Application-Oriented Approach for Detecting Cyberaggression in Social Media

Kurt Englmeier, Josiane Mothe

Application-oriented Approach for Detecting Cyberaggression in Social Media

Kurt Englmeier[1], Josiane Mothe[2]

[1] Schmalkalden University of Applied Science, Blechhammer, 98574 Schmalkalden, Germany
kurtenglmeier@acm.org
[2] Josiane Mothe, IRIT, UMR5505 CNRS, INSPE-UT2, Université de Toulouse, France
Josiane.mothe@irit.fr

Abstract. The paper discuses and demonstrates the use of named-entity recognition for automatic hate speech detection. Our approach also addresses the design of models to map storylines and social anchors. They provide valuable background information for the analysis and correct classification of the brief statements used in social media. Furthermore, named-entity recognition can help to tackle the specifics of the language style often used in hate tweets, a style that differs from regular language in deliberate and unintentional misspellings, strange abbreviations and interpunctuations, and the use of symbols.
We implemented a prototype for our approach that automatically analyzes tweets along storylines. It operates on a series of bags of words containing names of persons, locations, characteristic words for insults, threats, and phenomena reflected in social anchors. We demonstrate our approach using a collection of German tweets that address the vitally discussed topic “refugees” in Germany.

Keywords: Hate Speech Detection · Named-Entity Recognition · Social Anchor · Storyline.

1 Introduction

Detection of cyber-aggression and hate speech is still a complex task. It requires the careful analysis of a variety of human factors in language that reach beyond the words used in hate speech itself. Here, we give an impression to what extend named-entity recognition can be used in order to identify and classify regions of aggressive utterances in statements and to discriminate them against regions of profane utterances. In general, there is an actor creating aggressive statements that address a target person (prominent person or victim of cyberbullying, etc.) or group (refugees, Jewish people, Muslims, etc.).

We discuss our approach based on a collection of German tweets, mainly related to the topic “refugees”. We also explain the role and importance of an analysis along the storyline of tweets and of including information about social anchors as roots of storylines. The work presented here, including the prototype used for demonstration purposes. Nevertheless, it demonstrates the potential of named-entity recognition for

hate speech detection. All in all, we see our approach as useful complement for part-of-speech- or ontology-based strategies.

2 The Problem

Today, applied hate speech detection mainly relies on key word analysis. In fact, there are many comments that use outright and clearly visible offensive terms: “Ich bin dafür, dass wir die Gaskammern wieder öffnen und die ganze Brut da reinstecken. (I am in favor of opening the gas chambers again and putting the whole brood in there).” These and similar statements can be located easily in an automatic way. There are clear key words indicating offensive and inciting statements. The correct classification of this statement as hate speech is unquestionable, even if we consider it in isolation.
However, hate speech detection is more than just keyword spotting. The features to discover are manifold (type of language, sentiment, actor and target detection, and so on). Hate speech detection must also catch up with the specifics of the language applied in hate speech and the dynamic changes in our everyday language. The evolution of social phenomena and of our language makes it difficult to track all racial, abusive, sexual, and religious insults.
The language used in social media also has its own style. Many authors—in particular when emotionally agitated—don’t care or cannot care about correct spelling or punctuation. They use sometimes strange abbreviations or deliberately incorrect spelling to express their emotions or (much like in spam mails) to try to cheat automatic hate speech detection. Apart from the (sometimes) poor writing style, tweets also use references to background knowledge that needs to be taken into account in hate speech analysis.

The specifics of hate language start with these syntactic qualities that differ from regular texts such as in newspapers or books. Only in rare cases, the authors use outright offensive expressions. Sometimes, they try to “hide” their opinions and intentions by less obviously offensive terms. In many cases, offensive terms clearly address facets of events or phenomena, such as the holocaust for example, to indicate the author’s intention. Words, appearing innocuous in the first place, may reveal a clear act to stir up hate or to incite criminal acts after a closer look.

3 Our Approach

With our collection of tweets, we made the experience that

- outright offensive terms are much less used than we expected,
- tweets can only partly be classified in isolation, and
- we have to consider the complete storyline a tweet is embedded in.

We analyze storylines that started with a particular news trailing a series of comments reflecting opinions and opposing viewpoints. Only by viewing the whole storyline we are in the position to identify “toxic” words or expressions that look profane in the first place, but may refer to a context that emblematizes an aggressive or offensive act. Sadly, many such contexts reflect practices or methods of the Nazi regime.

Kaggle's Toxic Comment Classification Challenge differentiates six categories of toxicity that can be detected in hate speech: toxic, severe toxic, obscene, insult, identity hate and threat). The categories are not mutually exclusive. We add a further important category: inciting. Statements that incite others or intend to incite others to do a criminal act or to further propagate hate are among the most dangerous utterances in hate speech.

Named-entity recognition usually addresses the problem of extracting and classifying proper names in texts, such as names of people, organizations, or locations. In this context, an entity is an individual person, place, or thing in the world, while a mention is a phrase of text that refers to an entity using a proper name.

In the context of hate speech detection, named-entity recognition at first includes also extracting and classifying proper names of persons that are authors or targets of offense or aggression or names of locations that are focal points of hate inducing events. However, it also has to locate outright or disguised expressions of hate and offense.

In this paper, we concentrate on this aspect of named-entity recognition: We locate toxic terms and investigate their surroundings, their mentions, and classify them. In such a situation, terms like "train" or "stock car" may become toxic! Both words are not offensive in the first place. However, a mention like "We need again long trains for these refugees!" clearly refers to the trains that brought prisoners of all sorts to the concentration camps during the Nazi regime. The same holds for a phrase such as "Are there any stock cars left?" with the mentioning of refugees or supporters of refugees further up the storyline. In both cases, the mentions refer to the trains of extermination and propose the same fate for the target persons which the passengers of those trains met. Both words turn from "toxic" into "threat" or even "inciting" when considering their immediate surrounding and preceding storyline.

A toxic term or a set of toxic terms indicates the potential existence of a mention containing an offensive or aggressive act. However, here we have to be careful. Any sort of close negation can turn this potential into its contrary: "You are a fool!" (insult) vs. "I'm not such a fool and believe this story!" (profane). This is in particular the case in storylines that cover views and opposing viewpoints.

4 Related Work

Correctly detecting hate speech and discriminating it from humor or simply profane expressions is still a challenging task. Current approaches apply the full range of method established in text analysis, such as part-of-speech (POS), N-grams, dictionaries or bag-of-words (BOW), TF-IDF, sentiment detection, or ontology-based strategies.

In social media, humans use combinations of words, symbols (smiles etc.), and words that do not even exist in dictionaries. It is thus indispensable to learn significant expressions directly from the tweets. We incline our approach to the analysis of word N-grams [1], key-phrases [2], and linguistic features [3,4].

Many aggressive or offensive comments or posts originate from a certain event detailed in the news or in newspaper articles. Sometimes these comments are statements directly following a news. We may consider this news as anchor texts. In information

retrieval—in particular when analysis targets social media—, anchor texts are used as query replacements or query enhancements when authors refer to these texts by hashtags or links [5,6]. In contrast to traditional media that simply broadcast news, content in social media takes much more the form of a conversation or discourse. Lee and Croft [7] expand concept of anchor text further and consider texts that initiate conversation or discourse as social anchors. We believe that taking into account social anchors is indispensable for a correct interpretation of comments.

Example: "author_of_the_tweet: #kandel 8,5 Jahre Jugendstrafe für einen MORD! Wofür gab es die 1,5 Jahre Rabatt??? Ich kann gar nicht soviel fressen, wie ich kotzen möchte (#kandel 8.5 years of young custody for MURDER! What is the 1.5-year discount for??? I can't eat as much as I want to puke."

In this case, we consider "Kandel" as a social anchor. This includes first of all the anchor text, which can be one or even more news about the event and reports that follow up. A social anchor references a social phenomenon or event that is usually broadcasted by the news. In social media, there are one or more hashtags referring to discourses following this anchor event.

We take "Kandel" as the title of a social anchor that can be summarized, for example, by the key words (extracted from an anchor text) "event: fatal stabbing", "victim: German girl", "culprit: asylum seeker, refugee, charged with murder, jail: 8.5 years", "December 27, 2017", "Kandel, Germany". To achieve this summary, we may simply apply key word identification using TF/IDF or more sophisticated approaches for feature selection [8]. The example also shows that we probably have to collect more things than just key words. Much like in ontologies, there are qualities that further specify key items of the text.

