Policycorpus XL:
An Italian Corpus for the Detection of Hate Speech Against Politics

Fabio Celli[1], Mirko Lai[2], Armend Duzha[1], Cristina Bosco[2], Viviana Patti[2]
1. Research & Development, Gruppo Maggioli, Italy
2. Dept. of Informatics, University of Turin, Italy
fabio.celli@maggioli.it, mirko.lai@unito.it,
armend.duzha@maggioli.it, bosco@di.unito.it, patti@di.unito.it

Abstract

In this paper we describe the largest corpus annotated with hate speech in the political domain in Italian. Policycorpus XL has 7000 tweets, manually annotated, and a presence of hate labels above 40%, while in other corpora of the same type is usually below 30%. Here we describe the collection of data and test some baseline with simple classification algorithms, obtaining promising results. We suggest that the high amount of hate labels boosts the performance of classifiers, and we plan to release the dataset in a future evaluation campaign.

1 Introduction and Background

In recent years, computer mediated communication on social media and microblogging websites has become more and more aggressive (Watanabe et al., 2018). It is well known that people use social media like Twitter for a variety of purposes like keeping in touch with friends, raising the visibility of their interests, gathering useful information, seeking help and release stress (Zhao and Rosson, 2009), but the spread of fake news (Shu et al., 2019; Alam et al., 2016) has exacerbated a cultural clash between social classes that emerged at least since after the debate about Brexit (Celli et al., 2016) and more recently during the pandemics (Oliver et al., 2020). Despite the fact that the behavior online is different from the behavior offline (Celli and Polonio, 2015), we observe more and more hate speech in social media, to the point where it has become a serious problem for free speech and social cohesion.

Hate speech is defined as any expression *that is abusive, insulting, intimidating, harassing, and/or incites, supports and facilitates violence, hatred, or discrimination. It is directed against people (individuals or groups) on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth* (Erjavec and Kovačič, 2012). In response to the growing number of hate messages, the Natural language Processing (NLP) community focused on the classification of hate speech (Badjatiya et al., 2017) and the analysis of online debates (Celli et al., 2014). In particular, many worked on systems to detect offensive language against specific vulnerable groups (e.g., immigrants, LGBTQ communities among others) (Poletto et al., 2017) (Poletto et al., 2021), as well as aggressive language against women (Saha et al., 2018). An under-researched - yet important - area of investigation is anti-politics hate: the hate speech against politicians, policy makers and laws at any level (national, regional and local). While anti-policy hate speech has been addressed in Arabic (Guellil et al., 2020) and German (Jaki and De Smedt, 2019), most European languages have been under-researched. The bottleneck in this field of research is the availability of data to train good hate speech detection models. In recent years, scientific research contributed to the automatic detection of hate speech from text with datasets annotated with hate labels, aggressiveness, offensiveness, and other related dimensions (Sanguinetti et al., 2018). Scholars have presented systems for the detection of hate speech in social media focused on specific targets, such as immigrants (Del Vigna et al., 2017), and language domains, such as racism (Kwok and Wang, 2013), misogyny (Basile et al., 2019) or cyberbullying (Menini et al., 2019). Each type of hate speech has its own vocabulary and its own dynamics, thus the selection of a specific domain is crucial to obtain clean data and to restrict the scope of experiments and learning tasks.

In this paper we present a new corpus, called Policycorpus XL, for hate speech detection from Twitter in Italian. This corpus is an extension of the Policycorpus (Duzha et al., 2021). We selected Twitter as the source of data and Italian as the target language because Italy has, at least since the elections in 2018, a large audience that pays attention to hyper-partisan sources on Twitter that are prone to produce and retweet messages of hate against policy making (Giglietto et al., 2019).

The paper is structured as follows: after a literature review (Section 2), we describe how we collected and annotated the data (Section 3), we evaluate some baselines (Section 4), and we pave the way for future work (Section 5).

2 Related Work

Hate Speech in social media is a complex phenomenon, whose detection has recently gained significant traction in the Natural Language Processing community, as attested by several recent review works (Poletto et al., 2021). High-quality annotated corpora and benchmarks are key resources for hate speech detection and haters profiling in general (Jain et al., 2021), considering the vast number of supervised approaches that have been proposed (MacAvaney et al., 2019).

Early datasets on Hate Speech, especially in English, were produced outside any evaluation campaigns (Waseem and Hovy, 2016), (Founta et al., 2018) as well as inside such competitions. These include SemEval 2019, where a multilingual hate speech corpus against immigrants and women in English and Spanish (Basile et al., 2019) was released, and PAN 2021, that provided a dataset for the detection of hate spreader authors in English and Spanish (Rangel et al., 2021). Most Italian datasets in the field of hate speech have been released during competitions and evaluation campaigns. There are:

- the Italian HS corpus (Poletto et al., 2017),
- HaSpeeDe-tw2018 and HaSpeeDe-tw2020, the datasets released during the EVALITA campaigns (Sanguinetti et al., 2020),
- the Policycorpus (Duzha et al., 2021), the only dataset in Italian that is annotated with hate speech in the political domain.

