Application-oriented approach to hate speech detection.

Detection of hate speech is still a complex task. It requires the careful analysis of a variety of related factors that mostly reach beyond the ones covered by the utterances in the hate speech itself. In general, it is an actor that creates hate speech addressing a target person (a particular, mostly prominent person) or group (refugees, jewish people, partisans of the liberalisation of abortions, etc.). Hate speech usually pursues one or more purposes ranging from discrimination over flaming to extremism and radicalisation.

Correctly detecting hate speech and discriminating it from humour 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.

We are convinced that the correct detection of hate-related features in texts requires a bunch of analysis methods. Detection of significant keywords and named entities and their use patterns in text support the first classification of candidates for further analysis. Pattern analysis here explicitly includes obfuscation detection when offensive words are intentionally misspelled, like “@ss” or “sh1t”. We recommend to apply distance metrics or word pattern recognition in this situation and to tag these expressions as named entities to achieve transparent forms of obfuscated terms. Named entity recognition also serves to detect location- and time-related information (including URLs etc.) supporting the correct identification of actors and targets.

TF/IDF together with N-grams can help to detect significant keywords that constitute hate dictionaries (or bag-of-words) separated along the main hate-related topics such as religion, gender, ethnicity, disability, and the like.

Part-of-speech (POW) combined with dictionaries plays a decisive role. It detects **actor**, **intent**, **target**, and **intensity** in hate speech utterances: “I really disgust these people”. By analysing the sequence of utterances, POW links “people” with “refugees” if they are mentioned in close context beforehand. POW is the essential method to indicate the intent of the statement. It also helps to detect special stereotypes like (superiority of an actor or actor group) or the type of language (othering language, e.g.).

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 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.

Kaggle’s Toxic Comment Classification Challenge[1] 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.

We propose a **supervised learning approach** to identify the hate speech-related features we outlined so far above. The ultimative goal is the design of an hate speech detection application based on a multi-layered feature extraction and learning algorithm.

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

1) Seed words of keywords (including word N-grams) and named entities reflecting offensive, obscene, and inciting expressions, the intent of the hate speech. Similar bags of seed words are used for well known targets and intensity expressions. Named entities mainly address obfuscated expressions and abbreviations.
2) Investigation of the proximity of the seeds to identify further candidate words for the actor, target, and intensity in the utterance.

[1] https://www.kaggle.com/c/

3) We add suitable candidate words add to existing seed bags. The obtained words can be new ones or synonym expressions. For instance, “die Bundeskanzlerin”, “Merkel”, or “Angie” refer to same person.

The **classification process** consists of the following layers:

1) Preprocessing of the text: stop word elimination and named entity identification of obfuscated terms and abbreviations. We bypass lemmatization, because it clears too much useful information from the text.
2) The seeds are used to identify significant word patterns in texts. The higher the density of offensive or inciting words in one phrase or a sequence of consecutive phrases the more confident is the classification of this text passage as hate speech (offensive or inciting). All other phrases are considered to represent profanities.
3) If profane statements are combined with hate utterances, we investigate the polarity of the profane statement in order to correctly identify opposing statements to hate speech (“I don’t understand people who hate refugees”).

POW supports the identification of actors and targets across text passages.

The learning and classification process still represent work in progress.