The Role of Storylines in Hate Speech Detection

Kurt Englmeier[1]

[1] Schmalkalden University of Applied Science, Blechhammer, 98574 Schmalkalden, Germany
kurtenglmeier@acm.org

Abstract. The paper explains why it is necessary to consider offensive statements in the context of their respective narratives if we want to achieve a more accurate classification of hate comments. By considering the narrative, text analysis is more sensitive to a person's affective state that, in turn, helps to reveal the true orientation of the statements in its close proximity. The approach present here is mainly built on named-entity recognition for identifying the different text features we can encounter in hate speech. First, the statements often exhibit a writing style that differs from a regular one in deliberate and unintentional misspellings, strange abbreviations and interpunctuations, and the use of symbols. Central to text analysis here is the identification of toxic terms that clearly evidence the offensive and aggressive character of each statement. However, the biggest challenge is the recognition of emotions and affective state of the writer. Analysis described here underpins the design of a prototype that operates on statements along storylines using a series of bags of words for names of persons, locations, and groups as well as insults, threats, and emotions. First results from this work-in-progress show hate speech analysis of German tweets that refer to the vitally discussed topic "refugees" in Germany.

Keywords: Hate Speech Detection, Named-Entity Recognition, Text Anchor, Social Anchoring, Storyline.

1 Introduction

Offensive or aggressive utterances are usually part of a narrative. They can be evident, clearly stand out from their surrounding context, and can thus easily be detected as such. Machines scouring the social media for hate speech do not have much problems to discern these utterances of hate. Equipped with the right list of offensive expressions a search engine has an easy job to pick out hate comments. Often, however, hate speech comes in subtly nuanced forms that is easily understandable for human readers, but hard to interpret correctly for machines. Without knowledge about the social and cultural background, about facts and events the hate is referring to automatic hate speech detection will bypass many offensive and aggressive utterances. Without considering the semantic surrounding of the overt expression of hate, the machine may even misinterpret the one or the other offensive utterance.

Hate speech is not isolated or independent from context. It is embedded in the narrative of a person. Her or his narrative joins narratives of further persons constituting

a discourse. This discourse, in turn, is embedded in a socio-cultural context and rooted in one or more facts emerged or events happened in the past. These sources are external to the discourse and influence meaning and understanding of each utterance in each narrative.

A storyline is a coherent sequence of utterances from mutual narratives that root in things like an event, fact, or statement. It has a timeline that, however, is only of minor importance. Nevertheless, it is time-bound, but only in the sense that its triggering cause happened at a certain point in time. The cause of the discourse (with all its characteristics) and the different persons authoring their respective narratives are the main structural elements of the storyline. The first goal of our approach is to map out the discourse along the storyline. The second goal is to determine heuristics for correctly classifying utterances of hate speech.

The approach presented here is based on a collection of German tweets, mainly related to the topic "refugees". It addresses the role and importance of an analysis of statements along the storyline including the anchor texts that triggered the narratives of the storylines. The work presented here, including the prototype used for demonstration purposes, is work in progress of a small group at the Schmalkalden University of Applied Science. Nevertheless, it demonstrates the potential of named-entity recognition for hate speech detection. All in all, it may be also a useful complement for part-of-speech- or ontology-based strategies.

2 Related Work

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: "[…] are an abomination and need to be helped to go straight to Hell!" However, hate speech detection is more than just keyword spotting. If we want to correctly classify hate speech as such, we have to apply an approach that includes the broader context of each utterance. That does not mean that basic methods of information retrieval are useless. They are essential, when it comes to spot indicators of aggressive and offensive expressions and corresponding emotions or the affective state of the author of the statement. It is thus indispensable to incline our approach to the analysis of word N-grams [1], key-phrases [2], and linguistic features [3,4].

Narratives containing offensive and aggressive language reveal many emotions alongside the leitmotiv of the authoring person. Murtagh and Ganz [5] developed helpful approaches to track emotions. Opinion mining on social media is quite prominent in computer and social science [6]. The focus on opinions in discourses is addressed, for example, [7]. Utterances expressing opinions and, in particular, hate quite often reveal emotions. Hate speech analysis, thus, must consider results and work in textual affect sensing [8, 9] alongside discourse analysis. [10] developed a framework for therapist-patient discourse that is valuable in our context. His work has been summarized and discussed in [11].

Why is it so important to analyze the narratives? Firstly, many aggressive utterances pass undetected if we consider utterances in isolation. For example, the comment

"Person A: There are so many Muslims in our town." is neutral at first. However, the following comment "Person B: I hate them all!" reveals the hateful attitude of this person without doubt. On the other hand, we cannot automatically classify the second statement as hate speech without considering its surrounding. Just imagine the first statement was "Person A: There are so many snails in my garden." Then the second statement hardly qualifies for hate speech. It is important to know who or what is addressed by "them".

