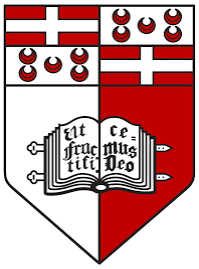
Automatic detection of hate speech in social media

Claudia Zaghi

MSc. Dissertation



Department of Artificial Intelligence
Faculty of Information and Communication Technology
University of Malta
2018

Supervisors:
Malvina Nissim, Faculty of Arts - University of Groningen
Albert Gatt, Institute of Linguistics and Language Technology - University of Malta

Submitted in partial fulfillment of the requirements for the Degree of
European Master of Science in Human Language Science and Technology

M.Sc. (HLST)
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own work, of work of another person, whether or not such work has been published" (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master's dissertation submitted is my, Claudia Zaghi, own work, except where acknowledged and referenced.

I, Claudia Zaghi, understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name: *Claudia Zaghi*
Course Code: CSA5310 HLST Dissertation
Title of work: *Automatic detection of hate speech in social media*

Signature of Student:

Date: 04/02/2019

ABSTRACT

Hate speech is "the use of aggressive, hatred or offensive language, targeting a specific group of people sharing a common trait: their gender, ethnic group, race, religion, sexual orientation, or disability" (Mish and Morse, Mish and Morse).

As the phenomenon is widely spreading online (Gagliardone et al., 2015), social networks and websites have introduced a progressively stricter code of conduct and regularly removed offensive content flagged by users (Bleich, 2014). However, the volume of data in social media makes it challenging to supervise the published content across platforms.

This research focuses on hate speech in Italian, aiming to automatically detect hateful content based on data scraped from different social media sites. Using the technique of distant supervision (Go et al., 2009), we automatically developed labeled datasets for machine learning experiments as well as hate-polarized word embeddings.

We tackled the challenge of hate speech detection by training a simple binary classifier, characterized by a Linear Support Vector Classification (SVC) model and n-grams features. We compared the performance of the classifier when trained and tested over manually and automatically annotated datasets, and resources containing hatred against a single target versus multiple targets.

The results of the study highlighted the effectiveness of manually labeled data (80% vs. 45% in F-1 score) and the versatility of distantly supervised data, as sections of automatically labeled data can be used to enrich small manually labeled datasets. The polarized word embeddings proved to be more predictive than off-the-shelf dense vectors (81% vs. 79% in F-1 score). Additionally, the experiments showed that the language of *haters* is very similar across targets.

CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I Sufi ci consigliano di parlare soltanto quando le nostre parole sono riuscite a passare attraverso tre cancelli.
Al primo cancello ci chiediamo: "Sono vere queste parole?"
Se lo sono, le lasciamo passare, se non lo sono, le rimandiamo indietro.
Al secondo cancello, ci domandiamo "Sono necessarie?
All'ultimo cancello invece chiediamo: "Sono gentili?

- Eknath Easwaran,

I would like to thank the LCT program for creating such a stimulating and useful opportunity for us linguists.

I would like to thank my supervisors, Malvina and Albert, for the trust and for teaching your courses with so much passion.
Flavio, Tommaso, and Xiaoyu, you also deserve a special thanks for the help and support you gave me throughout the composition of this work.

I would also like to take some space here to express my deepest love and gratitude to all the people who were by my side during the two most intense years of my life.

To my friends Petra, Livia, and Rebecca for always being present through the years regardless of the distance between us.

To my parents, my sister and all my family. Nothing of this would have been possible without your constant support, love and patience.

To Brandon, I cannot even put into words my gratitude. Thank you, grazie, grazzi, dank for leaving California and moving with me to Malta and The Netherlands. Without you, your daily support and jokes, I would have never made it this far. I can't wait to see what's ahead of us now.

1 INTRODUCTION

According to Statista (Statista, 2018), the daily social media usage of global Internet users amounted to 135 minutes (9% of someone's day) in 2017, up from 126 daily minutes in the previous year. People use social media for posting, sharing, and streaming content about their families, successes, political opinions, and lives. However, there is a universe consisting of people, who for one reason or another, publish hate speech, troll, and or cyber-bully (Delgado and Stefancic, 2004).

Hate speech is a widespread phenomenon, whose presence is so consistent to have become an accepted reality (Silva et al., 2016). In practice, it consists of offensive expressions addressed towards communities of people who share a common feature: from sexual, religious, dietary orientations to nationality and disabilities (Waldron, 2012).

Online haters are not relegated to a specific demographic. Instead, they are men and women of all ages and demographic type who, behind a screen, feel protected and comfortable to publish hate speech towards specific targets (Ziccardi, 2016).
However, the publication of hate speech is a dangerous and illegal practice that needs to be discouraged and eliminated using automatic and accurate tools (Brugger, 2002).

This dissertation focuses on automated hate speech detection in Italian. Our thesis is that automatically generated datasets could be as effective as manually annotated datasets. We also predict that hate speech addressed to groups of targets would utilize words related to those targets which could then be used in identifying hate.
Generally, the research aims at *addressing the issue of hate speech in Italian by proposing automatic solutions to detect and monitor online offensive content*.

The research goals that this dissertation aims at addressing are the following:

- provide a comprehensive overview of the topic, by defining hate speech, distinguishing it from related term and explaining how previous work has tackled the detection of this phenomenon.

- describe how we *annotated a dataset for the Italian language*.
  The majority of systems to perform text classification are supervised, thus requiring the manual annotation of training data. According to FigureEight [1], annotators are paid $0.13 per annotation, which means that the creation of a polarized word embedding dataset would have cost us over $130,000. This was out of the study's budget, so we began the search for an alternative method of annotation.

- investigate the advantages of using manually labeled (gold) data versus automatically labeled (silver) data, and find the advantages of us-

[1] https://www.figure-eight.com/company/

ing datasets with hatred addressed to a single target (e.g. the politician's community) versus hatred addressed to multiple targets (e.g. women or vegans).

- build a supervised learning model to perform hate speech detection. The large size of the data that we have gathered and want to explore restricted the number of algorithm to use. Thus, we narrowed the choice down to Linear Support Vector Machine, Logistic Regression and Naive Bayes (Buitinck et al., 2013).

  The chosen features to be added to the model include: a lexicon of hateful and offensive words features (De Mauro, 2016), polarized word embeddings that we generated, pretrained word embeddings [2] and word and characters n-grams.

  The use of n-grams has been successfully used in the research of hate speech detection (Davidson et al., 2017; Waseem and Hovy, 2016; Burnap and Williams, 2016; Magu et al., 2017) as it allows to achieve the baseline result (F-score 80%).
  We plan to add hate polarized lexicon and word embeddings as features to tune the predictive power towards the hateful content present in the social media content.

Thesis outline

Regarding the outline of our work, we dedicate the second chapter to find an accepted definition of hate speech and we provide an overview of the literature review on the topic.

In the third chapter, we investigate the datasets that we exploited in the machine learning experiments.

The fourth chapter is dedicated to the description of the model and our work on feature engineering.

The subsequent chapter consists of the outline and description of the results.

Finally, chapter six and seven contain the conclusion, the limitations and practical applications of our study.

[2] http://www.spinningbytes.com/resources/wordembeddings/

2 AUTOMATIC HATE SPEECH DETECTION: BACKGROUND

This section aims to provide a summary of the work conducted so far on hate speech detection.

We address the topic systematically, providing both theoretical and practical aspects and giving an overview of the most recent approaches. The first section concentrates on *Defining Hate Speech*, exploring the theoretical definitions of hate speech. We also discuss concepts, such as harassment and cyber-bullying, which are often mistaken for hate speech, with the aim to distinguish the topic of this thesis from related terms.

We proceed with the discussion of the previous research on the task of automatic hate speech detection that has been held on Italian and English, in the section *Previous work on Hate Speech in Italian* and *Previous work on Hate Speech in English*.

Finally, we consider the importance of researching hate speech detection and what are the possible difficulties that researchers might encounter when investigating the topic.

2.1 DEFINING HATE SPEECH

Hate speech is a controversial topic because its definition varies across time, place and, currently, also across online platforms (Waldron, 2012). This is the reason why we decided to dedicate a section to shed light on how hate speech has been perceived and defined so far. In Table 1 we propose a historical analysis of the laws that have spurred interest in this phenomenon around Europe and specifically, in Italy.

Historically, it has been noticed a transition from a general definition of hate speech to one that specifies hateful activities on social media, which is the topic of this research (Silva et al., 2016) . Offline and online hate speech have distinctive traits, online hate is characterized by:

- **Permanence and possibility of coming back** : online hate speech can be active for long periods of time. Hateful content, violating both people's public and private privacy, can become viral, triggering an avalanche of sharing across the internet (Mills, 2012).
- **Anonymity** : social media users have the ability to share content online without displaying their identity, believing anonymity allows them to post with impunity from both the platform guidelines and the law. Ziccardi (2016) reported on the phenomenon of hate speech in Italy. The writer explained that, even if social media allows Italian users to hide their private information, they tend to maintain their name and surname public when publishing hateful comments.

Additionally, the research showed that not only do users feel comfortable when publishing offensive content, but also the users that make positive use of social media platforms are getting more tolerant to the display of hateful manners.

Another important set of hate speech definitions come from the leading social media companies. Facebook, Twitter, and YouTube each have included in their guidelines a specific reference to hate speech. They clarify what they consider to be offensive content and how to report it. Table 2 reports the sections of the guidelines which refer to hate speech.

Hate speech and related terms

The research in the data scraped from social media draws attention not only to hate speech but also to related topics, which are often confused with the notion of hate speech. In 3, we aim at defining the differences among the terms closely related to hate speech.

After investigating the definitions of hate speech over time and in different sources, we have all the elements to find common patterns among them and arrive at a comprehensive description of the term.

- The target can be one or more individuals associated with a group that shares particular characteristics or the group itself.

Table 1: Comparison of hate speech definitions across time and institutions.

Source	Year	Definition
Council of Europe's Committee of Ministers	1997	Recommendation No. R (97). The recommendation defines *hate speech* as a term representing all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin(45).
Council of Europe's Committee of Ministers	2005	The term *hate speech* shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin. In this sense, hate speech covers comments which are necessarily directed against a person or a particular group of people.
ILGA-Europe	2010	Hate Speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups.
European Commission	2016	"Code of conduct on countering illegal hate speech online" to help users to flag illegal hate speech in these social platforms (Facebook, Microsoft, Twitter and YouTube), improve the civil discourse, and increase coordination with national authorities.

Table 2: Comparison of hate speech definitions across social media: Facebook, Twitter and YouTube.

Source	Definition
Facebook	We define hate speech as a directed attack on people based on what we call protected characteristics - race, ethnicity, religious affiliation, sexual orientation, sex, gender, gender identity and serious disability or disease.
Twitter	Users may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. The consequences for violating the Twitter rules vary depending on the severity of the violation. The sanctions span from asking someone to remove the offending Tweet before they can Tweet again to suspending an account.
YouTube	We encourage free speech and try to defend your right to express unpopular points of view, but we don't permit hate speech. Hate speech refers to content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on specific attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but if the primary purpose of the content is to incite hatred against a group of people solely based on their ethnicity, or if the content promotes violence based on any of these core attributes, like religion, it violates our policy.

- The presence of a common feature shared by the group, such as race, religion, ethnicity, nationality, sexual orientation or any other similar common factor that is fundamental to the identity.
- Hate speech, as a concept, refers to a whole spectrum of negative discourse, stretching from expressing, inciting or promoting hatred, to abusive expression and vilification, and arguably also to extreme forms of prejudice, stereotypes, and bias.
- The consequences arising from hate speech include disturbing public peace and order or inciting violence. Examples are incidents between groups in society, as well as hate crimes towards people previously targeted with online hate speech.

The definition of hate speech that this study adopts as the knowledge base is the following: Hate speech is a kind of expression designed to promote hatred by race, religion, ethnicity, national origin, gender, sexual orientation, social origin, physical or mental disability.

Table 3: Hate Speech and related terms. All the definitions were taken from Mish and Morse (Mish and Morse).

Source	Definition	Comparison with hate speech
Hate	The feeling of aversion for or extreme hostility toward a target without stated explanation for it.	Hate is a general expression of hatred, while hate speech has specific targets towards whom one addresses offensive content.
Cyberbullying	The electronic posting of mean-spirited messages about a person. Often done anonymously.	Hate speech does not include verbal attacks towards specific individuals. It is typically addressed towards a group of people or a member of a community. Personal attacks are not included in the definition.
Discrimination	Prejudiced or prejudicial outlook, action, or unfair treatment	Hate speech takes place only through verbal means.
Abusive language	The use of harsh, insulting language. It can include hate speech, derogatory language and also profanity.	Hate speech employs abusive language.
Profanity	Blasphemous or obscene language.	Hate speech can use profanity, but not necessarily.
Toxic language	Toxic use of the language is a synonym of aggressive language, used to hurt. It is rude and disrespectful and leads the interlocutors to leave the conversation.	Hate speech can be toxic, however, it is also able to trigger more discussion over a topic.
Harassment	The act of systematic and continued unwanted and annoying actions of one party or a group, including threats and demands. The purposes may vary, including racial prejudice, personal malice.	Hate speech does not include in its definition a temporal aspect.