The Italian HS corpus is a collection of more than 5700 tweets manually annotated with hate speech, aggressiveness, irony and other forms of potentially harassing communication. The HaSpeeDe-tw corpora are two collections of 4000 and 8100 tweets respectively, manually annotated with hate speech labels and containing mainly anti-immigration hate (Bosco et al., 2018). The Policycorpus is a collection of 1260 tweets manually annotated with hate speech labels against politics and politicians. We decided to expand it and produce a new dataset.

Hate speech is hard to annotate and hard to model, with the risk of creating data that is biased and making the models prone to overfitting. In addition to this, literature also reports cases of annotators' insensitivity to differences in dialect that can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. It is the case of African American English (Sap et al., 2019) but it potentially applies to Italian as well, as it is a language full of dialects and regional offenses.

Hate speech is intrinsically associated to relationships between groups, and also relying in language nuances. There are many definitions of hate speech from different sources, such as European Union Commission, International minorities associations (ILGA) and social media policies (Fortuna and Nunes, 2018). In most definitions, hate speech has specific targets based on specific characteristics of groups. Hate speech is to incite violence, usually towards a minority. Moreover, hate speech is to attack or diminish. Additionally, humour has a specific status in hate speech, and it makes more difficult to understand the boundaries about what is hate and what is not.

In the political domain we find all of these aspects, especially messages against a minority (politicians) to attack or diminish. We think that more resources are needed for the classification of hate speech in Italian in the political domain, hence we decided to collect and annotate more data for this task.

In the next section, we describe how we created the dataset and annotated it with hate speech labels.

3 Data Collection and Annotation

Starting from the Policycorpus, we expanded it from 1260 to 7000 tweets in Italian, collected using snowball sampling from Twitter APIs. As initial seeds, we used the same set of hashtags used for the Policycorpus, for instance: #dpcm (decree of the president of the council of ministers), #legge (law) and #leggedibilancio (budget law). We removed duplicates, retweets and tweets containing only hashtags and urls. At the end of the sampling process, the list of seeds included about 6000 hashtags that co-occurred with the initial ones. We grouped the hashtags into the following categories:

- Laws, such as #decretorilancio (#relaunchdecree), #leggelettorale (#electorallaw), #decretosicurezza (#securitydecree)
- Politicians and policy makers, such as #Salvini, #decretoSalvini (#Salvinidecree), #Renzi, #Meloni, #DraghiPremier
- Political parties, such as #lega (#league), #pd (#Democratic Party)
- Political tv shows, such as #ottoemezzo, #nonelarena, #noneladurso, #Piazzapulita
- Topics of the public debate, such as #COVID, #precari (#precariousworkers), #sicurezza (#security), #giustizia (#justice), #ItalExit
- Hyper-partisan slogans, such as #vergognaConte (#shameonConte), #contedimettiti (#ConteResign) or #noicontrosalvini (#WeareagainstSalvini)

Examples of collected hashtags are reported in Figure 1

Recent shared tasks (Agerri et al., 2021; Cignarella et al., 2020; Aker et al., 2016) promoted the use of contextual information about the tweet and its author (including his/her social media network) for improving the performance of stance detection. Here, with the aim to stimulate the exploration of data augmentation on hate speech detection, we shared additional contextual information based on the post such as: the number of retweets and the number of favours (the number of tweets that given user has marked as favorite - favours_count field) the tweet received, the device used for posting it (e.g. iOS or Android), the posting date and location, and an attribute that states if the post is a tweet, a retweet, a reply, or a quote. Furthermore, we collected contextual information related to the authors of these posts such as: the

Figure 1: Wordclouds of the hashtags collected with frequency higher than 2.

number of tweets ever posted, the user's description and location, the number of her/his followers and of her/his friends, the number of public lists that this user is a member of and the date her/his account has been created.

All these contextual information are respectively part of the "root-level" attributes of the *Tweets* and *Users* objects that Twitter returns in JSON format through its APIs. Additionally, we planned to explore the interests of the author collecting the list of her/his following (the users she/he follows) employing the *following* API endpoint. Moreover, for exploring the author's social interactions, we used the *Academic Full Search API* for recovering the list of the users that she/he has retweeted to and replied to in the last two years.