A broad range of tweets in our collection indicate that we probably need a broader concept of social anchor. In particular far-right populists often refer indirectly to the cruelties of the Nazis, mainly things and acts related to the murdering in concentration camps. Therefore, words like "gas", "oven", "furnace", "freight train", "chimney", etc. are potentially toxic. Therefore, we need to treat the facets of the Nazi barbarism also as social anchors.

5 Feature Detection in Storylines

In the end, we want to identify actor, intent, target, and intensity (or polarity) in hate speech utterances: "I really disgust these people". By analyzing the sequence of utterances, we can link "people" with "refugees" if they are mentioned in close context beforehand. Surface features help to indicate the intent of the statement, too. It also helps to detect special stereotypes like (superiority of an actor or actor group) or the type of language (othering or discriminating language, e.g.).

We propose a supervised learning approach to identify the hate speech-related features [9]. The ultimate goal is the design of a hate speech detection based on a multi-layered feature extraction and learning algorithm.

We start with bags of words containing names of persons (including synonyms) and locations. We are aware that there are promising approaches to automatically identify

names of persons and locations in texts using conditional random fields, for instance [10]. Here, we collect relevant names manually. Further bags of words contain toxic and severe toxic expressions and words indicating negation. Severe toxic words usually stand for insults like “fool”, “scumbag”, “idiot” and the like. The most interesting bag of words is the one containing words or expressions that reflect obscene or inciting statements or indicate identity hate or threat. It also contains profane expressions that specify otherwise toxic words as expressions of obscenity, inciting, identity hate, or threat. Words like “fire” or “gas”, for instance, are considered toxic. Combined with “send to” or “into” the whole expression becomes aggressive and inciting when referring to a target person or group.

Much like many established approaches for hate speech detection we propose a learning process consisting of the following layers:

1. Cleansing obfuscated expressions, misspellings, typos and abbreviations.
2. Identification of toxic words or expressions (including word n-grams and key phrases) in the tweets along the storylines. Investigation of the proximity of these expressions to further specify the toxicity of these expressions. The obtained words can be new ones or synonym expressions.
3. Add suitable candidate words to existing bags of words.

The first step—the cleansing process—addresses toxic words that are intentionally or unintentionally misspelled or strangely abbreviated:

- “@ss”, “sh1t”, “glch 1ns feu er d@mit”, correct spelling: “gleich ins Feuer damit”: “[throw him/her/them] immediately into the fire”.
- “Wie lange darf der Dr*** hier noch morden?”: “How long may this sc*** still murder? “Dr***” stands for “Drecksack (scumbag)”.
- “… die kuropten Politiker die ieben in saus und braus.”, correct spelling: “… die korrupten Politiker, die leben in Saus und Braus”: “… the corrupt politicians, they live in clover”.

We recommend to apply distance metrics or character pattern recognition in this situation and to tag these expressions as named entities in order to achieve transparent forms of obfuscated, misspelled, or abbreviated terms.

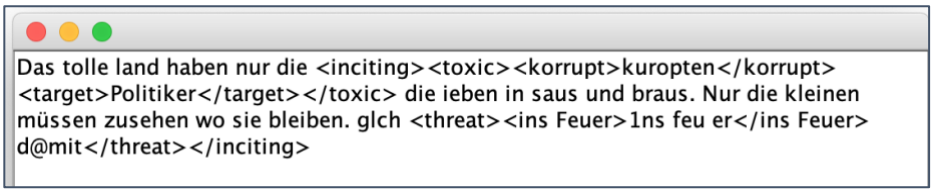

Fig. 1. Cleansing and tagging of a single tweet containing misspellings and toxic and inciting expressions as discussed in the example above.

The example of figure 1 shows a schema that addresses a target (“politicians”), one toxic expression (“corrupt politicians”) and one outright threat (“into the fire”). From

the style of the tweet, we probably assume that the toxic expression—as a general statement about politicians—is an insult. However, this is hard to determine in an automatic way without further information. In a country with a high level of corruption this statement even might be true. The close proximity of the toxic expression to the threat, that is, with only (presumably) profane expressions in between, clearly indicates an overall statement to incite somebody to do severe harm to politicians. Thus, we can conclude that the tweet has the character of being inciting. This conclusion can be achieved by the system in an automatic way. This schema works also for similar mentions when different targets addressed like a religious group, a minority, or a prominent person in conjunction with a threat. The threat in the example is to throw somebody (indicated by "damit") into the fire. The system will also indicate instances of similar patterns as "inciting" that mention different threats like "[send them] to the furnace".

The tweet of figure 1 can be classified as hate speech even without consideration of the preceding storyline the tweet is part of. However, there are cases when we need background information. Imagine the statement "send them by freight train to …" instead of "into the fire". "Freight train" in the context of hate speech has always a connotation with the holocaust. The cruelties of the Nazi regime provide important background information, we have to take into account in hate speech analysis. Sadly, each facet of these cruelties can be a social anchor, too.

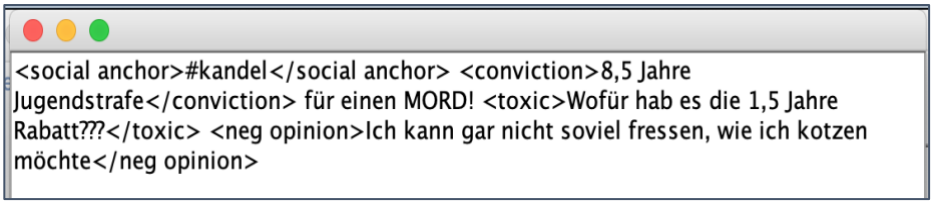

Fig. 2. Example of the analysis of a tweet with reference to the social anchor "Kandel".

With figure 2 we come back to our example of a social anchor "Kandel" as outlined above. The reference to this anchor with all its characteristics (facts) is important to correctly analyze this tweet. The anchor provides information on the crime of a refugee that sparked an intense social dispute the tweet is referring to. One mention of the tweet indicates a clear negative opinion. The close proximity to the fact (conviction) indicates the author's repudiation of the conviction. Multiple question marks are often used to express an opposite opinion to the fact rendered in the related phrase. Therefore, the system marks the expression as toxic.

However, for the system there are also limits: The tweet expresses a strong opposition against the court decision. The vulgar phrase indicates that, but also the sentence with the three question marks. In absence of the negative opinion, the question marks are the only weak signal pointing to the author's dismissive attitude. Of course, the term "discount" ("Rabatt") in the context of a judgment also reveals the author's objection. From the information we have so far, we cannot automatically infer any negative connotation of the word "discount".

6 Conclusion

In this paper, we gave an impression to what extend named-entity recognition can support automatic classification of hate speech in social media. The examples of offensive statements discussed here are quite typical for the ones we found in our collection of tweets. They also demonstrate that is hard to interpret and classify statements in the absence of social anchors. Even a storyline of a single author is often rooted in one or more social anchors. As long as we can retrieve sufficient information about these anchors, we are in the position to automatically and correctly detect semantic relationships that essentially support our classification process. From a particular author's storyline as series of her or his tweets we can deduce information on her or his attitude. However, there are limits. Many, probably important, utterances pass unnoticed the automatic process of hate speech detection if automatic hate speech detection systems lack the necessary context information. However, by indicating toxic terms or expressions we can support humans that fight against hate speech in social media. We can give them weak signals that point to offensive and aggressive language and make their work more efficient.

References

1. Ying, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and the 2012 International Conference on Social Computing, SocialCom 2012, pp. 71-80, Amsterdam, Netherlands (2012).
2. Mothe, J., Ramiandrisoa, F., and Rasolomanana, M.: Automatic Keyphrase Extraction Using Graph-based Methods. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 728–730 (2018).
3. Xu, J.-M., Jun, K.-S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 656–666 (2012).
4. Bollegala, D., Atanasov, V., Maehara, T., Kawarabayashi, K.-I.: ClassiNet—Predicting Missing Features for Short-Text Classification. ACM Transactions on Knowledge Discovery from Data (TKDD) 12(5), 1–29 (2018).
5. Anh, V.N., Moat, A.: The role of anchor text in clueweb09 retrieval. In Proc. of TREC, TREC'10 (2010).
6. Eiron, N., McCurley, K.S.: Analysis of anchor text for web search. In: Proceedings of SIGIR, SIGIR '03, pp. 459–460 (2003).
7. Lee, C.-J., Croft, W.B.: Incorporating social anchors for ad hoc retrieval. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 181–188 (2013).
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, pp. 1157–1182 (2003).
9. Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. de, Stringhini, G., Vakali, A., Kourtellis, N.: Detecting Cyberbulling and Cyberaggression in Social Media. ACM Transactions on the Web (TWEB) 13(3), pp. 1–51 (2019).
10. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields, Foundations and Trends in Machine Learning 4(4), pp. 267–373 (2012).