3 A Glance on Patterns of Aggressive Narratives

Text analysis as outlined here addresses statements emerging from events that triggered upset in diverse social media channels. The sources considered are tweets or comments that refer to the so-called refugee crisis in Germany, in general, and to specific events with refugees involved. News on such events trail aggressive or offensive comments or posts in the respective newspaper itself (mostly right-wing newspapers and channels) and further channels where the news has been re-published. In contrast to traditional media that simply broadcast news, narratives in social media form much more a discourse (or controversy) emerging from the event it is reflecting. News triggering a discourse or controversy has the role of an anchor text.

The narratives of the persons contributing to a discourse root in the anchor text, but also in their individual social and cultural context. This is important, because terms like "train" or "stock car" may sound trivial at first. However, in a mention like "We need again long trains for these refugees." it 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?". In both cases, the mentions refer to the trains of extermination and propose the same fate for the targets of their aggressive comments.

One of the discourses in our collection rooted in a fatal crime committed by a young refugee that afterwards has been sentenced for murder. The news about this crime is the anchor text, which may be expanded by one or even more news about follow-up events like the conviction. The different narratives emerging from that text express the repudiation of the political and justice system in Germany and great parts of the German society. Primarily, they expose a deep and undiscriminating rejection of all refugees, but in particular of these having the same nationality as the young offender. The negative and aggressive narratives also depict a clear picture of the debaters' social anchoring [12] that reflects their mental foothold gained from the world view of partisans of right-wing ideology. In that, their anchoring evidences their incapability to make accurate and independent judgements. The following statements are typical for this controversy.

Statement 1: "… #kandel1 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

1 "Kandel" is name of town in Germany where the crime happened.

(#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."

The debater repudiates the conviction as being too mild and emphasizes the rejection with a strong negative emotion. This emotion ("I can't …") punctuates the person's attitude in this issue. We can expect that it tends to cover also closely related aspects of this conviction, namely the nationality of the young criminal, the refugees, in general, and Germany's justice system.

Statement 2: "Wie viele Frauen müssen noch ermordet und vergewaltigt werden, bis unser Volk aufwacht und den Politikern Feuer unterm Hintern macht? ("How many women need to be murdered and raped until our nation awakes and makes fire under the politicians' bum". In this statement addressing the same case, the debater focuses on German politics that needs to be stirred into a different direction, by violence if necessary. This statement is of interest because reveals a certain openness of the debater for a violent, political change.

Statement 3: "… die afghanischen Frauen [sind] "Besitztum" der Männer, dürfen geschlagen werden, ohne Strafen zu fürchten." ("… Afghan women are in "possession" of men, may be beaten without fearing any penalty"). This is an example of a typical bias revealing inability or lack of willingness to differentiate between the individuals of a certain group. The comment may also represent an instance of social anchoring: the debater's intent to influence the audience towards a discriminatory stance.

The examples also suggest that we probably have to include cultural anchors, too. Far-right populists and their partisans often use terms and expressions borrowed from the cruelties of the Nazi regime, mainly things and acts related to the murdering in concentration camps. Therefore, words like "gas", "oven", "furnace", "freight train", "chimney", etc. are part of their aggressive language.

4 The Phases of Feature Extraction

In the end, text analysis wants to identify the debater or author of the narrative, target persons or groups, and the debater's leitmotiv (desires, need, and intents) and emotions. To identify the debater's narrative along the storyline is easy. The (real or fake) name of the author is one the few structural elements in tweets and similar messages beside the timestamp. The anchor text can be described using its key terms with or without annotations. We may take "Kandel" as the title of the anchor text and the following key words and annotations for its description: "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 [13]. The example also shows that we probably have to collect more things than just key words. Much like in ontologies, we should further qualify the identified key items by more general concepts. That, in turn, leads to a more general description of the event. Hate speech-related text features are probably best detected along a supervised learning process [14] over a series of phases:

1. Identifying structural elements of the discourse, its time frame, anchor text, and the different narratives of the debaters.
2. Cleansing obfuscated expressions, misspellings, typos and abbreviations by applying character patterns and distance metrics.
3. Application of different bags of words to locate mentions of persons, groups, locations etc.
4. Identifying outright discriminating, offensive, and aggressive terms.
5. Identifying emotions and measuring the affective state.
6. Measuring the toxicity of individual statements and narratives.

The process of phase 1 yields a linked list containing the individual statements with their time stamps and pointer to its author and anchor text.

Phase 2: The next step, the cleansing process, addresses terms 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)".

Phase 3: Text analysis uses here bags of words containing names of persons, locations, prominent groups, parties, and the like (including synonyms), even though there exist promising approaches for automatically identifying names of in texts based conditional random fields, for instance [15]. Bags of words for the prototype presented here are produced manually.