The enhanced Policycorpus has been finally anonymised mapping each *tweet_id*, *users_id*, and *mention* with a randomly generated ID. To produce gold standard labels, we asked two Italian native speakers, experts of communication, to manually label the tweets in the corpus, distinguishing between hate and normal tweets according to the following guidelines: By definition, hate speech is *any expression that is abusive, insulting, intimidating, harassing, and/or incites to violence, ha-*

tred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth. (Erjavec and Kovačič, 2012). Below We provide some examples with translation in English:

1. "*Un chiaro #NO all #Olanda che ci vorrebbe sì utilizzatori delle risorse economiche del #MES ma in cambio della rinuncia dell Italia alla propria autonomia di bilancio. All Olanda diciamo: grazie e arrivederci NON CI INTERESSA!!*"[1]

The first example is normal because it does not contain hate, insults, intimidation, violence or discrimination.

2. "*...Sta settimanale passerella dello #sciacallo #no #proprioNo! Ascoltare un #pagliaccio padano dopo un vero PATRIOTA un medico di #Bergamo non si può reggere ne vedere ne ascoltare. Giletti dovrebbe smetterla di invitare certi CAZZARIPADANI! #COVID-19 #NonelArena*"[2]

The second example contains hate speech, including insults like #clown and #jackal.

3. "*Dico la mia... #Draghi è un grande economista ma a noi non serve un economista stile #Monti... A noi non serve un altro #governo tecnico per ubbidire alla lobby delle banche! A noi serve un leader politico! A noi serve un #ItalExit! A noi serve la #Lira! #No a #DraghiPremier*"[3]

The last example is a normal case, despite the strong negative sentiment. It might be controversial for the presence of the term *lobby*, often used in abusive contexts, but in this case, it is not directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation or political conviction.

The Inter-Annotator Agreement is *k*=0.53. Although this score is not high, it is in line with the score reported in the literature for hate speech against immigrants (*k*=0.54) (Poletto et al., 2017) and indicates that the detection of hate speech is a hard task for humans.

All the examples in disagreement were discussed and an agreement was reached between the annotators, with the help of a third supervisor. The cases of disagreements occurred more often when the sentiment of the tweet was negative, this was mainly due to:

- The use of vulgar expressions not explicitly directed against specific people but generically against political choices.
- The negative interpretation of hyper-partisan hashtags, such as #contedimettiti (#ConteResign) or #noicontrosalvini (#WeareagainstSalvini), in tweets without explicit insults or abusive language.
- The substitution of explicit insults with derogatory words, such as the word "circus" instead of "clowns".

The amount of hate labels in the original Policycorpus was 11% (1124 normal and 140 hate tweets), strongly unbalanced like the Italian HS corpus (17% of hate tweets), because it reflects the raw distribution of hate tweets in Twitter. The HaSpeeDe-tw corpus (32% of hate tweets) instead has a distribution that oversamples hate tweets and it is better for training hate speech models. Following the HaSpeeDe-tw example, in Policycorpus XL we collected more tweets of hate, randomly discarding normal tweets to reach at least 40% of hate tweets in the corpus. In the end we have 40.6% of hate labels and 59.4% of normal labels, distributed between training and test set as shown in figure 2.

We note in the style of these tweets that there is a substantial overlap among the top unigrams in the two classes, as shown in Figure 3. We suggest that weak signals, like less frequent words, are key features for the classification task.

In the next section, we report and discuss the results of classification experiments.

[1] a clear #NO to the #Netherlands that would like us to be users of the #MES economic resources but in exchange for Italy's renunciation of its budgetary autonomy. To Netherlands we say: thank you and goodbye, WE ARE NOT INTERESTED !!

[2] ... There is a weekly catwalk of the #jackal #no #notAtAll! Listening to a Padanian #clown after a true PATRIOT a doctor from #Bergamo cannot be held, seen or heard. Giletti should stop inviting certain SLACKERS FROM THE PO VALLEY! #COVID-19 #NonelArena

[3] I have my say ... #Draghi is a great economist but we don't need a #Monti-style economist ... We don't need another technical #government to obey the banking lobby! We need a political leader! We need a #ItalExit! We need the #Lira! #No to #DraghiPremier

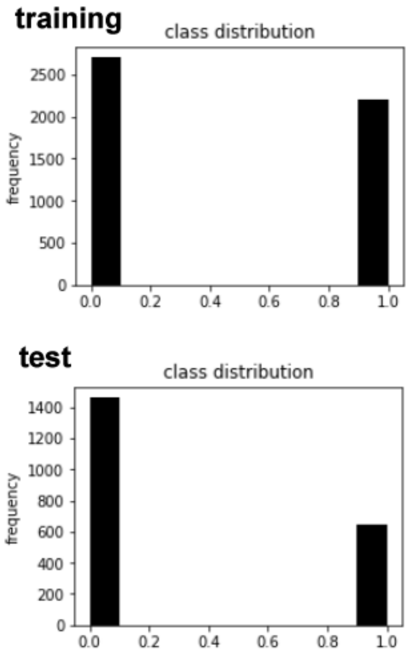

Figure 2: Distribution of classes in Policycorpus-XL training and test sets.