2.2 LITERATURE REVIEW ON HATE SPEECH DETECTION

This section is organized in the following way: first, we present the work that has been done with hate speech detection in Italian, then we look at how hate speech detection has been performed in English.

This literature review is organized systematically to define the features used in our machine learning hate speech detection algorithms: hate speech-specific features, and features that are used more generally in text classification.

Furthermore, it is crucial for our research to have an overview of the previous work based on data gathered via distant supervision. Therefore, we have also included a section that deals with the impact on text classification of this particular methodology of data collection.

Previous work on Hate Speech in Italian

In March 2018 Armando Cristofori, the World Speech Day ambassador, stated that no other European country, and likewise few other countries in the world, are showing such a growing presence of hate speech in social media

as Italy (Ansa, 2018). He also added that hate speech could be found in most threads, from politics to sports, showing hidden divisions within the country which could lead to bad turnouts. With this premises, hate speech has become progressively a matter of interest for Italian researchers in recent years.
An overview of the publication on Hate Speech in Italian is summarized in Table 4.

Table 4: Previous research on hate speech in Italian.

Year	Source	Human annotation	Topics	Type of research	Features	Paper
2018	Twitter	yes	Immigration	bibliography	-	Sanguinetti
2017	news	yes	-	bibliography	-	Bosco
2017	Facebook	yes	Immigration	statistical	Lexical, Morpho syntactic, Lexicon	Vigna
2017	Twitter	yes	Immigration	bibliography	-	Poletto
2016	Twitter	yes	Homophobia, Violence, racism, disability, Anti-semitism	statistical	sentiment analysis	Musto

The papers Poletto et al. (2017), Bosco et al. (2017) and Sanguinetti et al. (2018) discussed tools and resources that can be used in text classification to accomplish the task of detecting Hate Speech. They introduced essential annotation metrics and approaches to studying hate speech, but they have not yet made available their developed resources.

Musto et al. (2016) and Del Vigna12 et al. (2017) ran machine learning experiments to classify social media content to automatically assign the labels *Hate* or *Not hate*. However, Musto et al. (2016) did not provide details on the classifier nor reference to the results. The study aimed to find hateful tweets during a particular span of time and geolocalized them.

The research presented by Del Vigna12 et al. (2017), on the other hand, is the publication that most influenced our approach as they developed a corpus of annotated Facebook data addressed against the communities of Roma and immigrants.
The study reported struggles with annotating the dataset. They reached a poor (0.19) inter-annotator agreement and, therefore, they had to repeat the experiment with a smaller set of annotators and a reduced number of classes: *Hate* and *Not Hate*.

Del Vigna12 et al. (2017) confirmed the difficulties with annotating content according to the field of hate speech detection, as mentioned in Duarte et al. (2017), a report written for policymakers and researchers on how to study online hate speech.

The study presented two approaches to text classification, using both a machine learning and a neural network approach. The machine learning classifier was built on a combination of morphosyntactic features, sentiment polarity, and word embeddings. We can investigate the results obtained in the study in Table 5.

Table 5: Results from *Hate me, hate me not: Hate speech detection on Facebook.*

		Hate			Not Hate		
Algorithm	**Acc**	**P**	**R**	**F**	**P**	**R**	**F**
SVM	80.60	.757	.689	.718	.833	.872	.851
LSTM	79.81	.706	.758	.728	.859	.822	.838

Previous work on Hate Speech in English

The next section discusses how research in hate speech detection has been addressed outside the sphere of the Italian language.

2.2.1 Data Collection

The first step towards hate speech detection is data collection. Research, which does not employ publicly available datasets, can gather data from websites or social media.

On one hand, social media sites are repositories with large quantities of data. On the other, this content is noisy, multimodal and controversial to annotate, especially when conducting studies on hate speech (Duarte et al., 2017).

The process of data collection varies, not only according to which social media is chosen to investigate, but also to the modalities of data extraction. A recurring approach when collecting data for hate speech detection is the use of a lexicon of words that are considered hateful (Davidson et al., 2017; Waseem and Hovy, 2016; Burnap and Williams, 2016; Magu et al., 2017). Regular expressions (Magu et al., 2017) are also used as techniques to retrieve particular data from users known to have previously shared hate speech (Kwok and Wang, 2013).

The lexicon-based approach, however, suffers from shortcomings because it considers only tweets or comments when particular keywords are present, leading to an oversimplification of online content.
Nevertheless, the use of a lexicon functioned as the starting point for other types of data extraction as well. Waseem and Hovy (2016) expanded the vocabulary with co-occurring terms, improving the search for hateful content. Additionally, Ribeiro et al. (2018) used a lexicon of offensive words to identify Twitter haters and map them with their followers. The study aimed to look at hateful users rather than the content. However, this approach is not scalable to other social media platforms where one does not have the option to access the user's network of friends.

2.2.2 Annotation

In machine learning, research can adopt different approaches, such as supervised or unsupervised learning techniques and their semi-supervised and semi unsupervised variations.

Supervised learning approaches, which are the most popular choice to study the presence of hate speech in social media (Duarte et al., 2017), require to annotate the input data.
In the literature concerning hate speech, we found different types of manual annotation. The variation took place according to the scale and the budget of the study.

The labeling process would often be a task for the researchers involved in the studies (Waseem and Hovy, 2016; Kwok and Wang, 2013; Poletto et al., 2017), which had the added convenience of using expert annotators. Otherwise, external annotators (Warner and Hirschberg, 2012; Gitari et al., 2015) or crowd-sourcing services (Burnap and Williams, 2016) were employed to label the corpus.

2.2.3 Features

In this section, we summarize the main features used to tune the classifiers developed to detect hate speech. We divide the features into two parts, first the features that are generally applicable to text classification, such as word n-grams and part-of-speech (POS) tagging. On the other, we gathered features that were distinctively thought to cater to hate speech.

Let's first address the general text classification feature, which is mostly comprised of content-related features.

- n-grams, Baf of words (BoWs) (Waseem and Hovy, 2016; Kwok and Wang, 2013; Gitari et al., 2015; Greevy and Smeaton, 2004)
- word embeddings such as paragraph2vec (Park and Fung, 2017), GloVe (Pennington et al., 2014) and FastText (Badjatiya et al., 2017).
- POS tagging, ease of reading measures (Davidson et al., 2017; Burnap and Williams, 2016; Gitari et al., 2015; Warner and Hirschberg, 2012)
- other types of features include (i) attributes related to the user's activity, network centrality and the material he or she produced in our characterization and detection. (Chatzakou et al., 2017; Ribeiro et al., 2018) (ii) using the gender and the location of the creator of the content (Waseem and Hovy, 2016).
- sentiment analysis, topic modeling, semantic analysis (Agarwal and Sureka, 2017; Gitari et al., 2015).

Specific features were adopted in the previous research to tackle the challenge of detecting hate speech.

- **Othering Language** the expressions that create a marked division between two sides, "us vs. them". Typically, the side that recognizes

itself in *us* perceives to the superior part. Consequently, *them* is the weaker and subordinate part (Dashti et al., 2015). In the datasets that we employed, we saw several cases of othering language, both when considering the topics of immigration, veganism, and homosexuality. Haters tented to place a distance between themselves and the target of their hate. "immigrants have to go to their home', "take away children from their family. They are not real Italians", and, lastly, "We are the traditional family, you are not" are the translation of few instances that we found in our datasets.

- **Declarations of superiority** A more in-depth look at the relationship between superior and subordinate groups shows that declarations of superiority can also be considered hate speech. In this case, hate speech can assume the shape of defensive statements and disclosures of pride, rather than attacks directed toward a specific group (Warner and Hirschberg, 2012).

- **Stereotypes** The targets are often communities which share common traits and popular stereotypes. Warner and Hirschberg (2012) concentrated on the offenses towards such groups, detecting the expression used to address the stereotypes. Words, phrases, metaphors, and concepts around stereotypes are repetitive, and they can be considered the indicator of hate speech.

- **Perpetrator Characteristics** Studies connect the use of hatred with the user's personal characteristics, such as gender, age, geographical localization and ethnicity (Waseem and Hovy, 2016). Therefore, people's profiling can be used as additional clues when performing hate speech detection.

2.2.4 Models

The fourth step, after collecting, annotating the datasets and designing the feature engineering is the development of the classifier. The literature includes different approaches to tackle the difficult challenge of hate speech detection. The large majority of the models previously built follow the supervised learning approach: Naive-Bayes (Kwok and Wang, 2013), Logistic Regression (Waseem and Hovy, 2016; Davidson et al., 2017). Support Vector Machines (Warner and Hirschberg, 2012; Burnap and Williams, 2016; Magu et al., 2017; Badjatiya et al., 2017; Greevy, 2004; Davidson et al., 2017). Rule-Based Classifiers (Gitari et al., 2015), Random Forests (Burnap and Williams, 2016), GradientBoosted Decision Trees (Badjatiya et al., 2017) and Deep Neural Networks (Badjatiya et al., 2017; Pitsilis et al., 2018).

The results of the classifier per model are the following:

Table 6: Report on the most used models in hate speech detection and their corresponding F-score results.

Year	F1	Algorithm	Research
2013	.76	Naive-Bayes	(Kwok and Wang, 2013)
2016	.91	Deep Neural Networks	(Yuan et al., 2016)
2016	.73	Logistic Regression	(Waseem and Hovy, 2016)
2017	.90	Logistic Regression	(Davidson et al., 2017)
2012	.63	Support Vector Machines	(Warner and Hirschberg, 2012)
2016	.77	Random Forests, Support Vector Machines	(Burnap and Williams, 2016)
2017	.79	Support Vector Machines	(Magu et al., 2017)
2017	.93	Deep Neural Networks, GradientBoosted Decision Trees	(Badjatiya et al., 2017)
2015	.69	Rule-Based Classifiers	(Gitari et al., 2015)
2018	.88	Recurrent Neural Networks	(Pitsilis et al., 2018)

Previous work on text classification and distant supervision

In machine learning, supervised and unsupervised learning are the main adopted paradigms. The former requires that both input and output data are labeled so that the classifier learns how to map and predict from them. The latter, its unsupervised counterpart, considers unlabeled data with the aim of allocating it into labeled groups, according to shared patterns.

When considering the input data, both approaches suffer from limitations. The disadvantages of production of supervised learning lie in the time and resources needed to develop manually labeled training data. Unsupervised approaches can handle large amounts of data and extract as many numbers of relations. However, the lack of prior knowledge makes the results of the analysis impossible to be ascertained (Jurafsky and Martin, 2014).

Introduced as a new take on data annotation (Mintz et al., 2009; Go et al., 2009), *distant supervision* is used to automatically assign labels based on the presence or absence of specific hints, such as happy/sad emoticons (Go et al., 2009) to proxy positive/negative labels for sentiment analysis, Facebook reactions (Pool and Nissim, 2016; Basile et al., 2017) for emotion detection, or specific strings to assign gender (Emmery et al., 2017). In this research, we refer to data labeled via distant supervision as silver data, as opposed to gold, manually labeled data.

Such an approach has the advantage of being more scalable and versatile than pure supervised learning algorithms while preserving competitive performance. Better portability features distant supervision to different languages or domains, and it does not require extensive time and resources needed to train.

Apart from the ease of generating labeled data, distant supervision has a

valuable ecological aspect in not relying on third-party annotators to interpret the data (Purver and Battersby, 2012).

Moreover, distant supervision reduces the risk of adding extra bias, since it does not over manipulate the natural data. Go et al. (2009) also showed that machine learning algorithms (Naive Bayes, Maximum Entropy, and Support Vector Machine), trained on distantly supervised data could reach an accuracy of above 80%.

An interesting study on the infusion of portions of manually labeled data into distantly supervised data is presented in Pershina et al. (2014), whose approach achieved a statistically significant increase of 13.5% in F-score and 37% in the area under the precision-recall curve.

2.3 DIFFICULTIES IN DETECTING HATE SPEECH

In this section, we highlight different aspects that make the task of automatically detecting hate speech online difficult. First, we draw attention to the fact that there is no commonly accepted definition of the term *hate speech*, and secondly, we describe the types of data that are often mistaken for hate speech. Other possible limitations to the success of the task that we found in the literature review are the following:

- Annotators reach a very low agreement (33%) in hate speech classification (Kwok and Wang, 2013), demonstrating that this task would be harder for machines. (Del Vigna12 et al., 2017) had to run a second set of experiments due to the lack of a sufficient inter-annotator agreement reached during the first iteration.
- The difficulty at annotating content with the binary labels *hate – not hate* relies on the fact that the annotators should have a common cultural and social background (Raisi and Huang, 2016).
- Hate speech detection needs more appropriate means than a keyword look-up to be found.
- Hate speech is a longitudinal phenomenon which evolves with the language development. Its detection can be tricky when it comes to identifying offensive language against minorities and youngsters' new ways of communication (Nobata et al., 2016). Therefore, social media content is particularly interested by socio-linguistic phenomena (Raisi and Huang, 2016).
- Hate speech manifests in offensive and abusive language. If the offensive language can be associated with ungrammatical forms, the abusive expressions can be fluent, grammatically correct and mixed with sarcasm (Nobata et al., 2016).
- The progressive changing of policies and new restrictions on data collection are also affecting social media studies. Application programming interfaces (APIs), at the moment, allow registered user to create private applications and download public data. These progressively

tightening restrictions have a significant impact on this type of research.

3 DATA

The exploration of the literature review on hate speech detection highlighted two main issues that grounded our approach in this thesis.
First, we found only one study that dealt with the problem of automatic hate speech detection in Italian (Del Vigna12 et al., 2017) and second, we noticed the lack of resources to study hate speech in Italian.

Through the course of this study, we obtained a few, small, annotated datasets to perform hate speech detection on. However, when we began the project, we had no Italian datasets. Because we wanted our research to focus on Italian, we decided to develop our own annotated data set that suited the purposes of our supervised learning task.

In this chapter, we aim at explaining our take on distant supervision, focusing on the process that we used to gather and annotate data. Secondly, the large part of the chapter is used to clarify the distinctions between the datasets that we used for training and testing our classifier.

3.1 INTRODUCTION TO THE DATASETS

The following is an overview of the several data sets that we used, organized according to two criteria: source and target of hatred.

3.1.1 Dataset organized according to the source parameter

We scraped data from two social media sites, Facebook and Twitter, as well as from the video platform YouTube. The following visualization shows the data organized by quantity and source.

Figure 1 is a representation of all the datasets that we employed to train and test the classifier summed according to their source.
The picture shows that we downloaded most of the data from Facebook and YouTube. The choice to ground our research on these two platforms is due to the concentration of Twitter-based datasets in previous studies. As demonstrated in the literature review, most of the work on hate speech detection in Italian used datasets created from tweets (Sanguinetti et al., 2018; Musto et al., 2016; Poletto et al., 2017).

Facebook, YouTube and Twitter are the sources of the seven datasets we employed throughout the experiments. An overview of the battery of resources can be found in Table 7.

The Facebook dataset is composed of four subsections:

a) A manually labeled dataset provided in the context of the EVALITA 2018 task on Hate Speech Detection (`haspeede`).

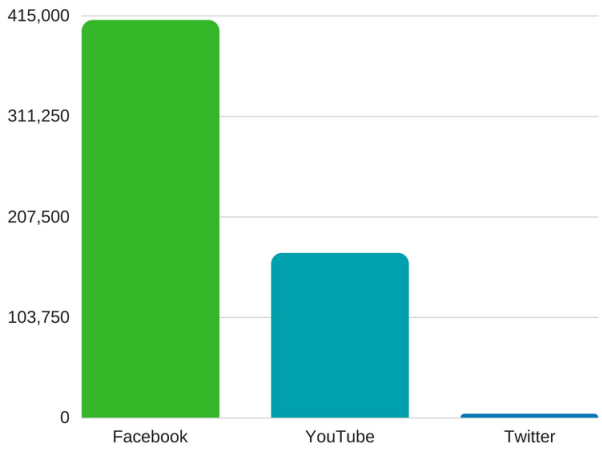


Figure 1: Summary of the datasets by source.

Table 7: Overview of the datasets according to its annotation type, usage and size in comments

Social media	Gold data	Silver data	Used for embeddings	Quantity
Facebook EVALITA 2018	yes	-	-	3,000
Facebook multi target	-	yes	yes	100,000
Facebook singe target	-	yes	-	189676
PSP	yes	-	-	12,153
Twitter EVALITA 2018	yes	-	-	3,000
Twitter Turin University	yes	-	-	990
YouTube	-	yes	yes	170,000

b) Two distantly supervised datasets gathered from specific Facebook pages according to previously determined proxies.

c) A dataset of social media messages manually annotated for offensive language and hate speech, the Political Speech Project (Bröckling et al., 2018). We will refer to this extra dataset henceforth as `PSP` dataset.

YouTube is the second largest dataset that we employ. It consists of a single dataset that we annotated using distant supervision.

We had two small Twitter datasets. First, a sample of 3000 tweets, obtained from taking part in EVALITA 2018 task on Hate Speech Detection (`haspeede`). Secondly, a small dataset of 990 Tweets that the University of Turin made freely available.

3.1.2 Dataset organized according to the Target of the hatred

The main target areas we covered are the following:

- hatred against immigrants
- hatred against women
- hatred against vegan people
- hatred against the LGBTQ community
- hatred against politicians

We studied the phenomenon of hate speech in two parallel ways: first, looking at how a triggering event catalyzes the creation of hate speech, and second, how hatred is holistically present in social media sites.
We made a distinction between these areas by creating two datasets. *Mattarella corpus* contains hatred against a single target, politicians. *Facebook multi-target corpus* addresses offensive content towards a variety of targets.
The limitations of using such noisy and automatically labeled datasets could result in a poor performance of the machine learning model, with particular struggles when recognizing hate speech across different types of targets.

3.2 DISTANTLY SUPERVISED DATASETS – SILVER DATA

Distant supervision is a method of annotating data that combines the advantages of bootstrapping with supervised learning (Mintz et al., 2009). Bootstrapping is designed to use as few training examples as possible. It first takes a small set of training examples, trains a classifier, and finally uses thought-to-be positive examples for retraining (Biemann, 2007). At the beginning of this project we did not have any small training sets to bootstrap, so instead, we fully employed the method of distant supervision.

The distant supervision approach is based on acquiring a large number of seed examples and automatically assigning labels based on the presence

or absence of specific proxies, such as emoticons (Go et al., 2009) and gender bias elements (Kiritchenko and Mohammad, 2018), or any other criteria that researchers believe is distinguishing.
Apart from the ease of gathering and annotating data, distant supervision has the convenience of not relying on third-party annotators to interpret the data (Purver and Battersby, 2012). The distant supervision approach has an advantage when creating a corpus for hate speech since previous work showed difficulties in reaching a satisfying inter-annotator agreement (Kwok and Wang, 2013).
However, the problem with distant supervision is that the labels are not gold standard and that they may be ambiguous or possibly even wrong. To reduce these inherent errors, we trained our classifier on both gold and distantly supervised data. We compared the two performances and verified their effectiveness.

We propose a unique take on distant supervision. We use the *sources, where the content is published online, as proxies, rather than gathering any hint of the label through the content itself*. For example, we scraped content from the Facebook pages of politicians known to have strict positions against particular topics, such as immigration. Their social media pages are a source of hateful posts and comments that we systematically collected over time.
The dataset generated via distant supervision is very versatile. We use the large corpus for both classification purposes and for the generation of polarized word embeddings to be used as features.

We developed three datasets via distant supervision:

a) The first dataset is a set of Facebook comments downloaded between May 27th and the 28th 2018. We chose this 48 hour window because the Italian people used social media as a medium to attack politicians. We called the political dataset the `Mattarella corpus`.

b) Second, we built another Facebook-based corpus, which addresses hateful content towards different communities of people. We call the resource `Facebook multi-target corpus`.

c) We created the third dataset from YouTube comments.

3.2.1 Single target data - Mattarella corpus

The 48-hour window between May 28, 2018, and May 27, 2018, was affected by an abnormal presence of online hate speech. The last days of May represented the final stage of the formation of the new Italian government. The running parties proposed a list of ministry members to the Italian President of the Republic, Sergio Mattarella, who did not accept one of the proposed members. The decision of the President of the Italian Republic created a governmental crisis and stagnation. Many newspapers, such as Il Corriere Della Sera (Breda, 2018), Il Giornale (Scafi, 2018), La Stampa (Minucci, 2018), and Il Fatto quotidiano (F.Q., 2018), reported that the online pages of the President of the Republic was experiencing a wave of hate speech . The postal police also discovered a considerable amount of hate speech directed

towards politicians, and they arrested a few people who threatened the life of the President of the Republic.

We systematically gathered the data from different news sources published within two days of the official speech held by the President of the Republic. We completed this data collection following Facebook's API terms of service and obtained 225,010 comments and 3,775,024 tokens (Table 8).
We also noticed a heavy use of hateful words and expressions while reading samples of the newly created dataset. For example, in the social media page of the newspaper *Il Corriere Della Sera*, the first comments to the speech held by the Italian President of the Republic were the following:

"Un altro stronzo che esercitato il diritto di propietà." - [Another asshole that exploited his right of property.]

"L'emerito ennesimo cretino, scarto di civile società." - [The emeritus piece of crap, garbage of the society.]

"Mettetegli una divisa del terzo Reich ed è perfetto." - [Put the third Reich uniform on him and it is perfect.]

Table 8: Comments extracted from May 28 to May 30th

Source	**Amount**
La Repubblica	70,024
Il Giornale	17,667
Il Corriere della Sera	35,163
Agenzia Nazionale Stampa Associata (ANSA)	14459
Il Manifesto	162
Il Fatto Quotidiano	78,222
La Stampa	6103
	225,010

3.2.2 Facebook multi-target dataset

To gather a dataset based on keywords, we selected a set of publicly available Facebook pages that had a good chance of promoting or being the target of hate speech, such as pages known for promoting nationalism (*Italia Patria Mia*), controversies (*Dagospia*, *La Zanzara - Radio 24*), hate against migrants and other minorities (*La Fabbrica Del Degrado*, *Il Redpillatore*, *Cloroformio*), and support for women and LGBT rights (*NON UNA DI MENO*, *LGBT News Italia*). In this latter case, we expected a plethora of both instigators and haters.
Using the Facebook API [1], we downloaded the comments from the posts present in these pages, as they are the text portions that are most likely to express hate. We collected over 1 million Facebook comments and almost 13 million tokens. The source and quantity of data that was extracted is reported in Table 9.

[1] https://developers.facebook.com/

Table 9: List of public pages from Facebook and number of extracted comments per page.

Source	Amount
Matteo Salvini	318,585
NON UNA DI MENO	5,081
LGBT News Italia	10,296
Italia Patria Mia	4,495
Dagospia	41,382
La Fabbrica Del Degrado	6,437
Boom. Friendzoned.	85,132
Cloroformio	392,828
Il Redpillatore	6,291
Sesso Droga e Pastorizia	8,576
PSDM	44,242
Cara, sei femminista - Returned	830
Se solo avrei studiato	38,001
La Zanzara - Radio 24	215,402
	1,177,578

From this large amount of data, we extracted 100000 random comments to be used as training data. We mirrored the proportion of the labels in the `haspeede` dataset. 54% of it was composed of hateful data that we scraped from sources which post and share hatred against specific communities and 46% of the total was gathered from a neautral source, the social media page of Agenzia Nazionale Stampa Associata (ANSA).

Being automatically annotated data, we do not know if the labels correctly represent the content, and consequently, if the distribution is kept in the desired proportions.

3.2.3 YouTube data set

The second platform that we included in the research is YouTube. The comments of a YouTube video can be reached via the YouTube API, by using the YouTube Comment Scraper project [2]. Given a YouTube video URL the user can request all comments for that video from the API. Therefore, we did not employ the source as a distinguishing proxy.

We decided what videos to focus on based on the findings of our research on Facebook data, where we noticed recurring targets of hatred. We narrowed down the topics that we thought to be heavily targeted by hate speech: women, the LGBT community, vegans and popular politicians (Table 11).

For YouTube, we created a control group out of comments scraped from popular music videos of Italian hits (Table 10).
We used keywords to find related videos on YouTube, and we downloaded the comments using the YouTube Comment Scraper. We set the distribution

[2] https://github.com/philbot9/youtube-comment-scraper

of the dataset to 54% not hateful and 46% hateful, based on the dataset provided by EVALITA 2018 haspeede.

Table 10: Not hateful comments from YouTube.

Artist	Song Title	Amount
Alessandra amoroso	Comunque andare	9,423
Fedez	Magnifico	9,643
	Vorrei ma non posto	43,929
Giorgia	Come la neve	2,885
	Credo	2,716
Marco Mengoni	Ti ho voluto veramente bene	9,015
	Guerriero	8,176
	L'essenziale	11,488
Vasco Rossi	Come nelle favole	3,762
		101,037

Table 11: Hateful comments from YouTube.

Theme	Source	Topic	Amount
women	Fanpage.it	Tiziana Cantone's funeral	397
	gli autogol	Diletta Leotta after foto leak	1,117
	rai	Belen Rodriguez goes to court	571
	la7	Interview with Selvaggia Lucarelli	86
	great menchi	Blogger has face plastic surgery	5,026
	cittadinapoli.com	Berlusconi calls Belen Rodriguez	1,373
	redazionenews	Interview with Miss Italia	442
	la7	Interview with Matteo Salvini and Laura Boldrini	4,088
life style	fanpage.it	Documentary on Fruitarians	1,267
	lambrenedettoxvi	Comparison fruitarians and carnivors	8,130
	rai	Interview with fruitarians	2,283
	rai	Interview with vegan family	2,821
	viavai	Interview with a vegan and a omnivore	9,507
immigration	fanpage.it	Documentary on immigrants in Italy	7,366
	fanpage.it	Castel Volturno immigrants' protest	1,446
	la7	Interview with Matteo Salvini and Cécile Kyenge	4,728
	luigi magenta	Cécile Kyenge goes buys expensive clothes	101
	funpage.it	In Milan restaurant against immigrants	1,076
	fanpage.it	Roberto Saviano debunks the myths around immigration	2,111
	matteo salvini	Matteo Salvini on immigration	3,654
	la7	Roman citizens vs the local Muslims	1,898
	la7	Documentary on Muslim women	1,326
	la7	Documentary of arranged marriages in Syria	1,215
			61,029

3.3 AVAILABLE DATASETS – GOLD DATA

3.3.1 PSP

The PSP dataset is part of a journalistic initiative to chart the quality of online political discourse in the EU. Almost 40 thousand Facebook comments and tweets between February 21 and March 21, 2018, were collected and manually annotated by an international team of journalists from four countries (France, Italy, Germany, and Switzerland). The original data set is organized as follows:

- **Language** : French, German, Italian, Swiss
- **Rating**: 0 - neutral, 1 - mildly offensive, 2 - offensive, 3 - highly offensive
- **Category**: Sexist, Anti-immigrant, Anti-muslim, Anti-semitic, Homophobic, Other, None
- **isPersonal**: The label **No** was assigned if the rating was zero, whereas **Yes** was used to indicate personally offensive content being addressed to politicians.

We extracted the data that reported *Italy* as the language label. In total, this section had 12,153 instances of Italian Facebook comments with a total of 27,601 tokens.
The label convention was normalized according to the need of this study. Our classifier makes a binary decision by assigning the labels *hate* or *not hate* to input text data. For this reason, we converted *1 - mildly offensive, 2 - offensive, 3 - highly offensive* to the label *hate*. Then, we assigned the label *not hate* when the original label was found to be *0*. We studied the presence of hateful content in this dataset. We also gathered the distribution of the hate across the target of the hatred messages. The findings are reported in Table 12.

Table 12: Distribution of labels for the PSP dataset.

Labels	Samples
not hateful	11,283
hateful	870

Among this section, the majority of the data was found not to be hateful. The annotators identified 7% of the data to be hate speech. This conclusion is in line with our expectations, as the data was scraped from a newspaper social media page, which is not as controversial as the personal pages of politicians (Kong et al., 2018).

3.3.2 Facebook and Twitter EVALITA 2018 datasets

The haspeede dataset consists of two subsets of data that are 3,000 instances each. The sources of the subsets are Facebook and Twitter. The distribution of hate within the datasets can be found in Table 14.

Table 13: Hate distribution across EVALITA 2018 data

Facebook	**Samples**
not hateful	1,618
hateful	1,382
Twitter	**Samples**
not hateful	2,029
hateful	971

3.3.3 Twitter corpus

Turin dataset is a collection of 990 manually labeled tweets concerning the topic of immigration, religion and the Roma community (Poletto et al., 2017). The distribution of labels in this dataset differs from the EVALITA 2018 dataset, with only 160 (16%) hateful instances.

Table 14: Hate distribution across the Turin dataset

Twitter	**Samples**
not hateful	830
hateful	160

3.4 PROCESSING

We applied minimal pre-processing to both the gold and silver data. For the gold data we normalized the following:

- we substituted the reference to users via the sign @name with @username
- we converted URLs to the string 'URL.'
- lowering the case of all the characters
- removal of Italian stop words

The intention was to preserve as much lexical information as possible, even if it contained grammatical errors.

3.5 MERGING DISTANT SUPERVISION AND ANNOTATED DATA

In the first chapter, we introduced hate speech detection as a complex task. We showed that the definition of hate speech is not clearly defined, and we presented the difficulties in developing a dataset that can be representative of the hateful content in real data.

Due to the lack of previous resources in Italian for such studies, we first gathered silver data via distant supervision. We assigned the *hate* label to content that we scraped explicitly from sources that promote hatred. A similar process was adopted for not offensive content. We created two datasets that allowed us to run machine learning experiments and create semantic tools quickly and cheaply. Also during the development of this project, we obtained a series of manually annotated gold datasets from shared tasks (`haspedee`) and study groups (`PSP`, `Twitter corpus`).

The literature proves that hate speech detection is particularly effective when the classifier is trained on manually annotated gold data. Therefore, we decided to train our classifier on a dataset that merged the two datasets we had available: silver and gold data.

Our hypothesis, supported by the research conducted by Pershina et al. (2014), is that not only would merging the two datasets increase the performance of the results obtained on gold data alone, but it would also create a dataset more in line with the guidelines provided by Duarte et al. (2017).

3.6 ADDITIONAL RESOURCES: WORD EMBEDDINGS

Distant supervision allows the development of semantic and distributive tools that require a large input dataset. We opted for the creation of word embeddings that could polarize the performance of the classifier towards the direction of the offensive language.

Word embeddings are dense and distributed representations which aim at enriching the semantic information. The linguistic theory behind the approach, namely the "distributional hypothesis" by Harris (1954), summarizes this concept with the following definition: *Words that have similar context will have e similar meanings*.

Mikolov et al. (2013) defined the skip-gram model to train word embeddings by maximizing the probabilities of words given their context windows. The approach offers word probability according to the embedding probability that condenses context information with the definition of a target word and a context word, of the same word. The probability of a target word is estimated by the cosine similarities between the target embedding and the content embeddings of its context words.

To semantically enrich the limited available annotated data and support the one gathered via distant supervision, we decided to use this vector based tool. We employed different types of embeddings in the research (Table 15), which can be differentiated into two kinds: pre-trained, off-the-shelf embed-

dings and newly trained embeddings over the broad set of data downloaded and labeled with distant supervision.

Table 15: Overview of the word embeddings used for the experiments

Source	Dimensions	N vocabulary
Twitter	52	2,196,954
Retrofitted	52	419,084
Hate oriented Facebook	300	381,697
Hate oriented YouTube	300	282,384
Merged Hate embeddings Facebook and Twitter embeddings	300	2,552,460

We use several embeddings made from different types of sources and dimensions, with the idea that such characteristics would influence the outcomes of our classification. We obtained Twitter embeddings from the website SpinningBytes [3], which made available embeddings with the dimensions of 300 by 52 trained on Twitter data. We developed polarized Facebook and YouTube embeddings, and we modified the existing Twitter embeddings, by merging them with our semantic tools or retrofitting them with the aid of a lexicon. We describe the development of the word embeddings in the chapter *Model*.

Coverage

We used different types of datasets to train and test our classifier. To check whether embeddings could be predictive, we had to consider the influence of semantic tools on the used datasets. A parameter that can be incisive on the results is the number of words shared by both the embeddings and the datasets. This parameter is called coverage, and we calculated it across the datasets that we used in the research (Table 16). We demonstrate that the Twitter embeddings has the broadest coverage among the other embeddings that we used.

Table 16: Word coverage: the number of tokens shared by datsets and word embeddings

Dataset	Twitter	Retrofitted	Facebook Hate	YouTube Hate	PCA
FB Gold News	18,130 (0.66 %)	16,416 (0.59 %)	7478 (0.63 %)	8,932 (0.32 %)	19,298 (0.7 %)
Mattarella	20,332 (0.75 %)	18,449 (0.68 %)	20819 (0.76%)	9,847 (0.36 %)	2,2110 (0.81 %)
EVALITA 2018	11,287 (0.68%)	10,485 (0.63%)	11,149 (0.67%)	6,396 (0.38%)	11,963 (0.72%)
Turin	3,964 (0.97%)	3,791 (0.67%)	3,879 (0.69%)	2,516 (0.45%)	4,129 (0.73%)
Facebook Silver sources	89,075 (0.62%)	66,101 (0.46%)	103,406 (0.72%)	16,305 (0.11%)	104,700 (0.73%)
YouTube	53,991 (0.55%)	44,764 (0.46%)	49,177 (0.51%)	17,409 (0.18%)	60,227 (0.62%)

[3] http://www.spinningbytes.com/resources/wordembeddings/

4 MODEL

In the previous chapter we presented our take on distant supervision and focused on the methodology we used to gather and annotate data.
We first downloaded a large amount of data by using the Facebook [1] and YouTube APIs. Secondly, we automatically assigned labels according to the source of the data.

Additionally, we went through each dataset that we planned to use during the classification task to both train the system designated to detect hate speech and test it. We created a number of datasets with different sources, compositions, sizes, and distributions.

The next section describes the definition of the machine learning model we used and our work on feature engineering.

4.1 MODEL

The task of hate speech detection using machine learning has previously been accomplished using rule-based methods or supervised classifiers. Rule-based methods (De Marneffe and Manning, 2008; Mondal et al., 2017; Pelosi et al., 2017; Xu and Zhu, 2010; Su et al., 2017; Palmer et al., 2017) heavily rely on lexical resources such as dictionaries, thesauri, sentiment lexicons, as well as syntactic patterns and POS relations.

Supervised approaches have shown to obtain good results, although they suffer from limitations as far as the size and domain of the training data is concerned. Support Vector Machine (SVM) and Convolutional Neural Network (CNN) classifiers turned out to be efficient algorithms for this task. A successful example of the SVM model is the system with word embeddings proposed by Del Vigna12 et al. (2017) and Term Frequency–Inverse Document Frequency (TF-IDF) n-grams present in Davidson et al. (2017), which showed competitive performances for this approach.

We also adopted a supervised learning approach to tackle hate speech detection. Supervised learning involves the presence of labeled data for both input and output variables (Jurafsky and Martin, 2014). Our take on distant supervision gave us an approximate dataset to quickly feed into the classifier, satisfying the requirements of supervised learning (Figure 2). The ultimate goal of the system is to learn from the information provided in the training data and approximate the found patterns to predict the output variables. In our case, the binary labels were *hateful not hateful*. represents the work-flow that our supervised learning algorithm followed to reach the final predictions.

We built a system to perform a binary task, made by a linear SVC model with unbalanced class weights using various linguistic features. We imple-

[1] https://developers.facebook.com/docs/graph-api

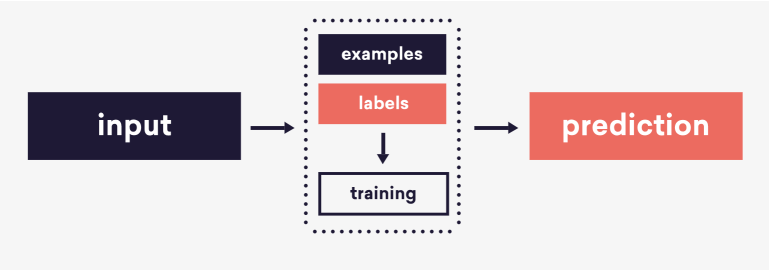


Figure 2: Work-flow of our supervised learning system.

mented the system using the Scikit-Learn Python toolkit (Pedregosa et al., 2011) using default values for the other hyper-parameters. We adopted this model for the size of the distantly supervised datasets and their unbalanced labels distribution.

4.2 FEATURES

Lexicon look up

We used two groups of surface features, namely: i.) unigrams and bigrams, and ii.) character n-grams in the range between 3 and 6. Additionally, we utilized a lexical surface feature to retrieve patterns at the word-level.

We expected that haters would address offensive language towards communities of people by using stereotypes. We hoped to detect such modular and repetitive expressions using a vocabulary of negative and hateful words. To do so, we generated the lexicon from two online resources: the article *words that hurt* written by the linguist De Mauro (2016) and a list of vulgar words available in Wikipedia [2].

The Italian linguist organized the thesaurus in 13 different groups, according to the target of the hatred (Table 17). In total, the article classified 195 words. This first lexical resource addressed the deeper level of hate against communities; however, it did not include indecent and offensive words, which we integrated using the lexicon found in Wikipedia.

[2] https://it.wiktionary.org/wiki/Categoria:Parole_volgari-IT

Table 17: Lexicon extracted from Tullio De Mauro article:

Type	Example	Translation
negative stereotypes	americanata	big/superficial thing
italian regional names	genovese	amaro
physical disabilities	orbo	deaf
psychical disabilities	cerebroleso	with brain issues
social economic differences	pezzente	poor
vegetables	finocchio	gay
animals	avvolotio	vulture
sexual parts	figa	female reproductive organ
sins	ghiotto	greedy
law	delinquente	outlaw
dispregiatives	aguzzino	tyrant
dispregiatives - suffix	donnaccia	woman without respect
dispregiatives - prefix	pseudo attore	pseudo actor
vulgarities	vaffanculo	go to hell

In total, we obtained a list of 1345 words. We catered for inflected forms of the lexicon by stemming (Porter, 2001) the list of words. We employed the dictionary as a lexical feature of our model. We first extracted the number of tokens present in each Facebook or YouTube comment, then we matched the tokens with the lexicon of hate words. If a match was found, we assigned a weight to the comment and counted it as a discriminating feature to tune the classifier.

Semantic features: word embeddings

The use of embeddings is widespread in the hate speech detection literature. We found examples that used embeddings to harvest semantic information from the input data (Del Vigna12 et al., 2017; Djuric et al., 2015; Nobata et al., 2016).

We also employed a set of embeddings to make the most out of the distantly supervised data. We studied two ways the impact of having word embeddings as a feature affected hate speech detection. i.) We aimed to determine the difference in performance between machine learning systems with or without the feature of word embeddings. ii.) We compared the influence of different dense vectors on text classification. The different embeddings that we employed are summarized in Table 15 present in Chapter *data*.

We entered the word embeddings in our model by converting the embeddings into a dictionary, whose keys are the lexicon and the values we selected the average of the dimension of the embeddings. Specifically, to use these word vectors in the SVM model, we mapped the content words in each sentence and we replaced them with the corresponding word embeddings values; afterwards, we computed the average value for each word embedding, in order to achieve a unique one-dimensional sentence vector with each word replaced with the corresponding embedding average.

We develop three types of word embeddings:

- Polarized embeddings
- Retrofitted embeddings
- Merged embeddings

Polarised Embeddings

A major focus of our contribution is the development of highly polarized, word embedding representations, trained on data specific to the hate speech content. Polarised embeddings are representations built on a corpus which is not randomly representative of the Italian language, rather, it is collected with a specific bias. In this context, we use data scraped from Facebook pages (communities) in order to create hate-rich embeddings.

The working hypothesis, grounded on previous studies on on-line communities (Pariser, 2011; Bozdag and van den Hoven, 2015; Seargeant and Tagg, 2018), is that each on-line community represents a different source of data, and consequently, their user-generated content can be used as proxies for specialized information.

DATA ACQUISITION We selected a set of publicly available Facebook pages that may promote or be the target of hate speech, which are reported in (Table 9). Using the Facebook API, we downloaded the comments to posts (as they are the text portions most likely to express hate), collecting a total of over one million comments for almost 13 million tokens.

MAKING EMBEDDINGS We built distributed representations over the acquired data. The embeddings have been generated with the `word2vec` skip-gram model (Mikolov, Chen, Corrado, and Dean, Mikolov et al.) using 300 dimensions, a context window of 5, and minimum frequency 1. The final vocabulary amounts to 381,697 words.

These hate-rich embeddings are used in models for hate speech detection. For comparison, we also use larger, generic embeddings that were trained on the Italian version of Wikipedia (more than 300 million tokens)[3] using `GloVe` (Berardi et al., 2015)[4]; the vocabulary amounts to 730,613 words.

As a sanity check, and a sort of qualitative intrinsic evaluation, we probed our embeddings with a few keywords, reporting in Table 18 the top three nearest neighbours for the words "immigrati" [migrants] and "trans". For the former, it is interesting to see how the polarised embeddings return more hate-leaning words compared to the generic embeddings. For the latter, in addition to hateful epithets, we also see how these embeddings capture the correct semantic field, while the generic ones do not.

We have here demonstrated that the creation of semantic tools, rather than using exclusively pre-trained resources, allows the researcher to set the most preferred parameters and have control over both the quality and quantity of the input data.

[3] http://hlt.isti.cnr.it/wordembeddings/
[4] https://nlp.stanford.edu/projects/glove/

Table 18: Intrinsic embedding comparison: words most similar to potential hate targets.

Generic Embeddings	Polarised Embeddings
"immigrati" [migrants]	
immigranti (.737)	extracomunitari (0.841)
emigranti (.731)	immigranti(0.828)
emigrati (.725)	clandestini (0.823)
"trans" [trans]	
europ (.399)	lesbo (.720)
express (.352)	puttane (0.709)
airlines (.327)	gay (.703)

At this point we owned a set of embeddings, both pretrained and newly made from distantly supervised data. With these semantic resources trained on different datasets, we decided to merge general Twitter embeddings and polarized ones to create dense vectors, so that the model could learn from both semantic spaces.

Merging embeddings

The first approach that we adopted to merge word embeddings of different natures is the concatenation via PCA [5], following the example of Chen et al. (2013).

We used PCA to ensure that resulting joined embeddings had the dimension of the smaller of the two embeddings. The advantage of this approach is that we did not need to pad the vectors to normalize the different sizes of the used embeddings (Xu et al., 2013).

We started off with two sets of embeddings, in our case Twitter embeddings, with 2,196,954 words and polarized embeddings with 381,697 words, and then applied the PCA dimensionality reduction.

The method determined the mutual vocabulary, present in both input embeddings, and the vocabulary that we found in only the one or the other set of embeddings. We proceeded with the concatenation of the embedding corresponding to the mutual vocabulary by choosing the one coming from the embeddings smaller in size by joining all embeddings into a single matrix to be fed into Scikit-Learn's PCA. As output, we obtained embeddings in a dictionary-like structure available for look-up, while keeping the vocabulary covered by the embeddings as a set. The final embeddings resulted having 300 dimensions and 2,552,460 words in the vocabulary.

Retrofitted embeddings

The second approach that we utilized was meant to maximize two of our resources: the polarized lexicon of *bad words* and the large Twitter embed-

[5] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

dings. We adopt the technique of *retrofitting*. Proposed by Faruqui et al. (2014), retrofitting is a method for tuning the vector space representations exploiting relational information obtained from semantic lexicons, such as the list of *bad words*. The retrofitting tool requires two input files: a lexicon, whose format has to be a sequence of words permuted for the length of the list, and an embeddings file.
The approach does not make any assumptions about how the input vectors were constructed and it also does not mention the optimal size of the lexicon to be used. An example of our input lexicon is reported here:

immigrato clandestino rigugiato terrorista
terrorista immigrato clandestino rifugiato
rifugiato terrorista immigrato clandestino
clandestino rifugiato terrorista immigrato

5 RESULTS AND DISCUSSION

The previous chapter introduced the binary classifier that we developed to perform hate speech detection and it explained our work on feature engineering.

In this section, we cover the results of our classifier when trained and tested on several types of datasets. We ran a series of experiments on different datasets built on manually annotated data, distantly-supervised data, and a mix of both.

Using distant supervision as a technique for labeling data is a new approach to hate speech detection. There is currently no precedent in the field for using distant supervision in this way. We believe it is not a popular approach because hate speech is a difficult task to label, even for trained annotators.
As stated in the guidelines for policymakers (Duarte et al., 2017), it is tough to recreate a data set for hate speech, since both the distribution and the target of hate speech varies through people's perception, time and place.

Our study aims to highlight the affordances provided by annotating a section of a dataset before combining it with a larger dataset extracted via distant supervision. We make the hypothesis that gold data, combined with the additional silver data, can reach higher performance than both gold and silver data when used individually.

The results of the experiments are reported in this order: (i) baseline (ii) baseline and feature engineering (iii) infusing gold and silver data (iv) end to end example with testing across datasets.

5.1 COMPARISON WITH THE STATE-OF-THE-ART

Before delving into the outcomes of this research, we first will compare the results that we obtain from our classifier and the one developed by Del Vigna12 et al. (2017), which represents the state of the state-of-the-art study for hate speech detection in Italian. In table 19 we show the two classifiers, trained on nearly the same size of data, but with different gold datasets. Del Vigna12 et al. (2017) employed 3,575 instances, and we used 3,000. Regarding the testing procedure, we both used the cross-validation method, with ten folds.

Table 19: Comparison of results from Del Vigna12 et al. (2017) and our system.

Model	Researcher	Source train	Size	CV	Features	Accuracy	Fscore (Macro)
SVC	Del Vigna	Facebook	3,575	10	Lexical + surface features	72.95	-
SVC	Our system	Facebook	3,000	10	char/word ngrams	80.5	79.6

The comparison shows that our approach is valid, as it improves the current state-of-the-art by 7 points in the F-score macro measure. Therefore, we can proceed with the exploration of the results of our classifier across different types of datasets.

5.1.1 Baseline

To begin the systematic analysis of the results, we start with covering the performance of our system as a baseline. We used three types of algorithms typical of supervised learning: LinearSVM, Naive Bayes and Logistic Regression. We used the minimum amount of features: TF-IDF word (1-2) and character (2-4) n-grams.

The training datasets included the resources containing only a single target of hatred, namely PSP dataset and the Mattarella corpus. The EVALITA dataset, the Facebook multi-target dataset, and the YouTube dataset were representatives of multi-target resources.

The testing process was systematic: we tested the classifier according to the domain of the training set. We exclusively used gold data for testing to bring uniformity to the experiments. We recognize, however, that distantly supervised datasets, when tested on gold data (held out data), could perform worse than tested on a section of their own dataset. We cannot test the YouTube training set on a test set from the same source because we did not utilize manually annotated corpora.

First, we considered gold data, which was the dataset created from the PSP and EVALITA data. Secondly, we proceeded with experiments on distantly supervised datasets.

Table 20: Results from baseline models trained and tested on Facebook EVALITA.

Model	Size	Test	CV	Features	Accuracy	Fscore (macro)
Naive Bayes	3,000	-	10	char/word ngrams	.765	.759
SVC	3,000	-	10	char/word ngrams	.805	79.6
Logistic Regression	3,000	-	10	char/word ngrams	.793	.784
Testset: EVALITA				**Distribution: *no*: 335 *yes*: 265**		
Naive Bayes	2,400	600	-	char/word ngrams	.79	.784
SVC	2,400	600	-	char/word ngrams	.798	.796
Logistic Regression	2,400	600	-	char/word ngrams	.793	789

Table 21: Results from baseline models trained and tested on PSP dataset.

Model	Size	Test	CV	Features	Accuracy	Fscore (macro)
Naive Bayes	10,357	-	10	char/word ngrams	.925	.481
SVC	10,357	-	10	char/word ngrams	.932	.583
Logistic Regression	10,357	-	10	char/word ngrams	.927	.507
Testset: L'Espresso				**Distribution: *no*: 2197 *yes*: 203**		
Naive Bayes	9,507	2,400	-	char/word ngrams	.915	.482
SVC	9,507	2,400	-	char/word ngrams	.928	.643
Logistic Regression	9,507	2,400	-	char/word ngrams	.917	.507

The training sets used in Table 20 and 21 are different in size and distribution. The EVALITA counts 3,000 samples, while the PSP dataset 12,153. Additionally, 8% of the first dataset is composed of offensive instances compared to the 1% of the PSP corpus.

We treated the datasets equally by training the classifier on three-fourths of the datasets and testing on the remaining one-fourth. The selection of the test set was random. We split the given gold datasets so that the test set for the EVALITA corpus would count 600 instances and the PSP corpus would count 2,400.

The high accuracy results in 21 hints that the system is overfitting across models. We correlate this phenomenon with the imbalanced distribution of the test set. The last fourth of the EVALITA dataset is balanced as it represents both classes of similar proportion (335 hateful vs. 265 hateful samples). On the other hand, the PSP dataset has the class *not offensive* under-represented, 200 instances out of 2,400 were found to be offensive. Overall, we noticed that the Linear SVM model was most effective.

Secondly, we consider the distantly supervised datasets:

Table 22: Results from baseline models trained and tested on Facebook multi-target dataset.

Model	Size	Test	CV	Features	Accuracy	Fscore (macro)
Naive Bayes	100,000	-	10	char/word ngrams	.733	.725
SVC	100,000	-	10	char/word ngrams	.766	.765
logistic regression	100,000	-	10	char/word ngrams	.771	.771
Testset: EVALITA FB				**Distribution: *no*: 1097 *yes*: 903**		
Naive Bayes	100,000	2,000	-	char/word ngrams	.45	.45
SVC	100,000	2,000	-	char/word ngrams	.677	.443
logistic regression	100,000	2,000	-	char/word ngrams	.473	.471
Testset: EVALITA FB + TW				**Distribution: *no*: 2465 *yes*: 1535**		
Naive Bayes	100,000	4,000	-	char/word ngrams	.463	.463
SVC	100,000	4,000	-	char/word ngrams	.570	.410
logistic regression	100,000	4,000	-	char/word ngrams	468	.465

Table 23: Results from baseline models trained and tested on Mattarella corpus.