Phase 4: Further bags of words contain toxic terms [16], that is, discriminating, offensive, and aggressive expressions and words ("fool", "scumbag", "idiot" and the like). Words like "fire" or "gas", for instance, are also considered toxic if they appear in combination with "send to" or "into" and refer to a target person or group. Initially, we may consider any occurrence of such a term as potentially toxic. The toxicity is approved if no immediate negation reverses the polarity of the expression.

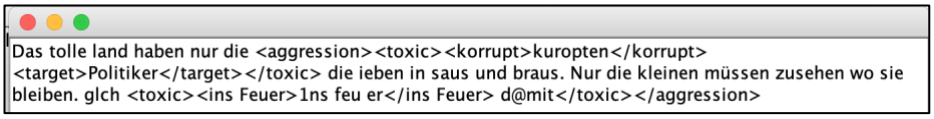

Fig. 1. The potentially toxic expression ("corrupt politicians") turns the initially profane expression ("into the fire") into an aggressive statement.

The example of figure 1 shows how two potentially toxic expressions turn the statement into an aggressive one. The close proximity of the toxic expression to the threat, that is, with only (presumably) profane expressions in between, clearly indicates the author's wish to do severe harm to politicians. 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 example also shows some typical misspellings or intentional typing errors.

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. This background is just as important as the anchor text.

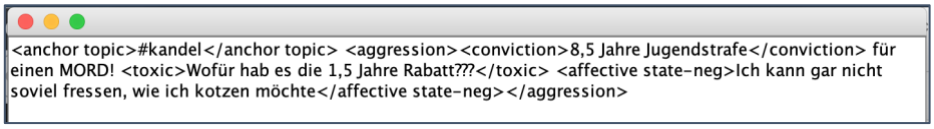

Fig. 2. Example of an expression of a negative affective state

Phase 5: In hate speech, we encounter many expressions of positive or negative emotions. These expressions are an important indicator of the overall affective state of the author in relationship to the discourse or the facts as described in the anchor text. The last phrase in figure 2 (“I can't eat as much as I want to puke.”) insinuates a negative affective state of the author. The reference to the anchor text addressing the details of this event is important for the correct classification of this tweet. The anchor text (“Kandel”) provides information on the crime of the young offender and his conviction. The close proximity of the fact to the author’s negative affective state reveals her or his repudiation of the conviction. We may take the affective state as a special indicator that puts a negative or positive impact on its surrounding, which can be toxic statements or facts from the anchor text or the immediate statements from the other debaters.

Phase 6: The final measurement of the toxicity combines the evaluations obtained from individual statements with related affective states.

5 Conclusion

A series of processes had been described for automatic classification of hate speech in social media. Even though the prototype that implements these processes represents work in progress, it demonstrates the usefulness of interpreting and classifying statements along the discourse storyline. The system cleanses the text and spots outright offensive and aggressive terms and the names of persons, locations, etc. Furthermore, from the author’s statements we can deduce her or his attitude and affective state that is important when it comes to correctly classifying statements in a broader context. In the long run, it will be beneficial if we can depict a more comprehensive and precise picture of social anchoring in this context.

The objective of the system is to support humans that fight against hate speech in social media. Systems can automatically perform a lot of the mundane detecting work, but there are limits. Assisting humans in this cumbersome work must also be part of the design of hate speech detection machines.

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. Murtagh, F., Ganz, A.: Pattern Recognition in Narrative: Tracking Emotional Expression in Context. Journal of Data Mining & Digital Humanities (2015).
6. Wright, A.: Our Sentiments, Exactly. Communications of the ACM 52 (4), pp. 14-15 (2009).
7. Dietz, L., Wang, Z., Huston, S., Croft, W.B.: Retrieving Opinions from Discussion Forums. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp 1225–1228 (2013).
8. Liu, H., Lieberman, H., Selker, T.: A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th international conference on Intelligent user interfaces, pp 125-132 (2003).
9. Neviarouskaya A., Prendinger H., Ishizuka M.: Textual Affect Sensing for Sociable and Expressive Online Communication. In: Paiva A.C.R., Prada R., Picard R.W. (eds) Affective Computing and Intelligent Interaction. ACII 2007. Lecture Notes in Computer Science, vol. 4738 (2007).
10. Schneider, P.: Language usage and social action in the psychoanalytic encounter: discourse analysis of a therapy session fragment. Language and Psychoanalysis, 2 (1), pp. 4-19 (2013).
11. Murtagh, F.: Mathematical Representations of Matte Blanco's Bi-Logic, based on Metric Space and Ultrametric or Hierarchical Topology: Towards Practical Application. Language and Psychanalysis, 3 (2), pp. 40-63 (2014).
12. Meub, L., Proeger, T.E.: Anchoring in Social Context, Journal of Behavioral and Experimental Economics (55), 2015, pp. 29-39.
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, pp. 1157–1182 (2003).
14. 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).
15. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields, Foundations and Trends in Machine Learning 4(4), pp. 267–373 (2012).
16. Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional Neural Networks for Toxic Comment Classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp. 1–6 (2018).