Hate-Speech Detection using HurtLex lexicon for Arabic and Czech Languages

Miran M. Rashid
Salim Fares
Eva Nováková
mohamm11@ads.uni-passau.de
fares01@ads.uni-passau.de
novako04@ads.uni-passau.de

Figure 1

ACM Reference Format:
Miran M. Rashid, Salim Fares, and Eva Nováková . 2020. Hate-Speech Detection using HurtLex lexicon for Arabic and Czech Languages. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In 2010, fewer than 1 billion people were signed on. That number has since tripled according to USA Today Network. Social media, giving the impression of freedom of speech, allow people to share their opinions and feelings in immense speed and reach. This also brings negative consequences, as social media has become a powerful tool for spreading hateful opinions, often reflecting social and political tensions [1]. Many users are taking advantage of social media to express their hatred and hostility, with a possibility to influence other people, too. In addition, there is a strong connection between hate speech and hate crime [6], therefore it is strongly required to develop effective methods for automatic detection of hate speech. Popular social networks, like Twitter and Facebook, are facing criticism for not doing enough to prevent hate speech [2].

The definitions of **Hate Speech** (HS) slightly vary [1–7], but they unite in one point - hate speech is any communication that targets disadvantaged person, social group or minority in a way that is potentially harmful to the object or directly disparages it. This speech attack is typically based on characteristics such as gender, race, colour, ethnicity, sexual orientation, nationality, social status etc. More generally, HS [4] can be defined as any type of communication which is abusive, insulting, humiliating, intimidating, harassing or leads to violence or discrimination.

Although offensive language is very often a part of hate speech, the presence of an offensive language in a speech does not necessarily imply that this speech is a HS, which makes the auto-detection pretty complicated [2]. Nevertheless, more lexicons of offensive language and other expressions, that are not offensive themselves but still have an offensive potential, have been created and shown useful, e.g. HurtLex [1] or Hatebase.org. Using these lexicons, we can effectively detect potentially offensive speech or even HS [2].

HurtLex is a multilingual lexicon of categorised hate words (or words to hurt), that was created semi-automatically from an Italian hate lexicon[1]. HurtLex was created by extracting the lemmas from this lexicon and after that, with the help of MultiWordNet and BabelNet dictionaries, translated into over 50 other languages. Thanks to the dictionaries many of the lemmas maintained their semantic definition even after the translations, though many mistakes can only be fixed manually.

There are 17 different HS categories in HurtLex which can be used for detecting different kinds of hate speech. Every specific HS

Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] This lexicon was developed by Tullio De Mauro.

attack, such as misogyny, is mostly employing just a few chosen categories of HurtLex [4]. This gives a potential for varied applications of HurtLex.

2 TARGETS/PLAN

The main aim of our project is to exploit the knowledge of our native languages, Arabic and Czech, to improve their versions of HurtLex. Firstly, we will remove all irrelevant expressions (responsibly verifying our understanding), which mistakenly entered the lexicon due to the automatic part of its creation. Secondly, we wish to enhance the lexicons with expressions that don't have equivalent in Italian and were therefore not added. Our intention is to make the lexicons more authentic, corresponding to the real present-day languages, covering the vocabulary from different generations and social classes (teenage, adult, low intelligence, aggressive or racist individuals…). Thirdly, we will focus on the categories assigned to the lemmas in HurtLex and verify their correctness. Finally, following the instructions of $Dr.ValerioBasile$, the lead author of HurtLex, we will add a new column in the Czech HurtLex to enhance the lexicon with a feature called **offensiveness score**. In this column we will adjust a real value between 0 and 1 to each lemma from HurtLex, indicating its score for the offensiveness. To demonstrate the improvement of the HurtLex lexicons, we will test the original and improved lexicons on an annotated corpora.

Once the Lexicons are successfully improved, we will search for multi-word expressions - hurtful phrases and idioms, which also represent HS, and we will therefore add them into HurtLex lexicons.

Lexical detection methods usually have low precision in detecting HS [2], classifying every text containing terms from lexicon as a HS. We will try to show that the improved HurtLex can be used more broadly than for simple quantitative and qualitative evaluations. Following [2–4, 7], we will test several models to find the best one for this task.

For *Arabic*, we will use a corpus that contains tweets and the following annotations: the hostility type (column: tweet sentiment), hostility directness (column: directness), target attribute (column: target), target group (column: group) and, annotator's sentiment (column: annotator sentiment).

For *Czech*, there is no existing corpora annotated in context of HS, therefore we are currently trying to get a corpora using Twitter API. We might resort to annotating a corpora ourselves or at least using the corpora for a different purpose, for example as a source for more HS expressions.

REFERENCES

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

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media.*

[3] Hossam Faris, Ibrahim Aljarah, Maria Habib, and Pedro Castillo. 2020. Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. 453–460. https://doi.org/10.5220/0008954004530460

[4] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, Vol. 2150. CEUR-WS, 234–241.

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

[6] 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.* 88–93.

[7] Alžbeta ŠTAJEROVÁ. 2019. Automatizovaná detekce ofenzivního jazyka a nenávistných projevů v přirozeném jazyce. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Martin Fajčík.