Figure 3: Wordclouds of the unigrams most associated to the normal and hate classes respectively. It shows a substantial overlap among the top unigrams in the two classes.

4 Baselines

In order to set the baselines for the hate speech classification task on Policycorpus-XL, we tested different classification algorithms. We are using a 70 train and 30 test percentage split, the training set shape is 4900 instances and 300 features, while the test set shape is 2100 instances and 300 features. The 300 features are the normalized frequencies of the 300 most frequent words extracted from tweets without removing the stopwords. Table 1 reports the result of classification.

algorithm	**balanced acc**	**macro F1**
majority baseline	0.500	0.37
naive bayes	0.783	0.78
decision trees	0.763	0.76
SVMs	0.788	0.79

Table 1: Results of classification with different algorithms.

We used Scikit-Learn to compute a majority baseline with a dummy classifier, that assigns all the instances to the most frequent class (normal tweets), a naive bayes classifier, a decision tree and Support Vector Machines (SVMs). The best performance for the classification of hate speech has been achieved with the SVM classifier, that has a very high precision (0.94) and poor recall (0.60). All the algorithms a The results are in line with the scores obtained by the systems on the HaSpeeDe-tw 2020 dataset at EVALITA, and we believe that there is still great room for improvement with the Policycorpus-XL, as we exploited very simple and limited features.

5 Conclusion and Future Work

We presented a large corpus of Twitter data in Italian, manually annotated with hate speech labels. The corpus is an extension of a previous one, the first corpus annotated with hate speech in the political domain in Italian.

Given the rising amount of hate messages online, not just against minorities but more and more against policies and policymakers, it is urgent to understand the phenomenon and train classifiers that could prevent people to disseminate hate in the public debate. This is very important to keep democracies alive and grant a free speech that is respectful of other people's freedom.

We plan to distribute the corpus in the next edition of EVALITA for a specific HaSpeeDe-tw task.

Acknowledgments

The research leading to the results presented in this paper has received funding from the PolicyCLOUD project, supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 870675.

References

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. VaxxStance@IberLEF 2021: Going Beyond Text in Crosslingual Stance Detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*. CEUR-WS.org.

Ahmet Aker, Fabio Celli, Adam Funk, Emina Kurtic, Mark Hepple, and Rob Gaizauskas. 2016. Sheffield-trento system for sentiment and argument structure enhanced comment-to-article linking in the online news domain.

Firoj Alam, Fabio Celli, Evgeny Stepanov, Arindam Ghosh, and Giuseppe Riccardi. 2016. The social mood of news: self-reported annotations to design automatic mood detection systems. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 143–152.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Fabio Celli and Luca Polonio. 2015. Facebook and the real world: Correlations between online and offline conversations. *CLiC it*, page 82.

Fabio Celli, Giuseppe Riccardi, and Arindam Ghosh. 2014. Corea: Italian news corpus with emotions and agreement. In *Proceedings of CLIC-it 2014*, pages 98–102.

Fabio Celli, Evgeny A Stepanov, Massimo Poesio, and Giuseppe Riccardi. 2016. Predicting brexit: Classifying agreement is better than sentiment and pollsters. In *PEOPLES@ COLING*, pages 110–118.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Sardistance@evalita2020: Overview of the task on stance detection in italian tweets. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Aachen, Germany, December. CEUR-WS.org.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Armend Duzha, Cristiano Casadei, Michael Tosi, and Fabio Celli. 2021. Hate versus politics: detection of hate against policy makers in italian tweets. *SN Social Sciences*, 1(9):1–15.

Karmen Erjavec and Melita Poler Kovačič. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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

Fabio Giglietto, Nicola Righetti, Giada Marino, and Luca Rossi. 2019. Multi-party media partisanship attention score. estimating partisan attention of news media sources using twitter data in the lead-up to 2018 italian election. *Comunicazione politica*, 20(1):85–108.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Sara Chennoufi, Hanene Maafi, and Thinhinane Hamitouche. 2020. Detecting hate speech against politicians in arabic community on social media. *International Journal of Web Information Systems*.

Rakshita Jain, Devanshi Goel, Prashant Sahu, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Profiling hate speech spreaders on twitter. In *CLEF*.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence*, pages 1621–1622.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110.

Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, et al. 2020. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources & Evaluation*, 55:477–523.

Francisco Rangel, GLDLP Sarracén, BERTa Chulvi, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling hate speech spreaders on twitter task at pan 2021. In *CLEF*.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Overview of the evalita 2020 second hate speech detection task (haspeede 2). In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439.

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

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Dejin Zhao and Mary Beth Rosson. 2009. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM.