachronic corpora, which is unfortunately not available as of now. This fact proves the importance of efforts such as the "Contro l'Odio" project in pursuing diachronic studies and the necessity of similar projects in multiple languages.

Our annotated data are also naturally unbalanced, with non-hate speech examples representing most of the dataset. It is commonly known that the performance of machine learning approaches is strongly influenced by the class unbalance [59, 69, 69, 25], and consequently, it could be very interesting for the future to investigate the impact of automatic balancing techniques or the addition of new training data on the robustness observed in the models we analyzed. In the next chapter I will explore the challenge of hate speech from the perspective of the topics discussed in online debate and how measuring the dynamic of online debates around real life events can help enhance the performance of detection and monitoring systems.

Chapter 4

Hateful contents and the public discourse on social media: a measurement of the role of topic shift

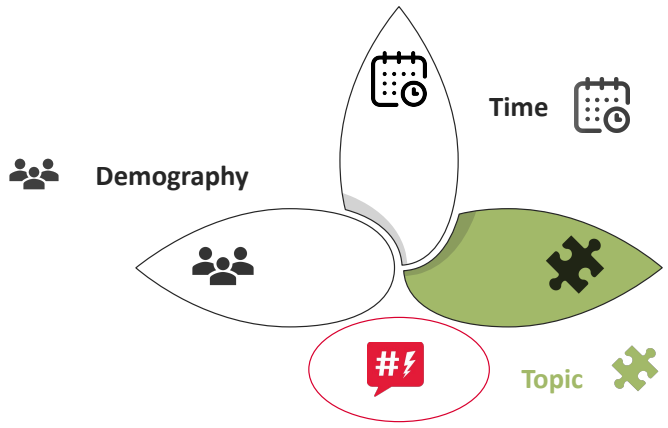


FIGURE 4.1: Thesis diagram.
Credits: Templated designed by PresentationGO.
Icon from Flaticon.com

In the previous sections I mentioned how we assumed that the public discourse on social media was centered around the debate of news from both online and traditional media outlets, thus leading to a rapid shift in the most discussed topics. This intuitive explanation needed a data-driven investigation, to better understand the impact of said topic shift on the performances of hate speech detection and monitoring systems. The optimal circumstances for this analysis occurred in March 2020 when the first European Covid-19 lockdown became a reality in Italy and dominated both the online and offline public discourse.

The goodness of hate speech prediction systems, and of NLP algorithms in general, is rooted in how well they capture and model all the relevant characteristics of

language in the context of a specific phenomenon and how it evolves over time.

In recent years social media have become one of the predominant sources of linguistic data and venue for noticeable phenomena in the domain of NLP tasks. Among all the possible features used to describe language, this section concentrates on the evolution of specific topics in online discussions on Twitter around a specific subject. The aim is to characterize how the online conversation on the Italian Twitter relatively to the first lockdown imposed in a European country in 2020, following the Covid-19 outbreak, shifted very quickly from one major news to the next one.

The lockdown due to the response to COVID19 pandemic in 2020 is a truly unprecedented event in recent history. A sizeable number of governments around the world imposed similar very restrictive measures on their citizens with respect to movements, social gatherings, health prescriptions in public places, outdoor activities and workplaces. The beginning of this pandemic was characterised by a quick succession of news reports on both news cases and institutional advice and rules on how to navigate everyday life as the crisis was unfolding in the entirety of the world. This combination of factors offered the optimal scenario in which gather data to perform analysis on the rapid topic shift in online conversations on Social Media, as this factor is crucial for any classification algorithms when deployed on NLP tasks on data deriving from such sources. Here the focus is more specifically on the identification of most popular debated topics and outburst of negative messages as a proxy to investigate which measures or behaviour generated the highest volume of negative emotional responses in the general population in Italy, the first EU country to impose strict lockdown measures, rapidly followed by many other countries. By combining multiple classification methods we gathered insights into which governmental measures generated the most debated online conversation but we also conclude for the need of deeper investigation on how to build ad hoc corpora and methods to investigate specific linguistic phenomena as online conversation with rapid topic shift following the flow of news coming from both online and traditional media outlets.

4.1 The Dataset

I mentioned in the previous chapter how the TWITA [15] datasets is a very useful resource when it comes at investigating hateful messages on social media. At the beginning of the pandemic in 2020 it seemed interesting to filter out this collection of tweets in order to obtain a Covid-19 related set of tweets in Italian. This new resource, named 40wita [1] [14], was created by means of a filtering with a set of Covid-19 related keywords. The very first set of said keywords were based on the constant monitoring of Twitter top trends and hashtags extracted from a dedicated website [2] . This information were extracted starting from February 2020 and updated three times per day, to ensure a certain degree of variability and limit the possible impact of bias.

This first list of keywords was then manually adjusted and the final version encompasses 42 items in Italian, listed in Table 4.1. For better clarity we provided a translation in Table 4.2.

The filtering on TWITA was run from 1^{st} February 2020 to 30^{th} April 2020 and resulted in the collection of 3309704 tweets. The dataset is publicly available on request, but for privacy reason all the tweets have been dehydrated. Twitter's Terms

[1] https://osf.io/n39ks/

[2] https://getdaytrends.com

Keywords in Italian				
covid	COVID19Italia	abbracciauncinese	CuraItalia	600euro
covid19	redditodicittadinanza	ionosonounvirus	ciricordiamotutto	CineINPS
covid-19	eurobond	ionomifermo	oggisciopero	COVID19Pandemic
corona virus	coronabond	aperisera	chiudiamolefabbriche	ringraziarevoglio
coronavirus	restiamoacasa	covindustria	chiudetetutto	iononrinuncioalletradizioni
quarantena	preghiamoinsieme	italiazonarossa	andràtuttobene	cercareDi
autoisolamento	NoMes	bergamoisrunning	INPSdown	
auto-isolamento	#milanononsiferma	percheQuando	l'italianonsiferma	
iorestoacasa	bergamononsiferma	stateacasa	apritetutto	

TABLE 4.1: 40wita Dataset Keywords.

Keywords translated in English				
covid	COVID19Italia	hughachinese	CuraItalia	600euro
covid19	basicincome	iamnotavirus	werememebereverything	CineINPS
covid-19	eurobond	idonotstop	striketoday	COVID19Pandemic
corona virus	coronabond	aperisera	closethefactories	saythankiwant
coronavirus	stayhome	covindustry	closeeverything	idontgiveupontraditions
quarantine	praytogether	italyredzone	allwithbeallright	tryto
selfisolation	NoMes	bergamoisrunning	INPSdown	
self-isolation	#milanodoesnotstop	whywhen	italydoesntstop	
istayathome	bergamodoesnotstop	stayinghome	openeverything	

TABLE 4.2: 40wita Dataset Keywords translated into English.

of Service do not allow the full JSON for datasets of tweets to be distributed to third parties. As it is only allowed to share tweets IDs, it is fundamental to use tools (know as "hydrators") to retrieve the full JSON format for each tweet, which of course includes the text, the main source of information for NLP purposes.

4.2 Hate Speech prediction with AlBERTo

In Section 3 I analyzed how we applied AlBERTo to predict hate speech against immigrants in Italy on data extracted from TWITA [15]. Given that that network was already fine tuned on Italian social media language, I decided to use the same exact pre-trained network to predict hate speech in 40wita, in order to evaluate the goodness of the predictions.

The daily percentage of tweets labeled as hate speech in February, March and April 2020 is shown in the following Figure 4.2.

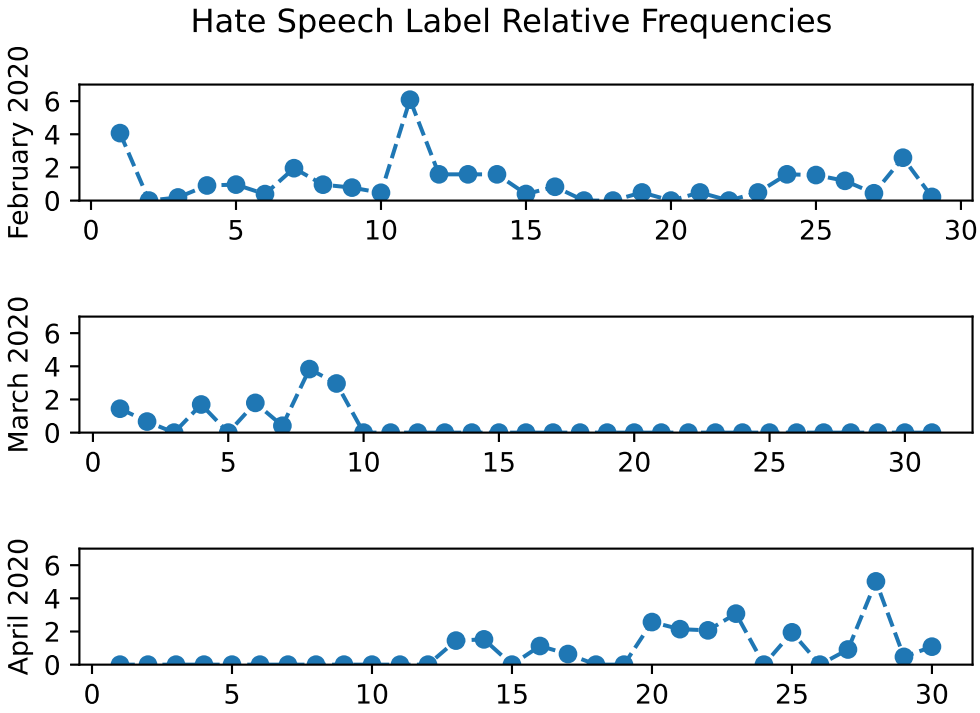


FIGURE 4.2: Daily percentage of tweets labeled as hate speech in February 2020 with AlBERTo

The full list of figures is provided in the table in Appendix A.

The results were far from satisfying, as the number of tweets predicted as hate speech were very very low, almost to the point were any kind of statistical analysis could not be trusted to bring meaningful contributions.

The time slices with almost negligible counting of HS message coincide with the weeks where the hard lock down was in place in Italy. On the other hand, the count rises in April where the first restrictions were about to be lifted soon, and this could have reflected into the type of language and topics most frequently discussed online. This finding is nevertheless aligned with the results in [52], despite the 40wita dataset having significant differences with respect to the training data of AlBERTo. Given all these considerations, we attempted a lexicon-based classification, which will be described in the next section.

4.3 Lexicon-based Abusive Speech prediction with HurtLex

In the previous section I showed that our pre-trained hate speech predicting algorithm AlBERTo is not effective at producing accurate enough results when it comes to 40wita. The studies on HS cited in Section 2 found much higher prevalence of hateful messages on Twitter-based datasets that, even though in a different context, it may induce us to conclude that an unknown but not negligible percentage of hateful messages were left undetected.

Hence, in order to get a broader insight of the potentially hateful messages in this dataset we resorted to perform the same task by means of a computational lexicon

of hate words. For this purposed we used HurtLex[3] [17], a multilingual computational lexicon of offensive, aggressive, and hateful words, derived from "Le Parole per Ferire", a lexicon of words to hate compiled by Italian Emeritus Professor of Linguistics Tullio de Mauro. This lexicon contains 17 different categories of hate words and for each of them a list of characterising words. The full list of categories and relative meaning and acronyms is listed in the following Table 4.3.

Label	Description
PS	negative stereotypes ethnic slurs
RCI	locations and demonyms
PA	professions and occupations
DDF	physical disabilities and diversity
DDP	cognitive disabilities and diversity
DMC	moral and behavioral defects
IS	words related to social and economic disadvantage
OR	plants
AN	animals
ASM	male genitalia
ASF	female genitalia
PR:	words related to prostitution
OM:	words related to homosexuality
QAS	with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
SVP	words related to the seven deadly sins of the Christian tradition

TABLE 4.3: HurtLex Lexicon Categories.

Each of the 17 HurtLex category consists in a list of lemmas and respective definitions[4]. We automatically counted the occurrence for each lemmas of each category in all the tweets of the dataset and used this metric to assign a predominant label for each tweet. The resulting distribution of the frequencies for each category over the whole 40wita dataset is shown in Figure 4.3.

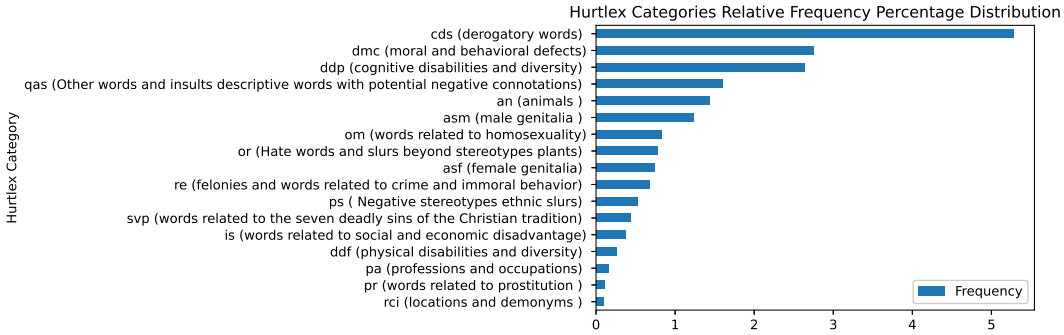


FIGURE 4.3: HurtLex Categories Frequency Distribution over the all 40wita Dataset

We can see from Fig. 4.3 that the predominant category of hate speech is represented by tweets containing derogatory words, which is a pretty general definition.

[3] http://hatespeech.di.unito.it/resources.html

[4] The full process of creating of this linguistic resource is described in [17] and is available at https://github.com/valeriobasile/hurtlex

To gain a deeper insight on how this classification has unfolded we analysed which were the most common words that classified a tweet into a specific category. For each of the top 3 categories I performed a raw count of the occurrences of the related lemmas and plotted only the ones with at least 10 occurrences. The results are presented in the following charts.

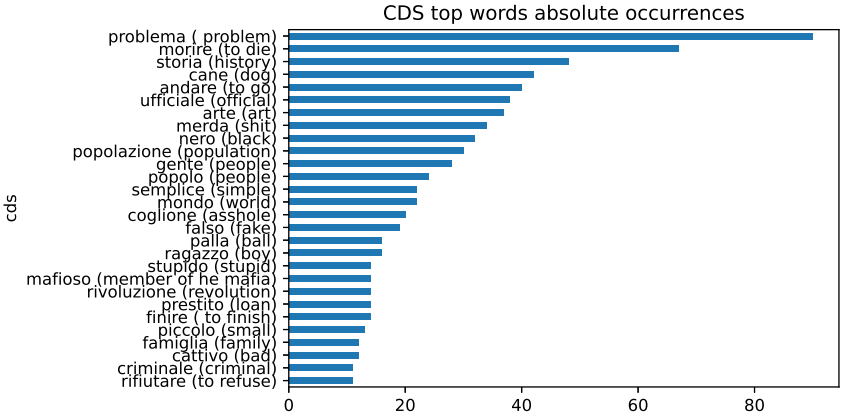


FIGURE 4.4: Words with at least 10 occurrences in tweets labeled as CDS (derogatory words)

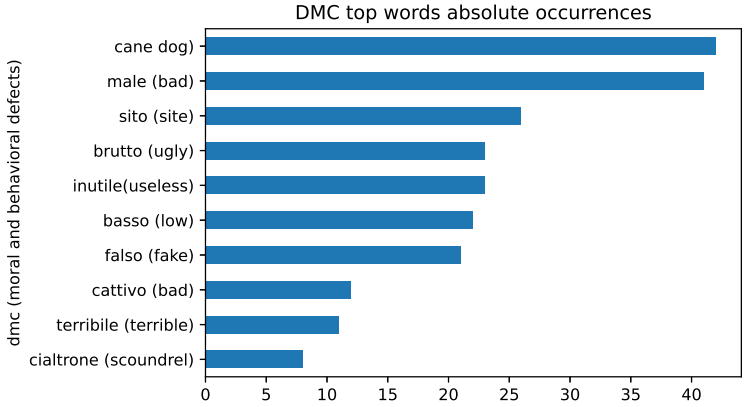


FIGURE 4.5: Words with at least 10 occurrences in tweets labeled as DMC (moral and behavioral defects)

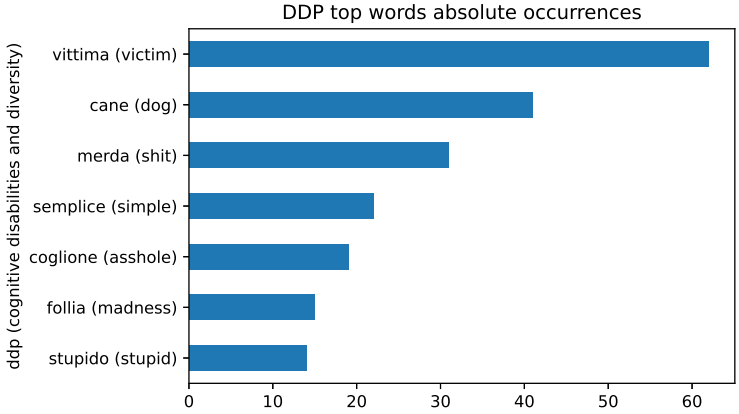


FIGURE 4.6: Words with at least 10 occurrences in tweets labeled as DDP (cognitive disabilities and diversity)

This insight is meaningful in showing why HurtLex presents some struggles in the accuracy of the classification of the tweets. The division into pre-defined categories turned out to be not as informative as we were hoping at the beginning. The words that determine whether a tweet falls into a category or another are very generic (e.g.,"problema"="problem", "storia"="history") or can assume very diverse meaning depending on the context (e.g., "cane"="dog"), and this contributes in creating noise in the classification of the tweets. The table presented in Appendix B contains a selection of tweets from the most common categories, to provide some clear example of the analysis I presented above.

An improvement on this would encompass a manual revision of the list of words for each category, in order to leave out the most generic one and leave only those who can guarantee a more accurate result. I also conducted a manual revision of the tweets classified in each of the HurtLex categories to gain more specific insight on the classification outcomes. I revised all the tweets belonging the categories with less than 30 tweets, while for the others I choose a random sample of 30, for consistency with the previous case. One of the most surprising findings was that in the category "rci - locations and demonyms", contrary to what one could expect given the global dimension of the pandemic, our data counter-intuitively showed that the debate was centered strictly around the measures taken in Italy and the differences between national and local rules. As an example of this, in the following I list some tweets in this category.

> Buongiorno #amici, visto che questo "governo", sta abbandonando i nostri produttori, agricoltori, contadini, pescatori, panettieri, commercianti, insomma, tutti! Aiutiamoli noi, partiamo dalla #colazione[...]!!!#IoMangioItaliano #italia #MadeInItaly #COVID19 [5]
> → *Goodmorning #friends, as this #government, is abandoning our producers, farmers, fishers, bakers, traders, in short everyone! Let's help them, starting from #breakfast!!! #IEatItalianFood #italia #MadeInItaly #COVID19*

[5]Note that tweets often contain orthographic mistakes and other errors such as missing punctuation.

> É solo cafona? O anche ignorante? Lo chiedo per un amico #covid_19italia #coronavirus Jole Santelli[6]: "Mi dispiace per Conte[7], ma io apro i bar e chiudo i confini" → *Is she just badly behaved? Or ignorant as well? Asking for a friend #covid_19italia #coronavirus Jole Santelli:"I am sorry for Conte, but I will open pubs and close the border"*

This lexicon-based approach, even though it did not lead to the desired outcome, it was still important to gain more information on our corpus and to gain experience for future similar work. In the next section we will focus on the most powerful classification tool that we employed on this dataset: the unsupervised topic modeling, carried through a Latent Dirichlet Allocation algorithm (from now on, LDA).

4.4 Latent Dirichlet Allocation Topic Modeling

The aim of this section is to describe a topic model classification on the 40wita dataset by means of a Gensim [105] implementation of the Latent Dirichlet Allocation Model (LDA, in the following). After observing that in the considered time frame the flow of news and new regulations issued by the government was relatively quick, I decided to divide the data into time slices consisting of 7 days each. This time granularity seems the one that offered the best opportunity to capture the quick reaction online to new rules and relevant news from media. After having run LDA on the data, I analyzed for each of the 13 time-slices, which were the most relevant words associated with the dominant topics for each time.

The list of the most interesting words associated to the dominant topic for each time-slice is presented in Table 4.4, both in the original version in Italian and in English translation. I made a conscious decision to avoid listing too common words in characterising the weekly topics, as it is unlikely to provide any specific information to that specific time-frame, but rather just adding to the background noise. I instead focused on highlighting words associated to very specific and punctual events that made it in the public debate just for a brief period of time (in the order of a couple of days). This is an attempt to show how, despite being a first approach to a topic modeling classification, the LDA is able to correctly identify very precisely spikes in the relevance of some specific topics of discussion from background noise on Twitter.

Time Slice	**Start Day**	**End Day**	**Relevant Words**
Time 0	02-01	02-08	toscano/capodanno/cinese/test/niccolo/quarantena/cinesi *tuscan/newyearseve/chinese/test/niccolo/quarantine*
Time 1	02-09	02-15	capodanno/toscano/cinese/asia/paura/spallanzani/jinping *newyearseve/tuscan/chinese/asia/fear/spallanzani/jinping*
Time 2	02-16	02-22	cinesi/influenza/veneto/sciacallo/focolai/ torino/lazio/chiudere/allarmismo *chinese/flu/veneto/shark/outbreaks/ torino/lazio/close/alarmism*
Time 3	02-23	02-29	zona rosso/intensivo/ceppo/turista/tifoso *red zone/intensive/strain/tourist/supporter*
Time 4	03-01	03-07	venezia/ cogogno/piemonte

[6]Governor of Regione Calabria, who announced public measures directly conflicting with national pandemic rules

[7]Prime Minister of Italy during 2020

Table 4.4 continued from previous page

Time Slice	Start Day	End Day	Relevant Words
			venezia/ cogogno/piemonte
Time 5	03-08	03-14	cina/cuba/bertolaso
			china/cuba/bertolaso
Time 6	03-15	03-21	supermercato/lecce/attività alimentari negozio
			supermarket/lecce/grocery shop
Time 7	03-22	03-28	mattarella/raffaele
			mattarella/raffaele
Time 8	03-29	04-04	pregare salvini
			to pray salvini
Time 9	04-05	04-11	campania/gualtieri
			campania/gualtieri
Time 10	04-12	04-18	calcio/campania
			football/campania
Time 11	04-19	04-25	oxford/vaccino
			oxford/vaccine
Time 12	04-26	04-30	brescia/ripartire/sierologico
			brescia/to recover economically/serology test

TABLE 4.4: Most relevant words associated to the dominant topic for each time-slice, in Italian and English.

This result is certainly interesting as it captured some specific topics of discussion and their shift over time. In week 1 for example, there are terms related to the origin of this virus in China and the reflection on the Chinese community in Italy, mainly located in Tuscany and that we put under the spotlight due to the celebrations of Chinese New Year in Early February 2020. Another noticeable example from the same week is "niccolo" which is the name of the first know patient suspected of Covid that made it to the news because he was flown back to Italy from China with a relevant media echo around his personal history and health conditions. This model was able to correctly identify the conversations around the first restrictions on movements following the first covid outbreak in Lombardy and Veneto (the two regions that were first and most hardly hit by the emergency) and then when the national lockdown was put in place, the gradual shift of the conversation towards the difficulties of normal life in such a new context, captured for example by "supermercato" and "attivitá alimentari", semantically related to every day shopping. Other punctual events that were captured by the model were the arrival of doctors from Cuba to face the emergency in the area surrounding Milan and the debate about Easter celebrations and restrictions in places of catholic workships ("pregare", "salvini"): a battle carried forward by the most right- wing parties. As powerful as this model is, it show a fundamental limit for our perspective and purpose. The relevant topics were punctual but not consistent over time, hence not comparable to study their evolution over time. To overcome this issue we implemented a Dynamic Topic Modeling, as presented in the next section.

4.5 Dynamic Topic Modeling

In the previous section I described the topic modeling obtained with the LDA but we faced the problem of lack of comparability of said topics over time. An interesting option to solve this problem is offered by the so-called Dynamic Topic Modeling [21]. The power of this model relies in the fact that it divides the datasets into custom time slices and extracts the same exact topics over all of them, allowing for the possibility of studying how topics evolve over time. The first step consisted in the exploration of the hyper-parameters space to evaluate which combination lead to a model that better predicted the topics contained in our corpus. Even though LDA is an unsupervised model and we have no gold standard to use as s benchmark for the goodness of the predictions, we can still rely on two some specific metrics to evaluate and compare the performances of models with different parameters setting. The first intrinsic evaluation metrics is the perplexity score: it captures the behaviour of the model towards data which were previously unknown by means of a normalised log-likelihood of a held-out test set. This metric, that can be described as a predictive likelihood basically describes how well the model represents or reproduce the statistics of the held-out data. However there are relevant studies (for example [30]) proving that perplexity and human judgement not only often do not correlate, but sometimes they even anti-correlate. For this reason a second metric was elaborated: the coherence score, to better model human judgement. This measure captures the degree of semantic similarity between the words related to each single topic (i.e. a measure of the likeness of their meaning). In our specific case there is no annotated corpus that can serve as a training set, hence we only explored the trend of the coherence score with reference to changes in the number of topics, chunksize of data, number of passes and evaluation score. After having evaluated the interdependence of all these variables on a sample of our data, I made the decision to move forward with this experiment with an LDA model with 5 topics and 20 words for each topics.

The extracted topics are listed in the following Table 4.5.

Topic No.	**Italian**	**English**
Topic 0	quarantena	quarantine
Topic 1	altro	other
Topic 2	lavoro	work
Topic 3	governo	government
Topic 4	sanità	healthcare

TABLE 4.5: Topics Extracted using the Dynamic Topic Modeling.

The DTM outputs each unlabelled topic as a list of words with a relevance value. This value, between 0 and 1, represents the probability of that single word to be affiliated with a specific topic. Based on the value it is possible, for each topic, to rank the most relevant words and see how they evolve over time.

The following Figures 4.7,4.8, 4.9 show the change in ranking for all the 20 words involved and is presented as a coloured heatmap, where the blue values represents words with higher ranking while the red ones the lower end of ranking.

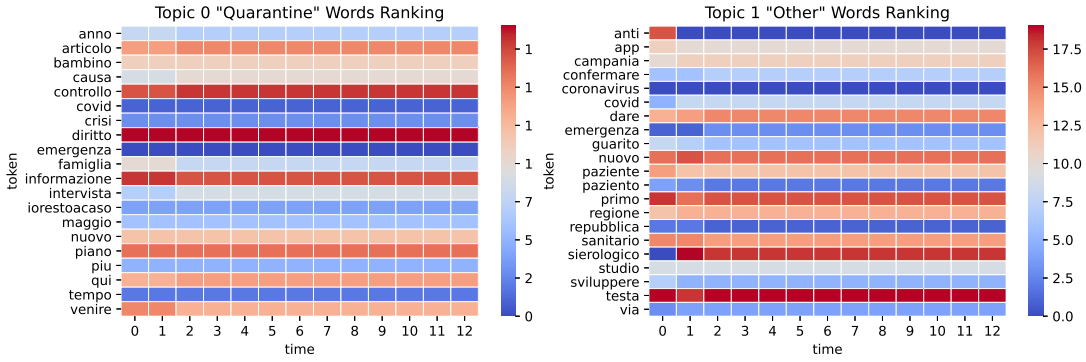


FIGURE 4.7: Time evolution of words relevance ranking for Topic 0 and Topic 1.

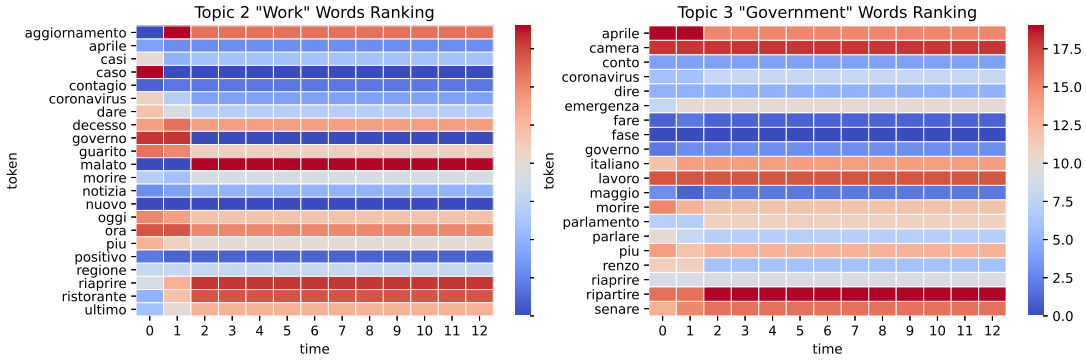


FIGURE 4.8: Time evolution of words relevance ranking for Topic 2 and Topic 3.

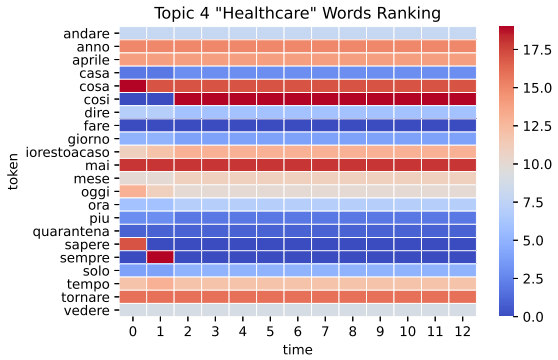


FIGURE 4.9: Time evolution of words relevance ranking for Topic 4.

There are two main insights we can gain from this visualization. The first one is that topics are lists of pretty common words, which proves how hard of a task topic detection is, because of the complexity and versatility of human language, where general words that can be used in different contexts with more or less similar meanings. The second insight we gained is that the biggest changes in the word ranking happen within the firsts time slices. A possible explanation may be tracked back to how this dataset was created. As described at the beginning of this section, the list of hashtags and trends used to filter the tweets was compiled in February and stayed fixed in time. This means that potentially more relevant tweets on the Covid-19

pandemic were left out because more relevant hashtags evolved over time but this changes did not have a mirror in the keywords used for selecting relevant tweets.

I mentioned that the most powerful feature of a DTM model is the extraction of exactly the same topics over all the different time slices. It is then very straightforward to analyse how the share of documents labeled as containing predominantly one of the topics evolve over time, giving hints on how predominant a certain topic is over the course of time. For each of the 13 time slices I computed the ratio of documents labeled as predominantly referring to each of the 4 topics and hence constructed the following charts in Figures 4.10, 4.11 and 4.12, that help visualizing the separate trends over time.

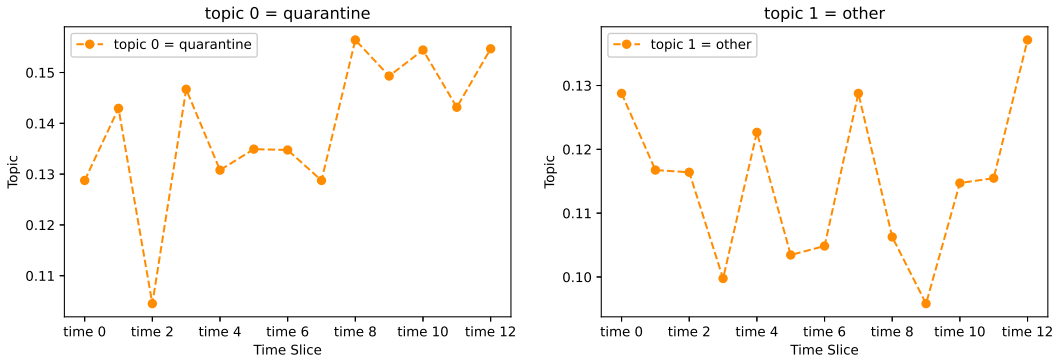


FIGURE 4.10: Time evolution of share of documents containing the Topic 0 and 1.

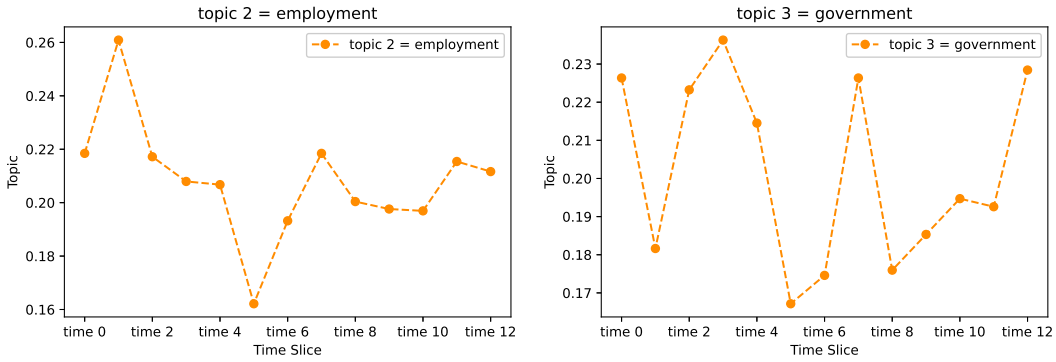


FIGURE 4.11: Time evolution of share of documents containing the Topic 2 and 3.

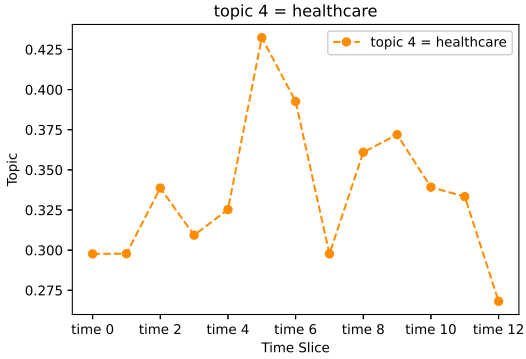


FIGURE 4.12: Time evolution of share of documents containing the Topic 4.

For an easier interpretation, we plotted in Figure 4.13 the normalized share of documents classified as containing each of the 4 topics in each time slices, to highlights the relative trends over time.

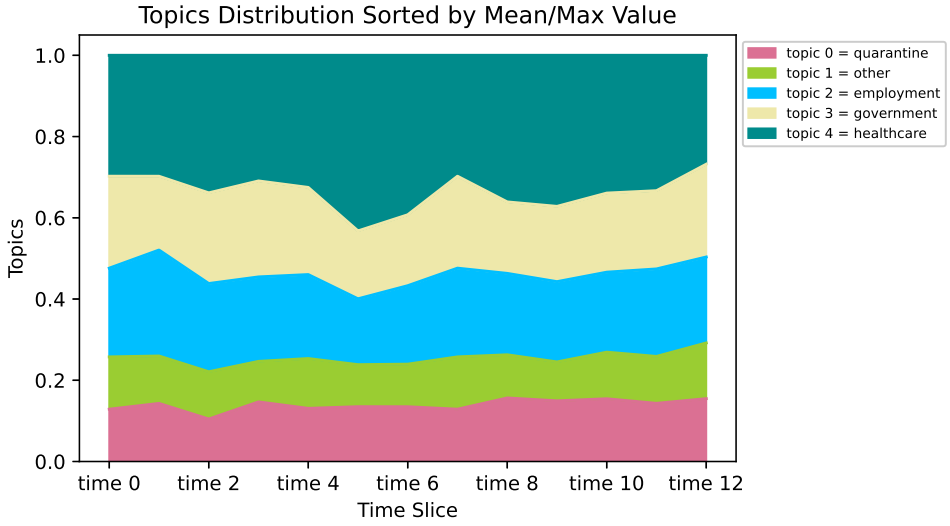


FIGURE 4.13: Evolution over time of mean and maximum values of the share of documents related to each of the 4 topics.

It is aligned with our intuition that the largest share of documents across time refer to topic no. 4 "healthcare". But more in details it is interesting to analyse the relation between the timestamp of the spikes and relevant Covid-19 events in Italy.

Topic	Time- Slice	Start Day	End Day	Relevant event
4-healthcare	2	16/2/20	22/2/20	public discussion around the first red zones in Veneto
	5	8/3/20	14/3/20	Announcement of the arrival of a medical task force from Cuba in Lombardy (14/3/20). Appointment of a special consultant for the emergency in Lombardy (16/3/20).
2-work	1	9/2/20	15/2/20	public discourse around the Chinese community in Italy
	5	8/3/20	14/3/20	Announcement of the arrival of a medical task force from Cuba in Lombardy (14/3/20). Appointment of a special consultant for the emergency in Lombardy (16/3/20).
	11	19/4/20	25/4/20	First positive news about the Oxford vaccine AstraZeneca
3-government	2	16/2/20	22/2/20	public discussion around the first red zones in Veneto
0-quarantine	3	23/2/20	29/2/20	first red zones issued in Lombardy and Veneto
	9	5/4/20	11/4/20	Economical measure announced. Public discourse around lifting the strict lockdown measures.

TABLE 4.6: Relevant Covid-19 events occurred around spikes in the chart.

The previous Table shows that the spikes in shares of documents related to the most predominant topic "quarantine" do follow temporally major events about public health announcement and measures. This proves the point of this research, which is that the discourse on Twitter not only it follows closely the most recent and relevant news but it quickly shifts from one topic to the other. In fact, all major peaks in Fig. 4.13 are followed by a descendant trend, indicating an immediate loss of predominance and hence an alternation of the dominant arguments of debates.

I decided to explore in a similar way also the temporal evolution of the share of tweets labelled with the HurtLex categorise described in Section 4.3.

For each of the time slices I computed the relative frequency of tweets labeled with every categories and then created a stacked plot of both the maximum values (shows in Figure 4.14) and the normalized mean values (shown in Figure 4.15) of their frequencies, to identify both peaks and categories that were consistently predominant through the time.

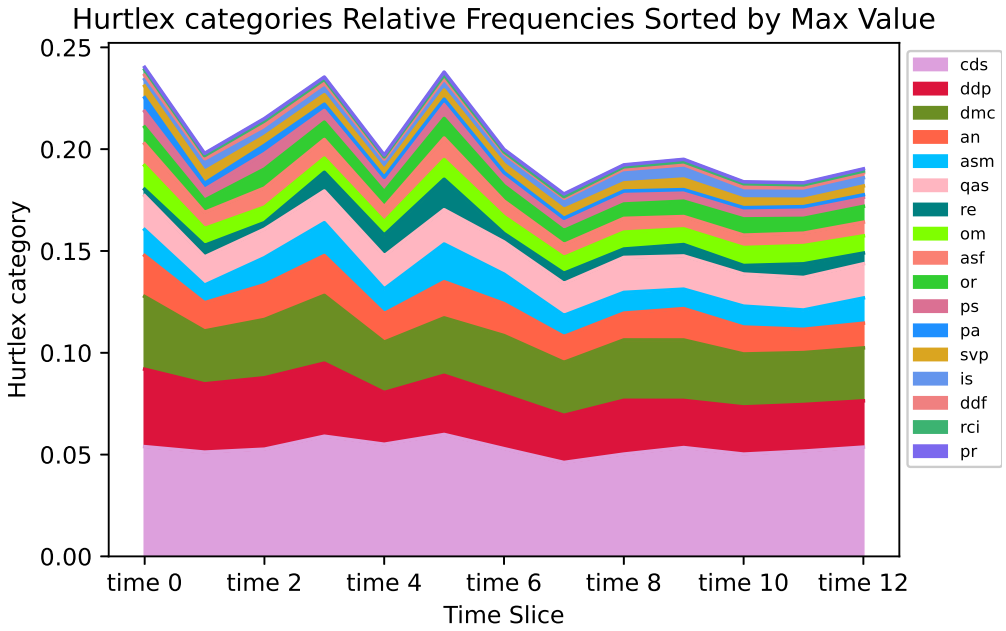


FIGURE 4.14: HurtLex categories maximum frequencies values over time.

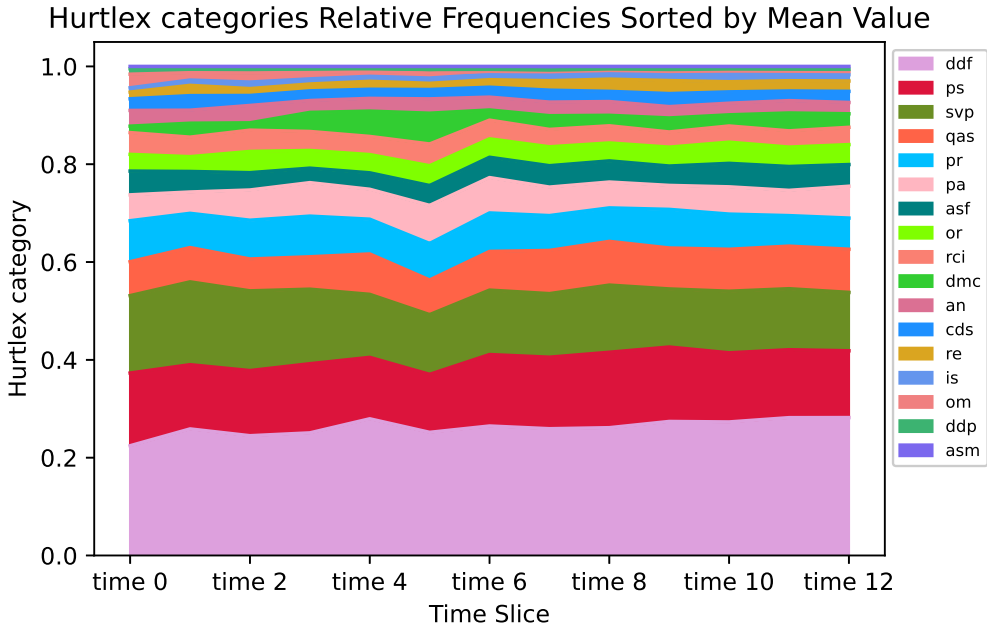


FIGURE 4.15: HurtLex categories mean frequencies values over time.

The relevance of the category "Cds - derogatory words" detected over the whole dataset as a whole in Figure 4.3 confirms its validity also at a weekly level of time granularity, of we look at the distribution over time of maximum relative frequencies values in 4.14. Looking at the chart as a whole it is important to notice that, in accordance with what we have already noticed before, the peaks occur in time slices 3 and 5, which respectively correspond the the issue of the first red zones in Italy and two major public health news regarding Lombardy, the hardest hit region of Italy in the first phases of the pandemic (see Table 4.6 for details).

It is relevant to notice that the peaks in both previous charts occur exactly in the same time slices as the peaks in Fig. 4.13 for the topics "quarantine" and "healthcare", showing that the most heated debates happened around public measures that affected directly and immediately on both the collectivity ("healthcare") and personal life ("quarantine"). Analysis the mean value of the frequencies, in Figure 4.15, we can see that categories rank differently from 4.14. More specifically we see that for example "ddf - physical disabilities and diversity" is by far the most consistent over time but it represents somehow a generic type of offensive language, not specifically correlated with the pandemic, and to some extend this could be considered as a noisy classification of tweets and it would be interesting to investigate further how to improve on this result. In the next section I will discuss how I attempted to use this information about the topics distribution to improve the classification of hate speech messages using AlBERTo.

4.6 Guided Latent Dirichlet Allocation

In the previous section I described how we extracted topics from our corpus in an automatic and unbiased way. In this section I will describe the attempt to use this information to enhance hate speech prediction on the 40wita dataset using AlBERTo. The first step consisted in creating a gold standard by manually annotating for hate speech a subset of 600 tweets from 40wita, 200 for each of the three months. The

annotation guidelines I followed for this task are exactly the same ones as described in Section 3.

Subsequently I used a guided LDA as a classification method, to label the tweets from Haspeede+ with chosen topics. Similarly to the choice made for DTM, I chose to leave one undetermined topic to collect all the tweet that the classifier was not able to correctly classify in one of the other, so the resulting topics, are, as previously: quarantine, work, government and healthcare. The guided LDA is a semi supervised algorithm that takes as input seeds in the form of words, that are believed to be relevant to the underlying topics in the text and hence serve as a guide to the model to converge to them. I chose to include seeds only for 4 instead of 5 to mirror the choice made with DTM where one very generic topic was simply labelled as "other", as it did not contain a set of words with strict semantic similarity. The list of keywords that were used as seeds for the guided LDA of course are derived from the relevants words that DTM extracted for each topic and were selected according to the following rules:

1. presence in all of the 13 time slices
2. consistent relevance values over time
3. exclusion of too generic words or not very informative (e.g.: "piú = more") or that assumes different meaning in different contexts
4. no specific hashtags (e.g.:"iorestoacaso = stayhome". It was widespread during the pandemic but virtually not used before, hence very likely not to be found at all in pre-2020 tweets)

The results set of keywords is listed in Table 4.7

Topic 0 - Quarantine	**Topic 2 - Work**	**Topic 3 - Government**	**Topic 4 - Healthcare**
quarantena (quarantine)	emergenza (emergency)	regione (region)	sanitá (healthcare)
casa (home)	fase (phase)	riaprire (re-open)	positivo (positive)
andare (to go)	lavoro (work)	governo (government)	contagio (contagion)
morire (to die)	crisi (crisis)	sindaco (mayor)	guarito (recovered)
	economico (economic)	sfida (challenge)	casi (cases)
	sanitario (healthcare)	ordinanza (order)	coronavirus
	decreto (law)	aperto (open)	decesso (decease)
	governo (government)	bare (coffins)	tampone (antigenic test)
		ristorante (restaurant)	paziente (patient)
			repubblica (republic)

TABLE 4.7: List of seed for the guided LDA Topic Modeling.

The choice of this particular version of LDA was due to the fact that, as the topics were by definition the same extracted before with the DTM, they are directly comparable with the ones used to label the 40wita dataset sample. The topics distribution in the Haspeede+ dataset followed the distribution presented in the following Table 4.8.

Quarantine	Work	Government	Healthcare
8.96	33.13	0.06	57.78

TABLE 4.8: Topic Distribution in the Haspeede+ Dataset, as computed with an guided LDA classification algorithm, based on the topics extracted from the 40wita dataset using a DTM model.

The topics distribution in the 40wita sample dataset followed the distribution presented in the following Table 4.9.

	Quarantine	Other	Work	Government	Healthcare
February 2020	0.36	0.03	0.09	0.19	0.35
March 2020	0.18	0.11	0.16	0.22	0.34
April 2020	0.14	0.08	0.10	0.28	0.42

TABLE 4.9: Topics Probability Distribution in the 40wita sample.

The next step consisted in optimizing AlBERTo by finding the optimal balance between the learning rate and the validation loss through a systematic exploration of the hyper-parameters space.

After having found the optimal set of parameters, and before running the prediction, given that our training set presents a significant class unbalance between presence and absence of hate speech, we also accounted for that by tweaking the appropriate weights while fitting the model to the data.

I present all the metrics of this experiment in Table 4.10: as anticipated the size of the test set is small, so it is not possible to draw statistically significant results from this analysis.

	precision	**recall**	**f1-score**	**support**
negative class	0.97	0.95	0.96	581.00
positive class	0.10	0.16	0.12	19.00
accuracy	0.93	0.93	0.93	0.93
macro avg	0.53	0.55	0.54	600.00
weighted avg	0.94	0.93	0.94	600.00

TABLE 4.10: Metrics for hate speech prediction with AlBERTo infused with information on topics from a guided LDA Topic Modeling.

The resulting rates of hate speech tweets per months are as follow in Table 4.11

February 2020	March 2020	April 2020
5.5%	1.5%	7%

TABLE 4.11: Hate speech tweets automatically labelled by AlBERTo in the 40wita sample dataset.

In the training set the predominant topics are healthcare and work (which makes sense because they are the ones which are less connected to the pandemic, as the data were collected way before the Covid-19 outbreak in 2020). The highest ratio of tweets labeled as hate speech are found in April, which is the month with the highest ration of tweets in the 40wita dataset labeled as relevant to the two aforementioned topics

The main insight from the exploratory approach is that combining deep learning model with information extracted from topic modeling sound certainly a promising way to enhance the accuracy of hate speech prediction, but a further investigation on size and characteristics of datasets is absolutely essential to gain better results.

4.7 Conclusions and final remarks

In this section I tried to tackle the challenge of measuring and quantifying the topic shift in the public discourse on Social Media, using as a case study the online debate on Twitter following the Covid-19 related lockdown in Italy in 2020, by means of a dedicated filtering of the TWITA [15] dataset. At first I tried to predict which messages contained hate speech using AlBERTo, with the same fine tuning as in [52] but the results were far from satisfying. We then tried a lexicon based approach using HurtLex [17]. We found that the dominant categories of negative messages were derogatory words, insults regarding moral or behavioural defects and cognitive disabilities or diversity. Nevertheless the accuracy of this classification was not very high, and analysing the words in the lexicon that determined the classification for the top 3 categories we realized that there are a lot of generic terms that contribute to a noisy classification. We concluded that a manual revision of the list of words per each category could reduce the noise in the tweets classification and hence improve the outcome of this task. I then moved to the most powerful classification tool that was used on these data: topic modeling. To start with, I run a Latent Dirichlet Allocation algorithm (LDA) and the dataset as a whole and analized the most relevant topic with a weekly time granularity. This method proved valid in extracting the conversation around specific relevant events that happened in Italy in the time from between February 2020 and April 2020. Unfortunately all these topic were not consistent over the whole time-frame hence I moved to a new model, the Dynamic Topic Modeling [21] that allows to classify the whole corpus, divided into suitable time slices consistently over time. After some optimisation of the model I settled for 5 topics which were labeled as "quarantine", "other", "work", "government" and "healthcare". Among them, "healthcare" is consistently the predominant in all of the 13 weekly time slices in our corpus and the peaks in the share of documents related to this topic happened around major announcement of public health measures. The stacked chart presented in the previous section shows that after ever major peak there is a descending trend which is a proof of our initial intuition of the fast shift in topics in the public debate. I also analysed in the same way how to share of documents labeled for each of HurtLex categories evolved over times and compares the two outcomes. It turned out that the most evident peaks in the HurtLex categories distribution happen exactly in the same time slices were the topics "quarantine" and "healthcare" have spikes as well, showing that the most heated debates happened around public measures that affected directly and immediately on both the collectivity ("healthcare") and personal life ("quarantine"). I then tried to use all the information gained so far to enhance the hate speech prediction performed by means of AlBERTo. Unfortunately this experiment did not lead to significant results due to the very small size of the resulting test dataset. Infusing deep learning model with information extracted from topic modeling sounds certainly a promising way to enhance the accuracy of hate speech prediction, but this path can be expanded in different ways. To start with, our dataset was collected with a fixed set of hashtags and keywords, while a more flexible and time-evolving approach could lead to a more insightful data collection. Secondly, our corpus was labelled using guidelines derived from another hate speech detection task, while ad-hoc rules and more annotators would certainly improve the quality of the gold standard.

When I described the lexicon approach, I mentioned that a potentially very interesting way to reduce the noise in the classification encompasses an in depth manual revision of the words in each category aimed at removing the most generic terms. In conclusion, our method helped us to prove that topics of discussion on social media

not only follow closely the flow of real life relevant events but also change rapidly over time. Further work on this issue is crucial in improving the performances of algorithms for NLP task on social media linguistic data, as such models need to be time robust to capture and learn as precisely as possible all the possible nuances on how language evolves over time. The finding described in this section, at the time or revising this thesis, are described in [50], which was accepted for an oral presentation at CliV-IT 2021, the Eighth Italian Conference on Computational Linguistics

In the next section will expand our perspective on the phenomena of hate speech online by analysing beyond the linguistic aspects and exploring how demographic data of the context in which those messages arise influence their characteristics and geographic distribution.

Chapter 5

Demography

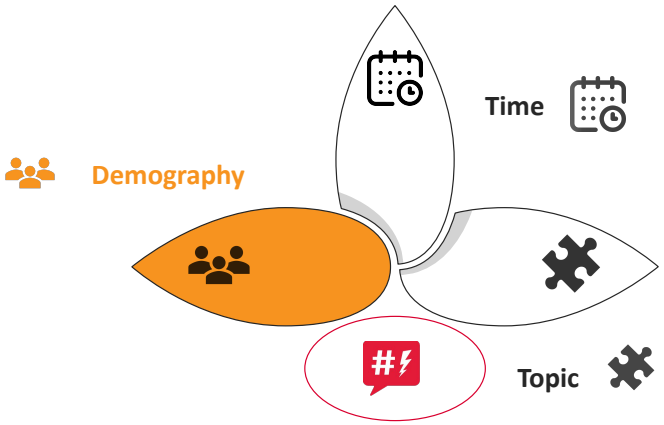


FIGURE 5.1: Thesis diagram.
Credits: Templated designed by PresentationGO.
Icon from Flaticon.com.

The fast growing penetration rate of the Internet[1] brought a radical change to the way people communicate, consume information and debate topics perceived as important, as we described in the previous section. The presence and integration of immigrants, in Europe and globally, is a widely debated issue in the political discourse, both offline and online.

In several circumstances, online discourse can mirror or anticipate real life events or situations that may lead to potentially dangerous episodes both for individuals and communities [90]. Therefore, the analysis of online contents plays an important role in the detection and prevention of critical events, by providing insights on the reality of immigrants integration in local communities.

In particular, as the danger of social media as a breeding ground for online hate speech against immigrants increases, the interest in developing artificial intelligence

[1] http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_tc_broad&lang=en

tools and resources to detect and analyze hate speech and prejudice against immigrants grows. These techniques have the twofold aim of understanding and monitoring the phenomenon and supporting the stakeholders, who daily address the problems related to immigrant integration, in designing data-driven strategies.

Following the definition of [95], *online hate speech* is characterized by a call or otherwise incitement to violent action. That is, messages could be *aggressive* or *offensive* while still not being considered hate speech. An example of a tweet conveying hate speech from the Italian corpus described later on is the following:

> Bombardare tutti i paesi islamici uccidento più bambini possibile perché cresceranno e ci taglieranno la gola per salvarci è inevitabile[2]
> → *Bomb all Islamic countries killing as many children as possible, because they will grow up and slit our throats to save us it is inevitable*

While by definition the messages considered as speech should contain some kind of call to action, the way they are phrased can be more or less explicit. In particular, the identification of the hateful communicative goal may require inference and access to world knowledge, as in the following examples:

> @laura boldrini per eliminarvi tutti voi e pulire litalia da rom e musulmani ci vuole il duce
> → *@laura boldrini to eliminate you all and clean italy from roma and muslim we need the duce*

> #virusrai2 al soluzione immigrazione è fare come fanno in Spagna, un etto di piombo per clandestino
> → *#virusrai2 the solution to the immigration problem is how they do in Spain, 100gr of lead for each illegal immigrant*

In these examples, to correctly identify the hate speech, a system needs to correctly link "Duce" to the Fascist regime and its political implications (first example) or infer the veiled implication *lead→shooting* (second example).

Detecting hate speech online may support the implementation of countermeasures to foster inclusion and fairness in our societies. Inequalities are indeed an increasingly spreading phenomena, which frequently imbue pervasive social media communications, and can have a non-negligible impact with respect to the exclusion of youth (cyberbullying), women (misogyny), and immigrants (xenophobia). Given the huge amount of user-generated content from microblogging platforms like Twitter, the interest is growing towards employing natural language processing and computational social science techniques to address the problem of detecting and monitoring the hate speech diffusion.

In this section I present an original approach to investigate immigration-related phenomena based on the integration of two sources of knowledge: one resulting from the application of automatic hate speech detection techniques to the analysis of spontaneous comments on immigrants from Twitter; a complementary one originating by a selection of relevant official survey-based statistical demographic data periodically released by national institutes, including a set of interesting traditional offline indicators on population.

[2]Note that tweets often contain orthographic mistakes and other errors such as missing punctuation.

The focus is on an European country - Italy - as a case study. Italy was recently affected by a natural population decline, that was completely offset by a net migration accounting for 108% of the total population change [23]. We argue that it is especially crucial to test our methods to study immigration and related phenomena in countries exhibiting such demographic characteristics. In particular, we aim at studying the emerging correlations between the indicators related to employment, education, and crime, and the presence of hate speech in the local online discourse. We present our findings along these lines, which suggest an interplay between economical and cultural factors and the expression of hate online, and somehow mirror the North-South socioeconomic divide in Italy.

5.1 Method

We propose an approach to the exploratory analysis of the socioeconomic landscape that leverages the official data provided by national sources as well as social media big data. We bridge these two sources of data by automatically labelling geo-tagged messages from social media (Twitter in particular) using a supervised NLP classifier. The social media dataset has already been described previously in this thesis as the Haspeede+ in 3.3, hence will not be described in here. In the following I present an overview of the steps of the method, their rationale and the results, which were published in 2019 in [51].

5.1.1 Automatic hate speech classification

We automatically label the whole dataset described in the previous section, classifying each message according to the presence of hate speech. We employ a Support Vector Machine (SVM) classifier with a one-hot unigram representation as feature vector. The model is similar, and inspired by, the one developed for the real-time automatic annotation of hate speech Twitter messages against immigrants in the context of the "Contro l'odio" (Against Hate) project [3], where a Web platform has been developed to support hate speech monitoring in Italy [26].

We implemented the SVM with the *Scikit-learn* linear model with a learning rate set to optimal and 27,642 features. We performed a 5 fold cross validation on the manually labelled dataset reporting a 87.1% overall accuracy, with .77 precision and .52 recall on the positive class (presence of hate speech). The final result of the automatic labelling is shown in Table 5.1.

[3] https://controlodio.it/. Online since December 2018.

Year	# tweets	# tweets HS	% tweets HS
2012	1,024	12	1.00%
2013	2,817	83	2.95%
2014	6,653	354	5.32%
2015	2,986	463	15.51%
2016	1,576	167	10.60%
2017	6,880	570	8.28%

TABLE 5.1: Rate of tweets labeled as Hate Speech, per year.

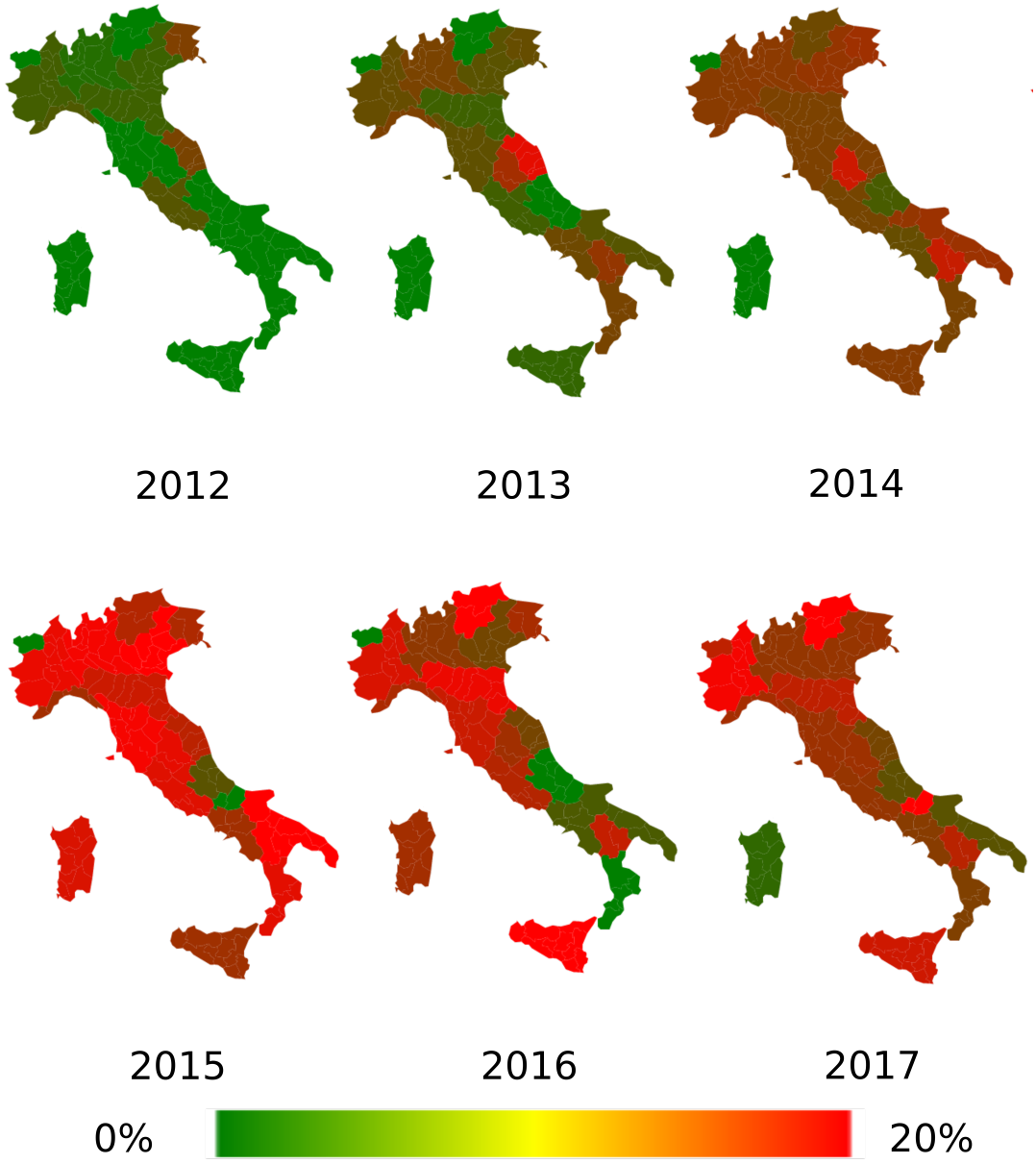


FIGURE 5.2: Percentage of messages automatically identified as containing hate speech, per year, in every Italian region. This maps represent the evolution over time of the rate of tweets automatically labeled as HS. Th red regions have a higher rate of such messages w.r.t. the overall number of geotagged tweets registered in that specific year.

There are noticeable variations of rate of detected hate speech across the years. This result is partially explained by the harvesting method of TWITA that has changed over time. Moreover, our classification model was trained on data drawn from the 2017 subset, therefore it may contain a bias towards certain topics rather than others. However, the online discourse (both in terms of topic and words used) about immigration has evolved over the years in social media, also following the refugee crisis of 2015 in Europe, and our results seem to mirror the unfolding of those particular historical events, such as for instance the civil war in Syria.

Analyzing the geographical distribution of the tweets in our dataset, we found them rather sparse from a geographical point of view and not uniformly distributed across cities. For this reason, we decided to aggregate our results at a regional level. This operation has a twofold benefit: it allows us to perform cross region analysis on the entire Italian territory but also it allows for a smooth integration with the demographic data described in Section 5.1.2. The temporal and geographical distribution of tweets labelled as hate speech is shown in Figure 5.2: it may appear in contrast with 5.1, but this is due to the aggregated data visualization that does not capture fine-grained regional variability over the years.

5.1.2 Demographic data

The counterpart of the data extracted from social media and labelled for hate speech is given by the rich dataset provided by the Italian National Statistical Institute (ISTAT) on socioeconomic indicators in Italy. We focus on three macro-indicators: employment, education and crime, each of which comprises several datasets on the ISTAT website[4].

Employment rate The first dataset we consider contains the figures about the employment rate of both Italian citizens and foreigners legally resident in Italy. The data span from 2012 to 2017 and are aggregated by gender, education level, and age. This set is built by weekly surveys of family samples, and officially released every three months. From the geographical point of view, this source does not provide a detail for each region, but the Italian territory is divided into three macro-areas: North, Center and South Italy (including also the two main islands: Sardegna and Sicilia).

Education degrees The second dataset contains the rate of people that hold one of the four degrees that characterize the educational system in Italy. The Italian scholar system is structured in a primary school for children aged 6 and lasting 5 years, 3 years of middle school ("scuola media"), followed by 5 years of high school education and a University degree with variable duration, according to the subject of study. Education is mandatory up to 16 years old.

We extract from the ISTAT database the number of people holding a specific degree, aggregated by gender and for ages equal or greater than 15 years old. The datasets span from 2004 to 2018 (we consider only those from 2012–2017) and is divided by macro-areas. This report provides individual statistics for Italians and for foreigners regularly resident in Italy, which is the focus of interest of the present work.

[4] http://dati.istat.it/ for the data on Italian citizens, and http://stra-dati.istat.it/ for data specific to foreigners resident in Italy.

Crime ISTAT publishes a dedicated dataset containing the number of people convicted for any given crime category, their citizenship and the number of victims, aggregated by macro geographic areas. For our study, we select data from 2012 to 2016 (the latter being the most recent data available). Although the set comprises a large number of crime types, not all of them is necessarily significant in the context of our study. We compute, for each crime category, the ratio between the rate of convicted Italians against the rate of convicted foreigners, for each year from 2012 to 2016. We then rank the crime categories according to this ratio and select the three categories with a higher imbalance towards foreigner convicts. The reason for this choice is that we are interested in crimes for which the number of foreigners convicted is significantly larger than the number of Italians consistently through all the years, to test the correlation between criminal activity typically associated with immigrants and cyberhate directed towards them. This procedure leads us to consider the following three types of crimes: counterfeiting, theft (particularly petty theft), and exploitation and aiding of prostitution.

5.2 Results

In the following paragraphs we present the results. The geographical maps use a common colour code: green indicates a correlation value closer to -1, meaning that the two indicators tend to correlate negatively (or *anticorrelate*), while in red regions the correlation value is closer to 1, indicating a positive correlation.

5.2.1 Employment rate

The first indicator we consider is the employment rate among foreigners registered as living in Italy and Italian citizens.

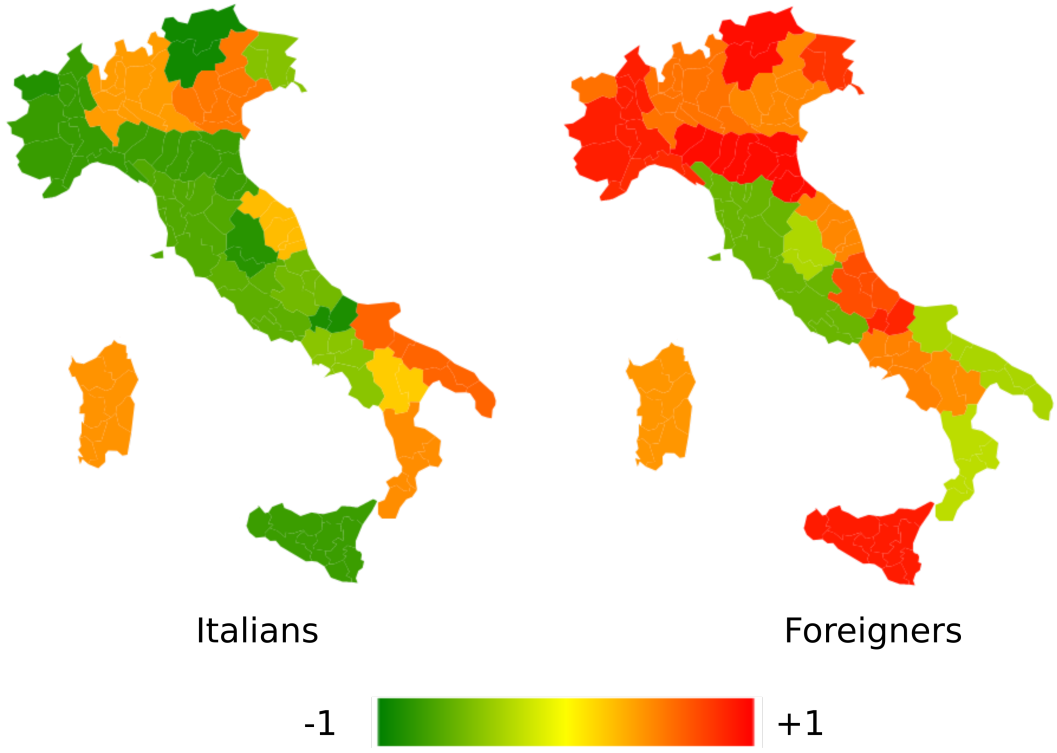


FIGURE 5.3: Pearson correlation index between the rate of tweets labelled as HS and the rate of employment among Italians (left) and foreigners (right) residents in the years 2012-2017.

We compute the Pearson correlation index between the rate of hate speech in tweets and the rate of employment for both foreigners and Italian citizens per region, across the data of all the years 2012-2017. For each region, we compute the correlation of two six-element series, namely the detected hate speech rate in that region over the six years, and the employment rate over the same span of time. The results are shown in Figure 5.3. We computed the correlation for the employment of men and women separately, finding virtually no difference in gender distribution. Therefore, for brevity, we report only the data that refer to the male population. The maps show a clear difference between the Italians dataset and the foreigner datasets, with roughly half of the regions having correlations of different sign, indicating a higher correlation between online hate speech and more foreigners in the workforce. Furthermore, the regions showing higher correlation values are clustered in the north (the richer macro-region) or are highly populated areas (such as the southern regions of Campania and Sicilia). This may indicate a correlation between hatred against immigrants expressed online and more competitive local job markets.

5.2.2 Education

We tested whether there is a relation between the rate of HS and the average level of education across Italy. The hypothesis is that the amount of cyberhate generated in a certain geographic area is related to the overall level of education of the population. We extract from the ISTAT database the number of people by scholarly degree, per each region, for all the years from 2012 to 2017, and for both the Italian and foreigners datasets.

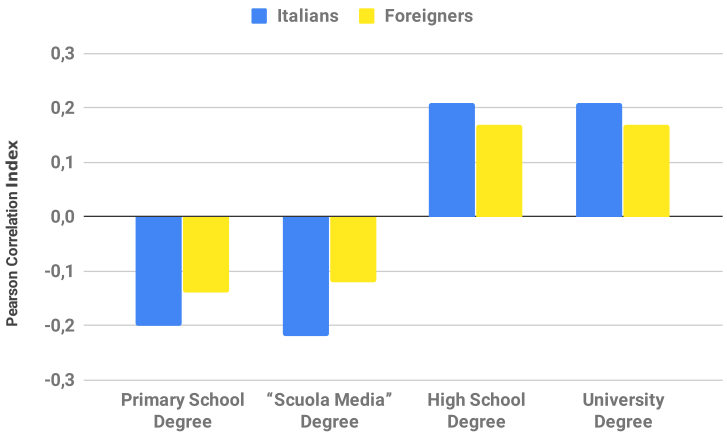


FIGURE 5.4: Pearson correlation index for rate of HS and rate of people with a specific degree per region across all years and regions.

We compute the Pearson correlation index between the rate of HS and the rate of people with a specific degree per region, both for all regions in each year and across all years in one region in order to investigate the presence of both temporal and geographical patterns. We found no particular pattern at this level of granularity. However, aggregating the data both temporally and geographically, the results, presented in Figure 5.4, show a clear pattern of correlation between the level of education and the presence of hate speech in the online discourse. In particular, the rate of online hate speech correlate positively with the level of education of the population, that is, the higher the number of people with high-level degrees (and the less people with lower-level degrees), the higher the rate of detected hate speech in social media. Interestingly, the correlation with the level of detected hate speech (with both positive and negative sign) is stronger with the level of education of the native population than with the level of education of foreigners. One possible interpretation for this result is in the same vein as the speculation on the employment rate. Higher rate of high-level education degrees tend to create a more competitive local job market, where foreigners with higher education can be seen as competitors.

5.2.3 Crime

Crime data is distributed by ISTAT at the macro-region level. We focus on the three crimes categories described in 5.1.2 (counterfeiting, theft, and prostitution), in the period from 2012 to 2016. For each Italian region, we compute the correlation between the detected hate speech rate and the number of foreigners convicted for each of the three crime types in the relevant macro-region, across all the years.

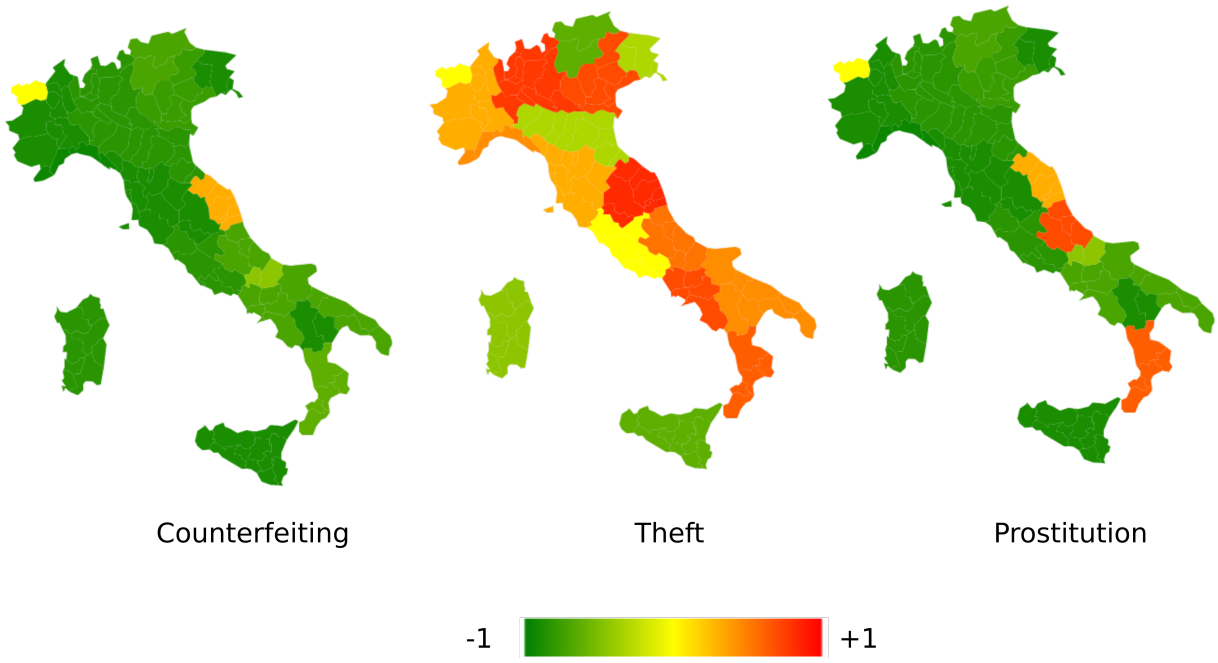


FIGURE 5.5: Pearson correlation index between HS rate and the number of convicted foreigners for counterfeiting (left), theft (center), and prostitution-related crimes (right).

Unsurprisingly, there is a positive correlation with theft-related crimes, suggesting that the perception of an unsafe environment contributes to the generation of cyberhate. On the other hand, counterfeiting and prostitution seem to correlate negatively, or at least not significantly with hate speech. A possible interpretation could rely on the fact that these crimes might be perceived as less impacting on personal safety.

5.3 Conclusions and final remarks

From the perspective of a semi-automated exploratory analysis of cyberhate and its relation to socio-economic factors, our first results show interesting patterns, encouraging us to further pursue the development of the proposed methodology, implement the more recent techniques of deep learning and compare the performances of different methods.

However, the technological challenges involved in a large-scale study generate a high-level of uncertainty, which we aim to address by retrieving a larger set of annotated tweets.

Interestingly, our results indicate a higher correlation between online hate speech and more foreigners in the workforce, and also a higher correlation between online hate speech and the level of education of migrants. Such signals may help to shed some light on the phenomena well-described in the most recent social studies on sociology of migration in Europe, where “the unresolved tension between the economicism of the European approach to labour migration and the philosophy of rights and solidarity" is highlighted [126]. The social expectations about the role of migrant workers is often based on a concept of complementarity between autochthonous and foreign work. When we focus on Italy, survey-based social studies report that people express uncertainty toward a positive impact of immigration on the society, but the prevalent opinion is that immigrants are doing jobs that Italians

no longer want to do, so contributing to improve Italian economy. This scenario is of course perceived as advantageous for Italians, since migrant workers and locals are not competing for the same jobs. On the other hand, where higher education and better integration in the labour market of the migrant workers occur, this is possibly perceived as a threat to employment of Italians, that finds its voice in social media as hate speech. This could provide a possible interpretation of the signals we detected, that deserves further investigation in future work, based on the analysis of finer-grained statistical data on work type and employment status.

As reported by social studies the other factor that seems to play a role in the generation of cyberhate is personal safety, or at least its perception. This finds a reflection in the stark contrast between the positive correlation of online hate speech with the level of petty theft (perceived as a personal danger) and the lack of correlation with equally serious crimes (such as counterfeiting or prostitution), that are perhaps less perceived as a threat to people's safety.

As a future work it would be interesting to investigate the sociology literature in order to test hypotheses on the origins of hate crime, validating them by retrieving additional specific demographic data. Furthermore, a more in depth exploration of the interplay between cyberhate and other indicators and events, such as the outcome of national and local political elections could lead to interesting finding about any specificity about Italy (see for example [108, 18, 104].

The finding of this research (which were published in [51] would certainly lead to even more interesting insights if coupled with more detailed meta information about internet usage and the demographic of Twitter users in Italy. More specifically the evolution of internet penetration rate (especially post Covid-19 pandemic lockdown in 2020) could serve as a proxy for the accessibility of online services to the general population in Italy (as in [20]). This information can also provide a useful context to the data location sparsity that is crucial in geotagged data (see for example [32]). In Section 3 we explored how the accuracy of hate speech predictions is a very complicated issue, hence combining this results to other statistics and then draw conclusions is a very delicate process. In light of the considerations that were made, an increase level of fine tuning of algorithms training could certainly impact positively in the accuracy and insightfulness of findings. Last but certainly not least, the ultimate goal and hope of these analysis is to try and monitor and hopefully prevent both online and offline hate crimes. Therefore a fine grained temporal and geographical mapping of the evolution of real life reported crimes could shed more light on the issue).

Chapter 6

Conclusions

The availability of large annotated corpora from social media and the development of powerful classification approaches have contributed in an unprecedented way to tackle the challenge of monitoring users' opinions and sentiments in online social platforms across time. This thesis aims to explore in particular the phenomenon of hate speech messages on social media with a multiperspective approach summarized by the research questions introduced in Chapter 1.

1. **RQ 1:** How can we evaluate the temporal robustness of hate speech detection and monitoring systems for social media?

2. **RQ 2:** How can we investigate the rapid temporal shift of most debated topics on social media and leverage this information to gain more insight and eventually boost the temporal robustness of different hate speech prediction systems?

3. **RQ 3:** Hate Speech detection on social media is an online phenomenon that is rooted in offline real life, where socio-economic factors characterising geographical territories and people are key. How can we leverage information from traditional socio-demographics indexes about population of a country or region, and information on hate speech dynamics automatically extracted from social media to improve our understanding of interplay between economical and cultural factors and the expression of hate online?

RQ1 The attempt to answer this question is contained in Chapter 3, where I described the experiments that were designed for the evaluation of the temporal robustness of different hate speech prediction systems, with respect to language and topic change over time. We designed two different experiments: in the first case, we trained the models on data from a single month and tested it on the following month. In the second case, we injected information on the recent past (thus increasing the size of the training set) by using data from all the months preceding the one from which we draw the test sample. Unsurprisingly, injecting training data temporally closer to the test set sharply improves the prediction performance of AlBERTo compared with the SVM (partly answering our second research question), since the training data are very similar to the test data form a linguistic and topic perspective. On the contrary, our experiments show that increasing the size of the training set does not necessarily lead to equally improved performance. To provide a more complete analysis, we also repeated the experiments adding a larger training set from a distant time span. Our results show how this setting has a beneficial effect on the SVM, but a negative effect on the performance of the transformer model.

We applied our methodology to a real Italian case study. However, the experimental design is agnostic with respect to the language. Therefore, the approach can

be expanded from a multilingual perspective, provided the development of suitable diachronic corpora, which is unfortunately not available as of now.

RQ2 The second research question tackles the issue of topic shift in online debates and the interplay with real life events, trying to measure and quantify the speed of such shift, as a crucial factor for hate speech monitoring systems robustness to language changes over the course of short period of time.

We used as a case study the online debate on Twitter following the Covid-19 related lockdown in Italy in 2020, by means of a dedicated filtering of the TWITA [15] dataset. At first I tried to predict which messages contained hate speech using AlBERTo, with the same fine tuning as in [52] but the results were far from satisfying. We then tried a lexicon based approach using HurtLex [17]. We found that the dominant categories of negative messages were derogatory words, insults regarding moral or behavioural defects and cognitive disabilities or diversity. Nevertheless the accuracy of this classification was not very high, and analysing the words in the lexicon that determined the classification for the top 3 categories we realized that there are a lot of generic terms that contribute to a noise classification. We concluded that a manual revision of the list of words per each category could reduce the noise in the tweets classification and hence improve the outcome of this task. I then moved to the most powerful classification tool that was used on these data: topic modeling. To start with I run a Latent Dirichlet Allocation algorithm (LDA) and the dataset as a whole and analized the most relevant topic with a weekly time granularity. This method proved valid in extracting the conversation around specific relevant events that happened in Italy in the time from between February 2020 and April 2020. Unfortunately all these topic were not consistent over the whole time-frame hence I moved to a new model, the Dynamic Topic Modeling [21] that allows to classify the whole corpus, divided into suitable time slices consistently over time. After some optimisation of the model I settle for 5 topics which were labeled as "quarantine", "other, "work", "government" and "healthcare". Among them, "healthcare" is consistently the predominant in all of the 13 weekly time slices in our corpus and the peaks in the share of documents related to this topic happened around major announcement of public health measures. The stacked chart presented in the previous section shows that after ever major peak there is a descending trend which is a proof of our initial intuition of the fast shift in topics in the public debate. I also analysed in the same way how to share of documents labeled for each of hurtlex categories evolved over times and compares the two outcomes. It turned out that the most evident peaks in the HurtLex categories distribution happen exactly in the same time slices were the topics "quarantine" and "healthcare" have spikes as well, showing that the most heated debates happened around public measures that affected directly and immediately on both the collectivity ("healthcare") and personal life ("quarantine"). I then tried to use all the information gained so far to enhance the hate speech prediction performed by means of AlBERTo. Unfortunately this experiment did not lead to significant results due to the very small size of the resulting training dataset. Infusing deep learning model with information extracted from topic modeling sounds certainly a promising way to enhance the accuracy of hate speech prediction, but I believe that this method could be greatly improved in different way. To start with, our dataset was collected with a fixed set of hashtags and keywords, while a more flexible and time-evolving approach could lead to a more insightful data collection. Secondly our corpus was labelled using guidelines derived from another hate speech detection task, whole ad-hoc rules and more annotators would certainly improve the quality of the gold standard. This step would also guarantee a bigger training set for

AlBERTo, which would have certainly positive effects on the accuracy of the prediction.

For what regards the lexicon approach, we mentioned that a potentially very interesting way to reduce the noise in the classification encompasses an in depth manual revision of the words in each category aimed at removing the most generic terms. In conclusion, our method helped us to prove that topics of discussion on social media not only follow closely the flow of real life relevant events but also change rapidly over time. Further work on this issue is crucial in improving the performances of algorithms for NLP task on social media linguistic data, as such models need to be time robust to capture and learn as precisely as possible all the possible nuances on how language evolves over time.

RQ3 The last research question investigates how to leverage socio-demographic information about users who post abusive contents and the distribution of such messages.

This research, presented in 5 is based on my first paper [51] where, inspired by [85], we tried to combined the geographical distribution of hateful messages with more traditional socio-demographics indexes, following the rational that online hate is not a phenomena per se but, for a deeper insight on it it is useful to broader the investigation and analyse the offline characteristics of users who tend to be more prone to generate such texts.

However, the technological challenges involved in a large-scale study generate a high-level of uncertainty, which we aim to address by retrieving a larger set of annotated tweets.

Interestingly, our results indicate a higher correlation between online hate speech and more foreigners in the workforce, and also a higher correlation between online hate speech and the level of education of migrants. Such signals may help to shed some light on the phenomena well-described in the most recent social studies on sociology of migration in Europe, where “the unresolved tension between the economics of the European approach to labour migration and the philosophy of rights and solidarity" is highlighted [126]. The social expectations about the role of migrant workers is often based on a concept of complementarity between autochthonous and foreign work. When we focus on Italy, survey-based social studies report that people express uncertainty toward a positive impact of immigration on the society, but the prevalent opinion is that immigrants are doing jobs that Italians no longer want to do, so contributing to improve Italian economy. This scenario is of course perceived as advantageous for Italians, since migrant workers and locals are not competing for the same jobs. Paradoxical as it may seem, where higher education and better integration in the labour market of the migrant workers occur, this is possibly perceived as a threat to employment of Italians, that finds its voice in social media as hate speech. This could provide a possible interpretation of the signals we detected, that deserves further investigation in future work, based on the analysis of finer-grained statistical data on work type and employment status.

As reported by social studies the other factor that seems to play a role in the generation of cyberhate is personal safety, or at least its perception. This finds a reflection in the stark contrast between the positive correlation of online hate speech with the level of petty theft (perceived as a personal danger) and the lack of correlation with equally serious crimes (such as counterfeiting or prostitution), that are perhaps less perceived as a threat to people’s safety.

Future Works This thesis described a manifold approach to the crucial modern issues of hate speech detection in social media, highlighting strengths and weakness of each of the approach we followed. I think that the highlighted research questions could have deeper and more insightful answers by tackling in more details the following issues.

- **Unbalanced data:** Our annotated data are naturally very unbalanced, with non-hate speech examples representing most of the dataset. It is commonly known that the performance of machine learning approaches is strongly influenced by the class unbalance, and consequently, it could be very interesting for the future to investigate the impact of automatic balancing techniques or the addition of new training data on the robustness observed in the models we analyzed, following for example what presented in [62] .
- **Digital Divide:**The finding of this research would certainly lead to even more interesting insights if coupled with more detailed meta information about the demographic of Twitter users in Italy. More specifically the evolution of internet penetration rate (especially post Covid-19 pandemic lockdown in 2020) could serve as a proxy for the accessibility of online services to the general population in Italy and give some general context useful to better interpret the geographical distribution of HS messages volume over time.
- **Data Sparsity:** Geotagged data are commonly very sparse, hence different inference method have been developed over time (see for example [32]). The implementation of such technique would allow us a finer grained geographical mapping of HS messages.
- **Sociology Literature:** a deeper investigation of the sociology literature can be helpful in order to test hypotheses on the origins of hate crime, validating them by retrieving additional specific demographic data. Furthermore, a more in depth exploration of the interplay between cyberhate and other indicators and events, such as the outcome of national and local political elections could lead to interesting finding about any specificity about Italy (see for example [108, 18, 104].
- **Crime rates:** The ultimate goal and hope of these analysis is to try and monitor and hopefully prevent both online and offline hate crimes. Therefore a fine grained temporal and geographical mapping of the evolution of real life reported crimes could shed more light on the issue.

We hope to be able to pursue this research path in multiple direction in the near future in the hope to give a humble contribution in making social media a safe online venue for respectful debates about facts and opinions.

Appendix A

Daily percentage of HS tweets in 40wita

Day	February 2020	March 2020	April 2020
1	4.071	1.442	0
2	0	0.667	0
3	0.188	0	0
4	0.911	1.697	0
5	0.963	0	0
6	0.391	1.794	0
7	1.954	0.413	0
8	0.96	3.833	0
9	0.782	2.964	0
10	0.471	0	0
11	6.088	0	0
12	1.583	0	0
13	1.583	0	1.451
14	1.583	0	1.525
15	0.397	0	0
16	0.834	0	1.125
17	0	0	0.64
18	0	0	0
19	0.49	0	0
20	0	0	2.57
21	0.49	0	2.141
22	0	0	2.069
23	0.49	0	3.078
24	1.584	0	0
25	1.54	0	1.95
26	1.192	0	0
27	0.441	0	0.914
28	2.587	0	5.024
29	0.221	0	0.468
30	-	0	1.085
31	-	0	-

TABLE A.1: Daily percentage of tweets labeled as hate speech in February, March and April 2020, automatically classified by AlBERTo with same hyper parameters as in Section 3.

Appendix B

Hurtlex Categories samples

Hurtlex Category	Tweet Text
cds (derogatory words)	#immuni sai cosa se ne frega uno che muore di fame o di sto coronavirus, coglioni pagate le casse integrazioni e i 600 euro *#immuni who cares if one dies starving or from covid, ass* pay your ordinary layoff and the 600 euros*
cds	Il problema alla base di questa crisi siete voi che avete votato le promesse e non i fatti in questi anni, meditate depensanti #Fase2 #COVID19 #CrisiCoronavirus *The problem at the base of this crises is you who voted the promises and not the facts, think about it mindless*
cds	I cani sono gli unici felici della quarantena. Il mio cane ora non è mai solo, gioca tutto il giorno con qualcuno e riceve dolcetti da tutti *Dogs are the only happy ones about the quarantine. My dog is never alone, plays all the time with someone and gets treats from everybody*
cds	L'emergenza #Covid19 ha evidenziato la pochezza dello #Stato, la tracotanza della #scienza moderna e la diserzione della #Chiesa, lasciando l'intera popolazione in uno sgomento che sa di quiete prima della tempesta https://t.co/0mbpmtDgIL *The #covid emergency highlighted the shortcomings of the State, the arrogance of modern science and the defection of the Church, leaving the entire population in a state of dismay that feels like the calm before the storm*
cds	La mamma di un mio alunno ha detto: " in questo periodo, non vorrei mai essere Conte e la maestra " ha ragione . Siete una gran rottura di coglioni! #COVID19 #Conte #30aprile #Parlamento #scuola *The mother of one of my students say:" in these period, I'd rather not be Conte or the teacher". She is right. You are all a bunch of ass**
dmc (moral and behavioural defects)	Quando credete che la vita vi stia andando male, pensate a me che avevo contemporaneamente SEI mascara aperti prima che iniziasse la quarantena

Table B.1 continued from previous page

Hurtlex Category	Tweet Text
	When you think your life is bad, thinking of me who had six mascara opened at the same time before the quarantine
dmc	Coronavirus nel Milanese, una volpe entra in giardino e si rifugia nella cuccia dei cani https://t.co/l63aqteM6e *Coronavirus in Milan interland, a fox enters in a garden and seeks shelter in dogs house*
dmc	#app #immuni che ti dice sei vicino a positivo a #covid19 app inutile. Non ci devo arrivare vicino ad un positivo. Questo il punto. #agorarai *#app #immuni tells you if you were close to a #covid19 positive useless. I should have not been close to a positive persone. This is the point #agorarai*
dmc	Renzi in aula "Il #Coronavirus e' una bestia terribile che ha fatto 30mila morti. Onoriamo quei morti. La gente di Bergamo e Brescia che non c'e' piu', se potesse parlare ci direbbe di riaprire". *Renzi at the Parliament "The Coronavirus is a terrible beast that killed 30K people. Let's honor the deceased. People from Bergamo and Brescia that is not here anymore would ask to open everything.*
dmc	Sito utile per chi fosse interessato/a ai dati statistici su covid-19. https://t.co/BWs9SlG5ol *Useful website for those interested in statistical data about covid19*
ddp (cognitive disabilities and diversity)	Covid19 Sicilia, diminuiscono i nuovi contagi, solo 20, 763 guariti e nessuna nuova vittima | BlogSicilia - Ultime notizie dalla Sicilia https://t.co/OAGfpIO5op *Covid19 Sicily, decrease of cases, only 20, 763 healed and no new victim*
ddp	Coronavirus, Winston è il primo cane contagiato da Covid-19 negli Usa https://t.co/EyjhdIIDzF *Coronavirus, Winston is the first infected dog in the US*
ddp	chissà che durante la quarantena gli americani si diano una svegliata e smettano di votare trump vedendo come sta gestendo di merda la situa *maybe during the quarantine american people will wake up and stop voting trump after seeing the sh* management of the situation*
ddp	Una cosa molto semplice è rendere per me enormi vantaggi finanziari, piuttosto che piangere ogni volta per i virus Corona e Covid 19. Il 4 maggio 2020 l'Italia aprirà il blocco dai virus Corona e Covid 19. https://t.co/DWHp6Lw5ga *A very simple thing is make for me eaasy enormous financial advantages, instead of cry every time for Corona and Covid19 viruses. On 4th May 2020 Italy will re open the block from virus Corona and Covid 19*

Table B.1 continued from previous page

Hurtlex Category	**Tweet Text**
ddp	@matteorenzi mio padre lavorava in ospedale, è morto di covid a Torino e sono sicuro che ti darebbe del coglione per la stronzata che hai detto. *@matteorenzi my dad used to work in a hospital , and died from covid in Torino and I am sure that he will call you an ass* for the bull* you said*

TABLE B.1: A sample of tweets labelled as belonging to one of the three most frequent categories in the HurtLex lexicon.

Bibliography

[1] Martín Abadi et al. "TensorFlow: A System for Large-Scale Machine Learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*. Ed. by Kimberly Keeton and Timothy Roscoe. USENIX Association, 2016, pp. 265–283.

[2] Sweta Agrawal and Amit Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms". In: *European conference on information retrieval*. Springer. 2018, pp. 141–153.

[3] Sohail Akhtar, Valerio Basile, and Viviana Patti. "Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 1. Association for the Advancement of Artificial Intelligence. 2020, pp. 151–154.

[4] Rubayyi Alghamdi and Khalid Alfalqi. "A survey of topic modeling in text mining". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 6.1 (2015).

[5] Wafa Alorainy et al. "Suspended Accounts: A Source of Tweets with Disgust and Anger Emotions for Augmenting Hate Speech Data Sample". In: *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. Vol. 2. IEEE. 20128, pp. 581–586.

[6] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Vol. 9. Cambridge University Press, 1999.

[7] Aymé Arango, Jorge Pérez, and Barbara Poblete. "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)". In: *Information Systems* (2020), p. 101584.

[8] Lora Aroyo and Chris Welty. "Truth is a lie: Crowd truth and the seven myths of human annotation". In: *AI Magazine* 36.1 (2015), pp. 15–24.

[9] Imran Awan and Irene Zempi. "The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts". In: *Aggression and violent behavior* 27 (2016), pp. 1–8.

[10] Femi Emmanuel Ayo et al. "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions". In: *Computer Science Review* 38 (2020), p. 100311.

[11] Pinkesh Badjatiya et al. "Deep learning for hate speech detection in tweets". In: *Proceedings of the 26th international conference on World Wide Web companion*. 2017, pp. 759–760.

[12] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. "TRI: a tool for the diachronic analysis of large corpora and social media". In: *Proceedings of the 7th AIUCD Annual Conference Cultural Heritage in the Digital Age. Memory, Humanities and Technologies*. Bari, Italy, 2018.

[13] Valerio Basile. "It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks". In: *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*. Vol. 2776. CEUR-WS. 2020, pp. 31–40.

[14] Valerio Basile and Tommaso Caselli. *40twita 1.0: A collection of Italian Tweets during the COVID-19 Pandemic*. http://twita.di.unito.it/dataset/40wita.

[15] Valerio Basile, Mirko Lai, and Manuela Sanguinetti. "Long-term Social Media Data Collection at the University of Turin". In: CEUR Workshop Proceedings 2253 (2018), pp. 1–6. URL: http://ceur-ws.org/Vol-2253/paper48.pdf.

[16] Valerio Basile et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007.

[17] Elisa Bassignana, Valerio Basile, and Viviana Patti. "Hurtlex: A multilingual lexicon of words to hurt". In: *5th Italian Conference on Computational Linguistics, CLiC-it 2018*. Vol. 2253. CEUR-WS. 2018, pp. 1–6.

[18] Adam Bermingham and Alan Smeaton. "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)*. 2011, pp. 2–10.

[19] Shan Bin and Li Fang. "A survey of topic evolution based on LDA". In: *Journal of Chinese Information Processing* 24.6 (2010), pp. 43–49.

[20] Grant Blank. "The digital divide among Twitter users and its implications for social research". In: *Social Science Computer Review* 35.6 (2017), pp. 679–697.

[21] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[23] Eurostat Statistical Books. "People in the EU: Who are we and how do we live". In: *European Union, Luxembourg: Publications Office of the European Union* (2015).

[24] Cristina Bosco et al. "Overview of the EVALITA 2018 Hate Speech Detection Task". In: CEUR Workshop Proceedings 2263 (2018), pp. 1–9.

[25] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[26] Arthur T. E. Capozzi et al. "A Data Viz Platform As a Support to Study, Analyze and Understand the Hate Speech Phenomenon". In: *Proceedings of the 2nd International Conference on Web Studies*. Paris, France: ACM, 2018, pp. 28–35.

[27] Arthur TE Capozzi et al. "Computational linguistics against hate: Hate speech detection and visualization on social media in the" Contro L'Odio" project". In: CEUR Workshop Proceedings 2481 (2019), pp. 1–6.

[28] Jason Chan, Anindya Ghose, and Robert Seamans. "The internet and racial hate crime: Offline spillovers from online access". In: *Mis Quarterly* 40.2 (2016), pp. 381–403.

[29] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[30] Jonathan Chang et al. "Reading tea leaves: How humans interpret topic models". In: *Neural information processing systems*. Vol. 22. Citeseer. 2009, pp. 288–296.

[31] Irfan Chaudhry. "# Hashtagging hate: Using Twitter to track racism online". In: *First Monday* 20.2 (2015).

[32] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 759–768.

[33] Vladimir Cherkassky and Yunqian Ma. "Practical selection of SVM parameters and noise estimation for SVM regression". In: *Neural networks* 17.1 (2004), pp. 113–126. DOI: https://doi.org/10.1016/S0893-6080(03)00169-2.

[34] Michael Clyne, Michael G Clyne, and Clyne Michael. *Dynamics of language contact: English and immigrant languages*. Cambridge University Press, 2003.

[35] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[36] Michele Corazza et al. "A Multilingual Evaluation for Online Hate Speech Detection". In: *ACM Trans. Internet Technol.* 20.2 (2020). ISSN: 1533-5399. DOI: 10.1145/3377323. URL: https://doi.org/10.1145/3377323.

[37] Andrew M. Dai and Quoc V. Le. "Semi-supervised Sequence Learning". In: (2015). Ed. by Corinna Cortes et al., pp. 3079–3087. URL: http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.

[38] Ali Daud et al. "Knowledge discovery through directed probabilistic topic models: a survey". In: *Frontiers of computer science in China* 4.2 (2010), pp. 280–301.

[39] Thomas Davidson et al. "Automated hate speech detection and the problem of offensive language". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017.

[40] Michela Del Vicario et al. "Echo chambers: Emotional contagion and group polarization on facebook". In: *Scientific reports* 6.1 (2016), pp. 1–12.

[41] Fabio Del Vigna et al. "Hate me, hate me not: Hate speech detection on Facebook". In: *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. Venice, Italy, 2017.

[42] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423/.

[43] Gonzalo Donoso and David Sánchez. "Dialectometric analysis of language variation in Twitter". In: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 16–25. DOI: 10.18653/v1/W17-1202. URL: https://www.aclweb.org/anthology/W17-1202.

[44] Antoine Dubois et al. "Studying migrant assimilation through facebook interests". In: *International Conference on Social Informatics*. Springer, 2018, pp. 51–60.

[45] Kinda El Maarry, Kristy Milland, and Wolf-Tilo Balke. "A fair share of the work? The evolving ecosystem of crowd workers". In: *Proceedings of the 10th acm conference on web science*. 2018, pp. 145–152.

[46] M Feindt and U Kerzel. "The NeuroBayes neural network package". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 559.1 (2006), pp. 190–194.

[47] Ronen Feldman. "Techniques and applications for sentiment analysis". In: *Communications of the ACM* 56.4 (2013), pp. 82–89.

[48] Lee Fiorio et al. "Using Twitter data to estimate the relationship between short-term mobility and long-term migration". In: *Proceedings of the 2017 ACM on Web Science Conference*. ACM. 2017, pp. 103–110.

[49] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378.

[50] Komal Florio, Valerio Basile, and Viviana Patti. "Hate Speech and Topic Shift in the Covid-19 Public Discourse on Social Media in Italy". In: *to apper in the Proccedings of the 8th Italian Conference on Computational Linguistics, CLiC-it 2021*. CEUR-WS. 2021.

[51] Komal Florio et al. "Leveraging Hate Speech Detection to Investigate Immigration related Phenomena in Italy". In: *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. 2019, pp. 1–7. DOI: 10.1109/ACIIW.2019.8925079.

[52] Komal Florio et al. "Time of your hate: The challenge of time in hate speech detection on social media". In: *Applied Sciences* 10.12 (2020), p. 4180.

[53] Karën Fort, Gilles Adda, and K Bretonnel Cohen. "Amazon Mechanical Turk: Gold mine or coal mine?" In: *Computational Linguistics* 37.2 (2011), pp. 413–420.

[54] Paula Fortuna and Sergio Nunes. "A survey on automatic detection of hate speech in text". In: *ACM Computing Surveys (CSUR)* 51.4 (2018), p. 85.

[55] Rahul Goel et al. "The social dynamics of language change in online networks". In: Lecture Notes in Computer Science 10046 (2016), pp. 41–57. DOI: 10.1007/978-3-319-47880-7_3.

[56] Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: *arXiv preprint arXiv:1402.3722* (2014).

[57] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[58] Arepalli Peda Gopi et al. "Classification of tweets data based on polarity using improved RBF kernel of SVM". In: *International Journal of Information Technology* (2020), pp. 1–16.

[59] Xinjian Guo et al. "On the class imbalance problem". In: *2008 Fourth international conference on natural computation*. Vol. 4. IEEE. 2008, pp. 192–201.

[60] Robert Hecht-Nielsen. "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[61] J Hellrich. *Word Embeddings: Reliability & Semantic Change*. Amsterdam, The Netherlands: IOS Press, 2019. URL: https://books.google.it/books?id=92OwDwAAQBAJ.

[62] Delia Irazu Hernandez Farıas et al. "Irony detection in Twitter with imbalanced class distributions". In: *Journal of Intelligent & Fuzzy Systems* 39.2 (2020), pp. 2147–2163.

[63] Thomas Hofmann. "Unsupervised learning by probabilistic latent semantic analysis". In: *Machine learning* 42.1 (2001), pp. 177–196.

[64] Mokter Hossain and Ilkka Kauranen. "Crowdsourcing: A comprehensive literature review". In: *Strateg. Outsourcing* 8.1 (2015), pp. 2–22. ISSN: 17538300. DOI: 10.1108/SO-12-2014-0029. arXiv: 0803973233.

[65] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://www.aclweb.org/anthology/P18-1031/.

[66] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv:1508.01991* (2015).

[67] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *CoRR* abs/1605.05894 (2016). arXiv: 1605.05894. URL: http://arxiv.org/abs/1605.05894.

[68] Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. "Diachronic degradation of language models: Insights from social media". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 195–200.

[69] Nathalie Japkowicz and Shaju Stephen. "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[70] Hamed Jelodar et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11 (2019), pp. 15169–15211.

[71] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[72] Anne Kao and Steve R Poteet. *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[73] Gurvir Kaur and Er Parvinder Kaur. "Novel approach to text classification by SVM-RBF kernel and linear SVC". In: *International Journal of Advance Research, Ideas and Innovation in Technology* 3.3 (2017), pp. 1014–7.

[74] A Khurshid, L Gillman, and L Tostevin. "Weirdness indexing for logical document extrapolation and retrieval". In: *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. 2000.

[75] Barbara Kitchenham et al. “Systematic literature reviews in software engineering–a systematic literature review”. In: *Information and software technology* 51.1 (2009), pp. 7–15.

[76] Manfred Klenner et al. “Harmonization sometimes harms”. In: swisstext-and-konvens-2020, 2020.

[77] Klaus Krippendorff. “Estimating the reliability, systematic error and random error of interval data”. In: *Educational and Psychological Measurement* 30.1 (1970), pp. 61–70.

[78] “Let’s Agree to Disagree: Fixing Agreement Measures for Crowdsourcing”. In: *Fifth AAAI Conference on Human Computation and Crowdsourcing*. 2017. URL: https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/viewFile/15927/15258.

[79] Erez Lieberman et al. “Quantifying the evolutionary dynamics of language”. In: *Nature* 449.7163 (2007), pp. 713–716.

[80] Jonathan Mellon and Christopher Prosser. “Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users”. In: *Research & Politics* 4.3 (2017), p. 2053168017720008.

[81] Letizia Mencarini et al. “Happy Parents’ Tweets An exploration of Italian Twitter Data with Sentiment Analysis”. In: *Demographic Research, Special Collection on Social Media* 40.25 (2019), pp. 693–724. DOI: 10.4054/DemRes.2019.40.25.

[82] Stefano Menini et al. “A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 105–110. DOI: 10.18653/v1/W19-3511. URL: https://www.aclweb.org/anthology/W19-3511.

[83] Johnnatan Messias et al. “From migration corridors to clusters: The value of Google+ data for migration studies”. In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press. 2016, pp. 421–428.

[84] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: NIPS’13 (Dec. 5, 2013), pp. 3111–3119.

[85] Lewis Mitchell et al. “The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place”. In: *PloS one* 8.5 (2013), e64417.

[86] Fred Morstatter et al. “Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose”. In: *Seventh international AAAI conference on weblogs and social media*. 2013.

[87] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. “Natural language processing: an introduction”. In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 544–551.

[88] Stefanie Nowak and Stefan Rüger. “How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation”. In: *Proceedings of the international conference on Multimedia information retrieval*. 2010, pp. 557–566.

[89] Alexandra Olteanu, Emre Kıcıman, and Carlos Castillo. "A Critical Review of Online Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 785–786.

[90] Alexandra Olteanu et al. "The effect of extremist violence on hateful speech online". In: *Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM 2018)*. AAAI Press, 2018, pp. 221–230. URL: http://www.aaai.org/Library/ICWSM/icwsm18contents.php.

[91] Endang Wahyu Pamungkas and Viviana Patti. "Cross domain and Cross lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 363–370. DOI: 10.18653/v1/P19-2051. URL: https://www.aclweb.org/anthology/P19-2051.

[92] Demetris Paschalides et al. "MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech". In: *ACM Trans. Internet Technol.* 20.2 (2020). ISSN: 1533-5399. DOI: 10.1145/3371276. URL: https://doi.org/10.1145/3371276.

[93] Matthew E. Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/n18-1202. URL: https://doi.org/10.18653/v1/n18-1202.

[94] Fabio Poletto et al. "Annotating hate speech: Three schemes at comparison". In: *6th Italian Conference on Computational Linguistics, CLiC-it 2019*. Vol. 2481. CEUR-WS. 2019, pp. 1–8.

[95] Fabio Poletto et al. "Hate speech annotation: Analysis of an Italian Twitter corpus". In: CEUR Workshop Proceedings 2006 (2017), pp. 1–6.

[96] Fabio Poletto et al. "Resources and benchmark corpora for hate speech detection: a systematic review". In: *Lang. Resour. Evaluation* 55.2 (2021), pp. 477–523. DOI: 10.1007/s10579-020-09502-8. URL: https://doi.org/10.1007/s10579-020-09502-8.

[97] Marco Polignano et al. "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets". In: CEUR Workshop Proceedings 2481 (2019).

[98] Marco Polignano et al. "AlBERTo: Modeling Italian Social Media Language with BERT". In: *Italian Journal of Computational Linguistics - IJCOL* -2, n.2 (2019).

[99] Marco Polignano et al. "Hate Speech Detection through AlBERTo Italian Language Understanding Model". In: CEUR Workshop Proceedings 2521 (2019).

[100] Jipeng Qiang et al. "Short text topic modeling techniques, applications, and performance: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* (2020).

[101] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018). https://www.cs.ubc.ca/ amuham01/LING530/papers/radford2018improving

[102] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019), p. 9.

[103] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. Piscataway, NJ. Dec. 3, 2003, pp. 133–142.

[104] Jyoti Ramteke et al. "Election result prediction using Twitter sentiment analysis". In: *2016 international conference on inventive computation technologies (ICICT)*. Vol. 1. IEEE. 2016, pp. 1–5.

[105] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[106] Björn Ross et al. "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis". In: *CoRR* abs/1701.08118 (2017). arXiv: 1701.08118. URL: http://arxiv.org/abs/1701.08118.

[107] Marta Sabou et al. "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines." In: *LREC*. Citeseer. 2014, pp. 859–866.

[108] Erik Tjong Kim Sang and Johan Bos. "Predicting the 2011 Dutch Senate Election Results with Twitter". In: EACL '12 (2012), pp. 53–60. URL: http://dl.acm.org/citation.cfm?id=2389969.2389976.

[109] Manuela Sanguinetti et al. "An italian Twitter corpus of hate speech against immigrants". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018, pp. 1–8. URL: https://www.aclweb.org/anthology/L18-1443.

[110] Anna Schmidt and Michael Wiegand. "A survey on hate speech detection using natural language processing". In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 2017, pp. 1–10.

[111] Rion Snow et al. "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks". In: *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008, pp. 254–263.

[112] Spyridon Spyratos et al. "Migration data using social media: a European perspective". In: Publications Office of the European Union, 2018.

[113] Julia Maria Struß et al. "Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019, pp. 354–365.

[114] Nina Tahmasebia, Lars Borina, and Adam Jatowtb. "Survey of Computational Approaches to Lexical Semantic Change". In: *Computational approaches to semantic change* 6 (2021), p. 1.

[115] M Trupthi, Suresh Pabboju, and G Narasimha. "Sentiment analysis on twitter using streaming API". In: *2017 IEEE 7th International Advance Computing Conference (IACC)*. IEEE. 2017, pp. 915–919.

[116] Raghavendra Vijay Bhasker Vangara, Shiva Prasad Vangara, and VR Kailashnath Thirupathur. "A Survey on Natural Language Processing in context with Machine Learning". In: *The International journal of analytical and experimental modal analysis* XII (I 2020).

[117] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Long Beach, California: Curran Associates, Inc., 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[118] Bertie Vidgen and Leon Derczynski. "Directions in abusive language training data, a systematic review: Garbage in, garbage out". In: *PLOS ONE* 15.12 (Dec. 2021), pp. 1–32. DOI: 10.1371/journal.pone.0243300. URL: https://doi.org/10.1371/journal.pone.0243300.

[119] Zeerak Waseem and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[120] Ingmar Weber and Bogdan State. "Digital Demography". In: WWW '17 Companion (2017), pp. 935–939. DOI: 10.1145/3041021.3051104. URL: https://doi.org/10.1145/3041021.3051104.

[121] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. "Detection of abusive language: the problem of biased datasets". In: *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 602–608.

[122] Lars Wissler et al. "The Gold Standard in Corpus Annotation." In: *IEEE GSC*. 2014.

[123] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. "Inferring international and internal migration patterns from Twitter data". In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 439–444.

[124] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants". In: *Population and Development Review* 43.4 (2017), pp. 721–734.

[125] Emilio Zagheni et al. "Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks". In: *Annual Meeting of the Population Association of America*. 2018.

[126] Laura Zanfrini. *The Challenge of Migration in a Janus-Faced Europe*. Springer International Publishing, 2018.

[127] *Intra-and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing*. 2019, pp. 1138–1143.

[128] Yukun Zhu et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.