Model	Size	Test	CV	Features	Accuracy	Fscore (macro)
Naive Bayes	100,000	-	10	char/word ngrams	.687	.469
SVC	100,000	-	10	char/word ngrams	.893	.843
Logistic Regression	100,000	-	10	char/word ngrams	.887	.824
Testset: PSP				**Distribution: *no*: 2197 *yes*: 203**		
Naive Bayes	100,000	2,400	-	char/word ngrams	.702	.479
SVC	100,000	2,400	-	char/word ngrams	.652	.507
Logistic Regression	100,000	2,400	-	char/word ngrams	.655	.481

We first cross-validate the entire dataset across ten folds, to investigate the performance of the classifier when trained and tested on the same type of data. At this stage, we notice that the YouTube dataset has the highest scores, followed by the multi-target Facebook dataset and the Mattarella corpus.

Table 24: Results from baseline models trained and tested on YouTube.

Model	Size	Test	CV	Features	Accuracy	Fscore (macro)
Naive Bayes	100,000	-	10	char/word ngrams	.842	.540
SVC	100,000	-	10	char/word ngrams	.930	.863
Logistic Regression	100,000	-	10	char/word ngrams	.918	.830
Testset: EVALITA FB + TW				**Distribution: *no*: 2465 *yes*: 1535**		
Naive Bayes	100,000	4,000	-	char/word ngrams	.631	.448
SVC	100,000	4,000	-	char/word ngrams	.645	.625
Logistic Regression	100,000	4,000	-	char/word ngrams	.64	.622

Regarding the experiments with the tests sets, the F-score did not vary substantially across the distantly supervised datasets. The value ranged from 62% to 41%. It is interesting to notice that the YouTube data also obtained the best scores. In this case, not only did the classifier test on held out data, but it also tested on data coming from different domains (Facebook and YouTube). The results again highlight that the Mattarella dataset suffers from the poor distribution of the test set from PSP, as we registered very low precision (14%, 08%, 11%) for all the three instances of the experiment.

Across the runs, we compared two factors. First, we compared distantly supervised data to gold data. Overall, we can state that gold data has better results (F-score 80%, Table 20) than its automatically labeled counterpart (F-score 50%, Table 22). Additionally, we compared datasets containing hatred against a different number of targets. We can see (Table 23) that the Mattarella dataset has better performance (F-score 50%) than the multi-target dataset (F-score 46%).

5.1.2 Adding features to the baselines

The datasets that we used in the research were either manually or automatically annotated. The former tend to be small in size (the EVALITA dataset is 3000 instances). Distantly supervised datasets can be larger but they are also noisier. To make up for these limitations, we used two features: word embeddings and a lexicon lookup.

We used different kinds of word embeddings to enrich the classifiers with additional semantic information. The information spanned from general knowledge retrieved from Twitter to only focusing on offensive content. We expected to obtain different results when employing various types of embeddings as their vector space, size and dimensions profoundly differed. The feature of lexicon lookup is thought to capture stereotypes, insults and offensive words.

The influence of word embeddings on the results of the classifier trained on gold data is small. The EVALITA test set Fscore goes from 79.6% to 80%. We registered larger improvements in the results obtained from a classifier trained on PSP data: from an F-score of 64%, it reached a score of 67%. We think the influence of word embeddings is diluted when inserted into large datasets, such as the one that we employed in this set of experiments (Tables 27, 29 28).

Table 25: Results from SVC model trained and tested on Facebook EVALITA with features

Model	Size	Test	Embeddings	Features	Accuracy	Fscore (macro)
SVC	2,400	600	hate fb	char/word ngrams	.807	.806
SVC	2,400	600	hate yt	char/word ngrams	.800	.797
SVC	2,400	600	twitter	char/word ngrams	.800	.797
SVC	2,400	600	retrofitted twitter	char/word ngrams	.801	.807
SVC	2,400	600	twitter + hate fb	char/word ngrams	.795	.792
SVC	2,400	600	-	lexicon - ngrams	.800	.798
SVC	2,400	600	hate fb	lexicon - ngrams	.815	.812

Table 26: Results from SVC model trained and tested on PSP dataset with features

Model	Size	Test	Embeddings	Features	Accuracy	Fscore (macro)
SVC	9,507	2,400	hate fb	char/word ngrams	.93	.672
SVC	9,507	2,400	hate yt	char/word ngrams	.929	.651
SVC	9,507	2,400	twitter	char/word ngrams	.928	.658
SVC	9,507	2,400	retrofitted twitter	char/word ngrams	.929	.662
SVC	9,507	2,400	twitter + hate fb	char/word ngrams	.927	.663
SVC	9,507	2,400	-	lexicon - ngrams	.925	.602
SVC	9,507	2,400	hate fb	lexicon - ngrams	.928	.648

Table 27: Results from SVC model trained and tested on Facebook mutli-target dataset with features

Model	Size	Test	Embeddings	Features	Accuracy	Fscore (macro)
SVC	100,000	2,000	hate fb	char/word ngrams	.469	.468
SVC	100,000	2,000	hate yt	char/word ngrams	.451	.450
SVC	100,000	2,000	twitter	char/word ngrams	.462	.462
SVC	100,000	2,000	retrofitted twitter	char/word ngrams	.466	.465
SVC	100,000	2,000	twitter + hate fb	char/word ngrams	.461	460
SVC	100,000	2,000	-	lexicon - ngrams	453	.452
SVC	100,000	2,000	hate fb	lexicon - ngrams	.468	.468

Table 28: Results from SVC model trained and tested on Mattarella corpus with features

Model	Size	Test	Embeddings	Features	Accuracy	Fscore (macro)
SVC	100,000	2,400	hate fb	char/word ngrams	.652	.508
SVC	100,000	2,400	hate yt	char/word ngrams	.648	.505
SVC	100,000	2,400	twitter	char/word ngrams	.654	.50
SVC	100,000	2,400	retrofitted twitter	char/word ngrams	.652	.508
SVC	100,000	2,400	twitter + hate fb	char/word ngrams	.646	.504
SVC	100,000	2,400	-	lexicon - ngrams	.650	.506
SVC	100,000	2,400	hate fb	lexicon - ngrams	.652	.508

Table 29: Results from SVC model trained and tested on YouTube dataset with features

Model	Size	Test	Embeddings	Features	Accuracy	Fscore (macro)
SVC	100,000	4,000	hate fb	char/word ngrams	.631	.618
SVC	100,000	4,000	hate yt	char/word ngrams	.630	.618
SVC	100,000	4,000	twitter	char/word ngrams	.626	.615
SVC	100,000	4,000	retrofitted twitter	char/word ngrams	.627	.615
SVC	100,000	4,000	twitter + hate fb	char/word ngrams	.632	.618
SVC	100,000	4,000	-	lexicon - ngrams	.630	.616
SVC	100,000	4,000	hate fb	lexicon - ngrams	.631	.618

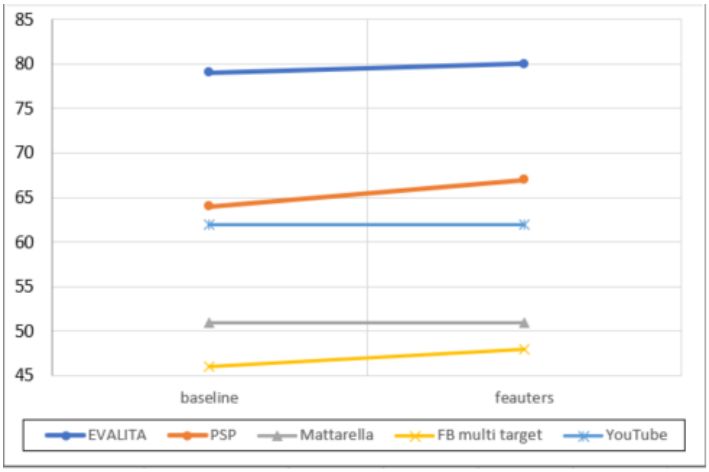


Figure 3: Representation of the obtained F-score macro across datasets.

In figure 3, we plot the results of the F-score (macro) from each of the five training sets employed in the experiments. The results highlight that gold data are the best resource to use when conducting hate speech detection studies. Furthermore, our assumption that corpora with a single target of hatred are less noisy than one built on several types of offensive expressions is validated.

The addition of word embeddings seems not to be predictive. However, we notice that the performance of systems with the newly developed hate embeddings are generally better.

5.1.3 Comparison of the performance between Single and Multi-target hate speech datasets and Infusion experiments

The primary goal of this study was to research how current supervised approaches can predict offensive content online. Due to the initial difficulties in finding annotated data that could suit our study, we decided to adopt distant supervision as the technique to retrieve and label data quickly and economically.

However, during the execution of this thesis, we noticed a growing attention towards the topic of hate speech. The 2018 Evalita [1], Germeval [2] and 2019 SemEval [3] shared tasks published challenges, as well as datasets, about hate speech detection and offensive language in social media.

As a result of this shared task on hate speech detection, we were able to gather a few manually labeled datasets and we expanded our research question to investigate the results of merging manually and automatically annotated datasets. We verified the effect that a portion of gold data had on instances of silver data after infusing the two. We did not merge all of the two datasets because hand-labeled gold data could easily be swamped by a more significant amount of distantly labeled silver data. Specifically, we infused instances extracted from the EVALITA dataset with samples taken from the Mattarella corpus and the Facebook multi-target corpus.

To understand the outcomes of the infusion, we begin with considering the performance of the classifier when trained on an increasing amount of gold and silver data alone. The test set that we chose for this task was 1200 instances from the EVALITA dataset to keep consistency through the experiments. The test set includes 798 not offensive samples and 402 hateful samples.

Table 30 is in line with our expectations concerning the idea of using distantly supervised data in text classification. The data is confirmed to be noisy and the more it is added to the dataset, the more the classifier loses the ability to correctly predict the binary labels. The comparison of these results demonstrates that we do not need to create separate resources with different types of targets to study online hate speech. However, overall we registered a better performance for the single target corpus.

[1] http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html
[2] https://projects.fzai.h-da.de/iggsa/
[3] http://alt.qcri.org/semeval2019/

Table 30: Silver data: performance of the model across different sizes of Facebook multi-target dataset

		Facebook Multi target		Mattarella	
mMdel	**Silver**	**Accuracy**	**Fscore**	**Accuracy**	**Fscore(macro)**
SVC	1,000	.655	.560	.665	.564
SVC	2,000	.618	.570	.633	.531
SVC	5,000	.590	.563	.641	.598
SVC	10,000	.523	.513	.598	.568
SVC	20,000	.492	.492	.601	.586
SVC	50,000	.463	.462	.556	.515

Table 31: Gold data: performance of the model across different sizes of Facebook EVALITA dataset

Model	**Gold Size**	**Accuracy**	**Fscore**
SVC	1,200	.45	.44
SVC	2,400	.475	.471
SVC	3,600	.763	.768
SVC	4,800	.797	.772

Table 31 represents the variations in results across different sizes of the dataset EVALITA. Opposite from what happened with the distantly supervised dataset, the more annotated data we exploit, the better performance we reach.

Table 32 reports the results of portions of silver and gold data summed together and tests on the 1,200 EVALITA instances.
If we compare the results of the infusion with the ones that we obtained when considering the gold and silver data alone, we can see that merging the two sources has positive effects on the outcomes only when it comes to merging 1,200 and 2,400 gold instances and small silver samples.

From Table 33 we can make the following observations: (i) in this context, training on small amounts of gold data is substantially more accurate than training on large amounts of distantly supervised data; (ii) overall, adding even small amounts of silver data to gold decreases performance. Few case (1,200, 2,400 gold + 1,000); (iii) also adding more gold data decreases performance, even more so than adding silver data, if the manually labelled data comes from a different dataset (thus created with different guidelines, and in this case with a different hate/non-hate distribution). Performance improves as expected when adding more data from the same dataset.

Table 32: Infusing silver and gold data.

			Multi target		Mattarella	
Model	**Gold**	**Silver**	**Accuracy**	**Fscore**	**Accuracy**	**Fscore(macro)**
SVC	1,200	1,000	.580	.558	.663	.561
		2,000	.568	.550	.636	.545
		5,000	.557	.544	.626	.577
		10,000	.480	.478	.576	.553
		20,000	.448	.448	.535	.520
		50,000	.480	.478	.55	.53
SVC	2,400	1,000	.565	.554	.671	.551
		2,000	.561	.551	.645	.535
		5,000	.556	.543	.647	.590
		10,000	.493	.490	.595	.564
		20,000	.455	.455	.564	.546
		50,000	.463	.462	.567	.553
SVC	3,600	1,000	.766	.740	664	.540
		2,000	.754	.722	.647	.590
		5,000	.763	.729	.651	.589
		10,000	.756	.718	.592	.559
		20,000	.729	.693	.572	.550
		50,000	.729	.702	.575	.559
SVC	4,800	1,000	.796	.770	.644	.534
		2,000	.795	.768	.652	.581
		5,000	.796	.764	.648	.602
		10,000	.784	.750	.585	.560
		20,000	.492	.492	0.58	.563
		50,000	.776	.744	.584	.570

Table 33: Infusing Facebook EVALITA with different types of data: EVALITA, Turin dataset and Silver data.

Gold	**Infusing with**	**Accuracy**	**Fscore**
3,600	1,200 EVALITA	.797	.772
3,600	1,000 Multi Target	.766	.740
3,600	990 Turin	.742	.688

5.1.4 End to end comparison of the datasets

For completion and clarity, we ran an end to end comparison of our experiments. It covered tests from the baseline to runs on held out data. We decided to run an end-to-end experiment that could summarize the aims of the study: investigate hate speech detection and datasets of different natures.

We chose a sample size of 4,800 instances to train the classifier, which was kept the same with the following parameters: LinearSVC, 1-2 word, 3-6 character TF-IDF n-grams and hate embeddings as features.

The example is organized as follows:

- we compared the results obtained training machine learning models on gold data from the PSP corpus and gold data from EVALITA dataset.
- we compared the results obtained training machine learning models the two Facebook based distantly supervised datasets: Mattarella corpus and the Facebook multi-target corpus.
- we merged a sample of 4800 instances of both distantly supervised datasets with three types of annotated data to verify whether different types of annotation of the same topic can influence the outcomes of the classifier.

The first step is the comparison of the manually annotated datasets (Tables 34, 35). We compared the performance of the classifier when trained on 4,800 instances of the PSP dataset and the same amount of data for the EVALITA Facebook corpus. We tested it on a separate held out sample (a Turin dataset of 990 instances), and we also cross-tested on a split of 1,200 cases, with 600 samples taken from the PSP and the other 600 extracted from the EVALITA datasets.

Table 34: Gold comparison: PSP vs. EVALITA

		Train:PSP		**Train:Evalita**	
model	**Test set**	**Accuracy**	**Fscore**	**Accuracy**	**Fscore(macro)**
SVC	Turin 990	.813	.473	.740	.538
SVC	EVALITA 1,200	.681	.484	.797	.772
SVC	PSP 1,200	.925	.613	.694	.538

Following the results that we previously obtained, we expected to find better results when the classifier was trained and tested on the data of the same nature. The outcomes of the classifier supported this expectation. Another unsurprising result was the output we obtained from the Turin test set. The test set was small, made of data retrieved from a different domain and it had an unbalanced hate-not hate distribution.

Secondly, we considered the comparison between Facebook data featuring single and multiple targets of hatred. Therefore, we compared distantly supervised datasets, which was the direct counterpart of the two gold datasets used above.

Table 35: Silver comparison: Mattarella corpus vs multi-target Facebook comments

		Mattarella		Multi-target	
Model	**Summed with dataset**	**Accuracy**	**Fscore**	**Accuracy**	**Fscore(macro)**
SVC	Turin 990	.721	.509	.607	.473
SVC	EVALITA 1,200	.530	.512	.606	.571
SVC	PSP 1,200	.544	.436	.684	.487

Distant supervision, as a labeling technique, allows one to gather large datasets. However, the labels assigned to the instances of a corpus may not represent reality.
To investigate the possible weakness of our annotation process and train the classifier on different types of datasets, we built two distantly supervised resources: one dataset containing hateful content against one single target and another one that gathered hatred against multiple targets.

Our initial assumption for making these resources was that a classifier trained on a polarized dataset would perform better than one trained on a resource with multiple topics. However, neither the results that we obtained in these experiments, or the results in the experiment that we ran before showed major significant improvement across the two datasets.

To better understand the reasons behind the predictive stagnation, we led a systematic error analysis of the result of the classifier when tested on the EVALITA test set with 1,200 instances, that we knew to be balanced across classes.

The error analysis highlighted the presence of three types of utterance: Facebook comments, Facebook posts, and article headlines. We found that 50% of the errors for both distantly supervised datasets were of type article headline. 30% of the samples were Facebook comments and the last 20% were posts. The error analysis demonstrated the classifier is not able to discern headlines about hatred and hatred itself. To name a few examples:

"Perchè nn vanno a casa loro a fare i mussulmani ?? #tuttiingabbia" - [Why don't they go back to their countries to be Muslims?? everyoneinprison]

"Ecco perchè anche i musulmani celebrano Gesù - TPI via tpi" - [This is the reason why Muslims celebrate Jesus Christ.]

Additionally, the error analysis demonstrated that the examples are highly correlated to the underground knowledge and political context. Many of the samples that our system mistakenly classified as non-offensive require extensive world knowledge, especially knowledge related to political events and cultural beliefs, to correctly annotate. While the classifier is able to correctly categorize examples such as:

"Nelle zone terremotate, in 300,000 almeno sono senza luce e/o acqua. Ma per negri, rom, islamici e clandestini, funziona tutto.." - [In the areas involved in the earthquake there are 300,000 people who do not have light or water. But for black people, roma, Muslims and illegal immigrants everything works good.]

It does not recognize subtler examples such as:

"Chiediamo al Comune di Torino un'operazione immediata di rimpatrio per tutti gli abitanti del campo rom di Via Generale Dalla Chiesa..via tutti e subito! " - [Let's ask the county of Turin to send all the roma currently settled in the camp of Via Generale Dalla Chiesa...let's do it now!]

"Terroristi ISIS nascosti tra i migranti in Sicilia! allarme! matteorenzi spiega questo! Avete stufato! Difendete l'Italia o andatevene! " - [ISIS terrorist hiding in Sicily! Watch out! @matteorenzi could you explain this phenomenon? we are tired of you! Protect Italy or leave the country!]

Such comments become offensive when settled in the political and historical context. The polarized embeddings should shift the semantic space for the classifier to learn from the prejudice against immigrants and the political situation. However, the current techniques for training the classifiers seem not to work.

Another linguistic trait the error analysis highlighted is the nature of the language used to express hatred. It is true that communities of people, such as women, gay, vegans, politicians, and immigrants are described using specific vocabulary (Del Vigna12 et al., 2017). For instance, we noticed that the word *immigrant* occurs in the same contexts as *terrorist*. However, haters use similar harsh expressions in their comments across targets.

We found that vegans and immigrants have many references to the idea of "reporting" (denunciare), "keeping babies away from" (tenete lontano i bambini), "not being true Italians" (non siete veri italiani). Or they are placed on the same level, as demonstrated in this example extracted from the EVALITA test set:

"QuintaColonna come volevasi dimostrare ..dopo i nazi vegani i musulmani ... c'è attinenza dunque. . . " - [#Quintacolonna, QED..after Nazi vegans, here come the Muslims.. There is a similarity.]

The classifier also has difficulty discerning what a headline is and what it an actual attack against specific communities of people are. However, since the results of the classifier on different test sets are very similar, we can infer that the linguistic component is active enough to leverage the predictive power of the classifier across test sets.

We believe that the model does not make better predictions when trained on a type of hatred, because it is learning the same information that it would gain from a richer dataset. We addressed the specific lexicon used to refer to communities by using a vocabulary of "bad words", however, when adding

the lexicon look up as a feature, the classifier performs worse than without the feature.

The last step in this example involves the infusion and the test on held out data. We merged a sample of 4,800 instances of both distantly supervised datasets with three sets of annotated data (Table 36) to verify whether different types of annotation of the same topic could influence the outcomes of the classifier. In this case, we did not register significant differences across datasets when adding gold data of different nature to a silver dataset. When adding 1,000 samples from the same datasets the results tend to be higher (Fscore: 56% for Mattarella corpus and 59% for Facebook multi-target).
The outcomes of the end-to-end example also show that adding more gold data decreases performance, even more so than adding silver data.

Table 36: Merging a sample of 4,800 silver samples with three types of annotated data.

		Mattarella		**Facebook multi-target**	
model	**summed with dataset**	**Accuracy**	**Fscore**	**Accuracy**	**Fscore(macro)**
SVC	Turin 990	.633	.543	.813	.473
SVC	EVALITA 1,200	.521	.517	.574	.556
SVC	PSP 1,200	.925	.577	.672	.599

6 CONCLUSION

Answers to research questions

In the previous chapters, we used distant supervision to annotate a large amount of data downloaded from YouTube and Facebook. We input the data into a binary classifier and we ran series of tests to investigate how our system behaved at detecting hate speech. The outcome of our classifier highlighted several results:

- We have provided a comprehensive overview of the research on hate speech and hate speech detection. We can conclude that hate speech detection is a difficult task to accomplish, especially with distantly supervised data.
- The main contribution of the thesis was the *annotation of tree datasets for the Italian language* and the development of polarised word embeddings. We developed two datasets from Facebook comments, one containing hatred against one target and the second containing hatred against multiple targets. We also created a dataset of data scraped from YouTube.
  The new resources are available at this GitHub repository:
  https://github.com/ClaZaghi/Hate-Speech-detection-in-Italian-social-media.
- Another crucial contribution of the thesis was our investigation of the use of manually labeled (gold) data versus automatically labeled (silver) data in a machine learning task such as hate speech detection.

  Silver data did not prove to be a successful alternative but it can be a complementary strategy to using purely gold data. However, we noticed that gold data, when infused with small portions of silver data, reaches better results than when used alone.

  Our final experiments also suggest that gold data is not better than silver data if it comes from a different dataset, as we noticed when we employed the manually labeled Turin dataset. This result highlights a crucial aspect related to the creation of manually labeled datasets, especially in the highly subjective area of hate speech and affective computing in general, where different guidelines and different annotators introduce large biases and discrepancies across datasets.
- We considered the aspects of using datasets with hatred addressed to a single target (e.g. the politician's community) versus hatred addressed to multiple targets (e.g. women or vegans), realizing that there is no need to control the amount of target of the hatred. Hateful content can contain specific traits according to the target it caters. However, haters spread hate using common patterns. We registered similar insults, stereotypes and offensive words across all targets of hate.

- We built a supervised learning model to perform hate speech detection using Vector Machine, Logistic Regression and Naive Bayes. We added the following features: a lexicon of hateful and offensive words features (De Mauro, 2016), and word and characters n-grams. Additionally, we developed a hate polarized lexicon and word embeddings to tune the predictive power towards the hateful content present in the social media content. We obtained results that are in line with the baseline of the Italian models so far implemented (F-score 80%).

Limitations

Our take on hate speech detection suffers from two limitations, namely, the type of data collection and the machine learning model that we adopted.

Distant supervision, as a labeling technique, has the convenience of quickly generating large portions of annotated data. However, distant supervision does not insure that the model's predictions are correct and representative of reality.

The uncertainty in our results pushed us to explore different kinds of datasets to have a better overview of the phenomenon. We considered manually and automatically annotated datasets and we created two separate corpora, one with hatred against a single group, and one with hatred against multiple targets.

The longitudinal exploration of hate speech in natural data led to interesting outcomes, such as that offensive content against different targets present common linguistic patterns. However, manually annotated datasets do not require one to employ this approach to validate a system's findings.

The second limitation of this study is the exclusion of neural networks in our model. Supervised approaches have been shown to obtain good results, although they suffer from limitations as far as the size and domain of the training data is concerned.

Simple SVM models with word embeddings (Del Vigna12 et al., 2017) and TF-IDF n-grams (Davidson et al., 2017) showed competitive performance compared to rule based approaches, however Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) classifiers (Del Vigna12 et al., 2017; Badjatiya et al., 2017; Gao and Huang, 2017; Chu et al., 2016) turned out to be the most efficient algorithms for this task, as they reached 90% in F-score.

Future work

We pointed out the limitations of using distant supervision as a technique to label datasets quickly and ecologically. We also stated that machine learning approaches are becoming obsolete since the neural network take on hate speech detection has shown to be very successful. Therefore, the improvements that still need to be done in this research have to be found in the data collection and the model development.

We found two possible ways to improve the distant supervision approach that was used in this research: First, we used one proxy to assign labels

automatically to the content downloaded from social media. If the content was extracted from pages promoting hatred against particular communities, we labeled it as hateful. Therefore, we used the source of the social media content as a discriminant variable.

However, the choice of proxies can be broadened so that it can capture a wider spectrum of hate-related content. The use of lexical proxies is prevalent in the literature which have proven the ability to capture phenomena such as insults and stereotypes (Davidson et al., 2017; Waseem and Hovy, 2016; Burnap and Williams, 2016; Magu et al., 2017).
Another improvement that we could adopt is a more sophisticated method of infusion of gold and silver data. We are currently taking a combination of gold data and silver data. However, when the silver data is extensive, the information provided by gold data could be diluted by noisy silver data.

Pershina et al. (2014) attempted to find the predicting relations and features present within a small portion of gold data and generalize them into training guidelines. This approach integrated the extracted rules into a latent variable model so that the larger silver dataset could learn from them.

Since hate speech is becoming a trending topic in research, it is possible to find small hand-labeled datasets. This trend could lead researchers to move from approximate techniques, such as distant supervision, toward adopting up-sampling or self-training techniques.

The second weak point of the thesis is the exclusive use of supervised learning algorithms. To be up-to-date with the current research and provide a complete overview of the tools to detect offensive content in social media, neural network approaches should have also been included.

7 ACTIVE COUNTERMEASURES AND PRACTICAL APPLICATIONS

The recent widespread dissemination of hate and intolerance across social media led to the birth of several active associations to monitor, study and counter this phenomenon. These associations operate both at a national and international level. Two examples of active movements in the European Union are NoHateSpeech [1] and the EMORE PROJECT [2]. If NoHateSpeech works to spread awareness and promote the systematic reporting of hate speech cases, the EMORE PROJECT is aimed at the development and testing of a knowledge model on online hate speech.

The EMORE PROJECT pointed out three phenomena related to hate speech occurring in Italy: (i) the identification of the communities who are the primary targets of online hatred including: Roma, Sinti, Caminanti, foreign citizens, Islamic communities and LGBT people. (ii) the causes that led to such behavior, such as the refugee's crisis, the use of personal data for profit by radical political parties and media, and finally, the discrimination towards African people and Muslims. (iii) how Italian people are dealing with this situation: even if online hate speech crimes are growing, many victims decide not to pursue legal action because the legal system is too tedious.

At the national level, there are other operative movements. `paroleostili` [3], is a social project aiming at monitoring and creating awareness on the violence words carry. Additionally, at the University of Turin there is a Hate Speech Monitoring organization @UniTO [4], which conducts studies and holds workshops on the topic of hate speech.

The purpose of this thesis was to develop a system to detect offensive content online. As a practical use of our research, in the next section, we propose several possible tools and applications to utilize our findings.

Several parameters have to be considered when designing such tools. The development could be catered to a specific platform, or it could work across platforms. Another important criterion is timing. Hate speech could be addressed before or after it has been publicly shared online.

Considering these parameters, several practical applications could include the following:

- Following the model of the Adblock plugin [5], available for most web browsers, a hate speech blocker could be created. The plugin would function not as censorship, but as a hate speech detection and hiding plugin. It could be implemented for underage individuals, so they won't think hate speech is a tolerable habit. It could also be used to

[1] http://www.nohatespeech.it/
[2] https://www.emoreproject.eu/
[3] http://paroleostili.com/
[4] http://hatespeech.di.unito.it/
[5] https://https://adblock-chrome.it.softonic.com/

filter hateful content on a page so the user does not need to waste their time viewing hateful content.

- Another application could be a plugin to work with a chat-box. When the user types, the plugin, with the help of a regression algorithm, could indicate to the user how hateful the content they are typing could be perceived. This tool is currently being developed by Google's incubator Jigsaw [6].

- The development of a plugin that allows users to quickly report content that they believe to be hateful against a particular group of people. The outcome would be twofold: the reporting and removal of the hateful data, and the creation of a dataset that can be fed into a binary classifier to perform stochastic studies. As mentioned several times, the main issue when studying hate speech is the data collection and this tool could be very valuable for future research.

In this study, we did not show any experiments with manually annotated data. We concentrated on the effects of distant supervision as a labeling technique. However, we developed and proposed an annotation interface at the TABU DAG [7], the annual international linguistics conference held at the University of Groningen.

The platform allows the users to annotate data randomly scraped from the internet via distant supervision and it is saved on a server. At the moment, the platform allows the annotation in both Italian and German. A link to the annotation platform that we called AnnotHate is provided here:

https://github.com/ClaZaghi/AnnotHate/

We have talked about methods to counter hate speech and possible applications to do it automatically. However, we have noticed a counteractive phenomenon, which is the perpetuation of hate speech in specific cases. (Ziccardi, 2016) witnessed the general acceptance of hatred online by users.

Another phenomenon that we observed is where the target accepts the content directed toward them and sometimes even encourages it.
A blatant example of this practice can be found on the public pages of women who work in the entertainment business in Italy.We observed that pages of female television personalities are dominated by thousands of sexually explicit comments towards the women themselves and the whole category of women. However, offensive comments are not blocked by the page owner, but instead they are used to trigger visibility and virality of their image. This is one of the main reasons why hatred is tolerated and accepted by these groups.

Another interesting phenomenon that we observed is that some Facebook pages select the most hateful comments and re-share them to give them more visibility.

Therefore, we see two parallel and opposite behaviors towards hate speech. On one hand, associations, movements, and research centers monitor hate speech and create plans to limit the spreading of offensive words. On the

[6] https://www.perspectiveapi.com/#/
[7] https://www.let.rug.nl/tabudag/

other hand, the same offensive words are a source of income for the owner of the social media content which triggers them.

These are two opposing forces that both are a source of harm, however, we hope that the massive campaign against hate speech can overcome the hilarity and entertainment fueled by hatred.

BIBLIOGRAPHY

Agarwal, S. and A. Sureka (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*.

Ansa (2018). World speech day, primi in unione europea per hate speech.

Badjatiya, P., S. Gupta, M. Gupta, and V. Varma (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760. International World Wide Web Conferences Steering Committee.

Basile, A., T. Caselli, and M. Nissim (2017). Predicting Controversial News Using Facebook Reactions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy.

Berardi, G., A. Esuli, and D. Marcheggiani (2015). Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Biemann, C. (2007). Unsupervised and knowledge-free natural language processing in the structure discovery paradigm. In *Ausgezeichnete Informatikdissertationen*, pp. 31–40.

Bleich, E. (2014). Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the usa and europe. *Journal of Ethnic and Migration Studies 40*(2), 283–300.

Bosco, C., V. Patti, M. Bogetti, M. Conoscenti, G. F. Ruffo, R. Schifanella, and M. Stranisci (2017). Tools and resources for detecting hate and prejudice against immigrants in social media. In *SYMPOSIUM III. SOCIAL INTERACTIONS IN COMPLEX INTELLIGENT SYSTEMS (SICIS) at AISB 2017*, pp. 79–84. AISB.

Bozdag, E. and J. van den Hoven (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology 17*(4), 249–265.

Breda, M. (2018). Mattarella, il giorno più amaro: le minacce sul web evocano l'assassinio del fratello.

Bröckling, M., V. Coquaz, A. Fanta, A. Langley, M. Munafò, J. Pütz, F. Sironi, L. Thüer, and R. Wazir (2018, May). Political Speech Project. https://rania.shinyapps.io/PoliticalSpeechProject/.

Brugger, W. (2002). Ban on or protection of hate speech-some observations based on german and american law. *Tul. Eur. & Civ. LF 17*, 1.

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux (2013). API design for machine

learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Burnap, P. and M. L. Williams (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science 5*(1), 11.

Chatzakou, D., N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pp. 13–22. ACM.

Chen, Y., B. Perozzi, R. Al-Rfou, and S. Skiena (2013). The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*.

Chu, T., K. Jue, and M. Wang (2016). Comment abuse classification with deep learning. Von https://web. stanford. edu/class/cs224n/reports/2762092. pdf abgerufen.

Dashti, A. A., A. A. Al-Kandari, and H. H. Al-Abdullah (2015). The influence of sectarian and tribal discourse in newspapers readers' online comments about freedom of expression, censorship and national unity in kuwait. *Telematics and Informatics 32*(2), 245–253.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

De Marneffe, M.-C. and C. D. Manning (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

De Mauro, T. (2016). Le parole per ferire. *Internazionale*.

Del Vigna12, F., A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi (2017). Hate me, hate me not: Hate speech detection on facebook. pp. 88–93.

Delgado, R. and J. Stefancic (2004). *Understanding words that wound.* Westview Press.

Djuric, N., J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pp. 29–30. ACM.

Duarte, N., E. Llanso, and A. Loup (2017). Mixed messages? the limits of automated social media content analysis. *Center for Democracy and Technology*.

Emmery, C., G. Chrupała, and W. Daelemans (2017). Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 50–55.

Faruqui, M., J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

F.Q. (2018). Attacchi troll a mattarella, pm roma indagano per attentato alla libertà del presidente: 'account creato su snodo milano.

Gagliardone, I., D. Gal, T. Alves, and G. Martinez (2015). *Countering online hate speech*. Unesco Publishing.

Gao, L. and R. Huang (2017). Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

Gitari, N. D., Z. Zuping, H. Damien, and J. Long (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering 10*(4), 215–230.

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

Greevy, E. (2004). *Automatic text categorisation of racist webpages*. Ph. D. thesis, Dublin City University.

Greevy, E. and S. Smeaton (2004). Text categorization of racist texts using a support vector machine. *7 es Journées internationales d'Analyse statistique des Données Textuelles*.

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

Jurafsky, D. and J. H. Martin (2014). *Speech and language processing*, Volume 3. Pearson London.

Kiritchenko, S. and S. M. Mohammad (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kong, H.-K., Z. Liu, and K. Karahalios (2018). Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 438. ACM.

Kwok, I. and Y. Wang (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

Magu, R., K. Joshi, and J. Luo (2017). Detecting the hate code on social media. *arXiv preprint arXiv:1703.05443*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Mills, A. J. (2012). Virality in social media: the spin framework. *Journal of public affairs 12*(2), 162–169.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics.

Minucci, E. (2018). Tsunami di commenti sui social, la crisi politica conquista twitter e facebook.

Mish, F. and J. Morse. Merriam-webster's collegiate dictionary, (springfield: Merriam-webster, inc.).

Mondal, M., L. A. Silva, and F. Benevenuto (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 85–94. ACM.

Musto, C., G. Semeraro, M. de Gemmis, and P. Lops (2016). Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pp. 307–308. ACM.

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pp. 145–153. International World Wide Web Conferences Steering Committee.

Palmer, A., M. Robinson, and K. K. Phillips (2017). Illegal is not a noun: Linguistic form for detection of pejorative nominalizations. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 91–100.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Park, J. H. and P. Fung (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Pelosi, S., A. Maisto, P. Vitale, and S. Vietri (2017). Mining offensive language on social media. In *CLiC-it*.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Pershina, M., B. Min, W. Xu, and R. Grishman (2014). Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Volume 2, pp. 732–738.

Pitsilis, G. K., H. Ramampiaro, and H. Langseth (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 1–13.

Poletto, F., M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *CEUR WORKSHOP PROCEEDINGS*, Volume 2006, pp. 1–6. CEUR-WS.

Pool, C. and M. Nissim (2016, December). Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, pp. 30–39. COLING 2016.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

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

Raisi, E. and B. Huang (2016). Cyberbullying identification using participant-vocabulary consistency. *arXiv preprint arXiv:1606.08084*.

Ribeiro, M. H., P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr (2018). Characterizing and detecting hateful users on twitter. *arXiv preprint arXiv:1803.08977*.

Sanguinetti, M., F. Poletto, C. Bosco, V. Patti, and M. Stranisci (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of LREC*.

Scafi, M. (2018). L'impeachment non fa paura. mattarella minacciato sul web.

Seargeant, P. and C. Tagg (2018). Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum. *Discourse, Context & Media*.

Silva, L. A., M. Mondal, D. Correa, F. Benevenuto, and I. Weber (2016). Analyzing the targets of hate in online social media. In *ICWSM*, pp. 687–690.

Statista (2018). Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes).

Su, H.-P., Z.-J. Huang, H.-T. Chang, and C.-J. Lin (2017). Rephrasing profanity in chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 18–24.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Warner, W. and J. Hirschberg (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pp. 19–26. Association for Computational Linguistics.

Waseem, Z. and D. Hovy (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93.

Xu, H., C. Caramanis, and S. Mannor (2013). Outlier-robust pca: the high-dimensional case. *IEEE transactions on information theory 59*(1), 546–572.

Xu, Z. and S. Zhu (2010). Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pp. 1–10.

Yuan, S., X. Wu, and Y. Xiang (2016). A two phase deep learning model for identifying discrimination from tweets. In *EDBT*, pp. 696–697.

Ziccardi, G. (2016). *L'odio online: violenza verbale e ossessioni in rete*, Volume 102. Raffaello Cortina.