Exploring Dimensions of Chinese Hate Speech on Twitter: A Feature Selection Task

Wai Ching Leung
Georgetown University
Wl607@georgetown.edu

Abstract

With the prevalence of hate speech online, hate speech detection has become an important task in the research community. This paper examines different types of features that fall into the following 5 categories: semantic features, sentiments, lexical features and linguistic features, and concludes that the most relevant features are associated to semantic similarity and profanity. The least indicative features are linguistic features: punctuations, sentence final articles tonal softening particles. In particular, the language of interest is Chinese (Simplified and Traditional) due to the limited research in this topic in Chinese.

1 Objective

Most of the previous studies on hate speech use similar features, such as n-grams, sentiments and part-of-speech tags, to train their models. Some studies (Gröndahl et al. 2018, Arango et al., 2019) found out that all of the state-of-the-art models do not generalize well to other datasets. This is associated to multiple factors: Overfitting, sampling bias, inconsistent labelling, and the lack of quality features.

The objective of this paper is to explore features based on the bag-of-word assumption. Although hate speech patterns go beyond surface form, for which, for example, external world knowledge and the ordering of words matter (MacAveney et al., 2019), surface features have been found to be useful for detecting abusive speech (Nobata et al., 2016), especially when used together with other types of features. Either presence or absence of a correlation of such features to hate speech can shed light on the characteristics of hate speech, and can be used to leverage future supervised classification task.

Since most of the hate speech studies focus on English hate speech, one goal of this study is to investigate whether features that are commonly used across different studies for English hate speech can apply to Chinese hate speech well. On top of the commonly used features, we will also explore some Chinese-specific features based on my understanding of the knowledge as well as my preliminary observation of the data collected.

2 Data

The dataset used contains a total of 4447 tweets. Out of all the tweets, 1240 of them are labeled as hate speech, the rest are labeled as non-hate speech. In order to retrieve relevant tweets, I built a Chinese hate speech keyword lexicon based on words listed on a Wikipedia page [1] that compiles derogatory terms. I filtered out terms that are not commonly used or outdated. In total, I collected 44 keywords that are often used describe a group of people based on religions, race, ethnicity. Since sexist derogatory terms in Chinese are often used in offensive speech, but not hate speech, they are therefore omitted in my research. This approach is aligned with the finding that most English sexist

[1] https://zh.wikipedia.org/wiki/%E6%AD%A7%E8%A7%86%E8%AF%AD

words are offensive only (Davidson et al., 2017). For the research purpose of this project, tweets are labeled as HS (hate speech) or NHS (non-hate speech). Although multinomial classification has been found to achieve better performance and capture nuances that could avoid common errors and confusion among annotators (Founta et al., 2018), I have simplified into 2 labels to ensure sufficient features can be captured, given the small size of the dataset.

3 Approach Background

In this section, I will provide a background information for the features I use for Chinese hate speech.

Semantic features: Many past studies on hate speech detection use word embeddings to assign similarity between training data and target data (Samghabbadi et al., 2017; Kshirsagar et al., 2018). A common approach to represent a sentence with vector is by averaging the word embeddings within it.

Keyword features: Waseem & Hovy (2016) found that the most predictive features in their task of hate speech detection are features that are associated to frequent words. In particular, profane and offensive words have been found to be useful features to classify hate speech. In fact, many studies create hate speech lexicon which consists of words that are often abusive or profane (Basile et al., 2019). While the presence of profane terms could be potentially a good indicator for hate speech, they have found heavy reliance of keywords could lead to high rate of false positives and false negatives (Davidson et al., 2017).

Sentiment features: Since hateful speakers often convey their speech with usually strongly negative emotions or aggressiveness, sentiment analysis is a common approach in hate speech classification task. Bauwelinck & Lefer (2019) looked into the performance of sentiment features for detecting hate speech, and concludes that sentiment features combined with surface features give the highest performance.

Linguistic features: Surface level and grammatical features are often used together with other features. Clarke & Grieve (2017) examine relevance of a great amount of linguistic features derived from part of speech, syntactic structures and communication styles for abusive language. They identify over 100 features that could potentially differentiate between racist and sexist tweets. For example, they look at the presence of question mark, preposition, different types of pronouns, etc. Many other features such as sentence length, capitalizations and the number of pronouns are also used (Schmidt & Wiegand 2017). Alorainy et al. (2019) examines the number of othering language, such as *'us-them'* for hate speech detection based on the *Othering and Intergroup Threat Theory* (OITT). They found out that incorporating othering features improves classification performance.

4 Methodology

Based on the previous studies and the linguistic patterns of Chinese, I propose that the following 10 features are associated to Chinese hate speech.

Similarity feature: In order to capture the semantics of hate speech, H.S tweets are clustered using word embeddings that are trained with Tencent AI Lab Embedding Corpus[2]. This is a 200 dimensional space with over 8 millions words or phrases. I observed the clusters and selected 4 out of the 8 clusters created. The clusters removed primarily contain irrelevant words, such as stop words, auxiliary verbs, pronouns or words that contain neutral sentiment, such as '*week'* or *'understand'*. Although preprocessing was done to remove non-content words, many function words still exist. Each word of a tweet is then compared against words in these cluster in terms of similarity. A similarity score 0.9 or above indicates that they are similar.

Out-of-vocabulary counts: Since hateful speakers often coin new words to express hateful comments or modify hateful expressions into a different form to evade detection, the number of out of vocabulary words could be a good indicator

[2] https://ai.tencent.com/ailab/nlp/en/embedding.html

of hate speech. The count is normalized by sentence length since longer sentences are more likely to have more oov words.

Punctuations: From my observation of hateful tweets in Chinese, I have notice that Chinese speakers online tend to use exclamation mark to emphasize their points of view or to express strong emotions. Besides directing hate speech in an aggressive way, they also use sarcasm or questioning to diminish their victims. Such a tactics is often accompanied by a question mark. Strong emotions can sometimes be reflected by the consecutive use of question mark or exclamation mark. Therefore, counting these two punctuations might be able to reflect expressiveness of a speaker, which might help differentiate HS vs non HS.

Sentence length: Intuitively, hateful speakers tend to be more expressive but has less content in their speech. A speaker who expresses their negative opinions towards a group of people with sufficient reasoning is often not considered a hateful speaker since hate speech also depends on the intention of speakers. Therefore, it is predicted that non hateful tweets are longer than hateful tweet on average.

Sentence-final particles: In Chinses, there is a unique category of words call 'sentence final particle'. They appear only at the end of the sentence, and do not carry any lexical meaning. One of the functions they have is to indicate mood/tone of speakers. The particles '*吧*','*呢*','*哦*','*啊*','*啦*' can be used to soften tone or express politeness. The number of these particles in a tweet is therefore expected to be negatively correlated to hate speech.

Assertiveness: Hateful speakers make over-generalization of a group of people, and/or appear to be more assertive when making a statement. Therefore, I compiled a list of 21 terms/expressions that could indicate over-generalization or assertiveness. For example, '全是' ('all are'), '最' ('all are')

Profanity proximity: Words that are closer together tend to have closer semantic relationship. For example, the subject or object of a verb is usually within 2 word-gram in Chinese. I assume that if an offensive expression is close to a proper noun, it has a higher likelihood to contain hateful elements. This is a binary feature to determine whether profane words appear within 5-word window before or after a proper noun.

Profanity count: Although profanity count often leads to false positives and false negatives, counting the number of offensive/abusive words can still be a good feature given their relatively high association to hate speech.

Sentiments: Sentiment polarity calculated to pure counts of negative or positive sentiment words collected from NTUSD: National Taiwan University Semantic Dictionary. The positive library has 2812 terms, and the negative library has 8276 terms.

Othering Language: To capture othering language in a tweet, I create a list of second and third person pronouns. The OITT theory states that the first person pronouns combined with second or third person pronouns is a sign of othering language. From my observation of hateful tweets, first person pronouns are often omitted as a way for speakers to distance themselves from their victims or the statements. Therefore, only second and third person pronouns are counted as a feature. Besides gender, Chinese pronouns also differentiate between objects and animals. For example 它 refers to an object *it*, while 牠 refers to an animal *it.* Since Chinese speakers often belittle their victims by calling them animals terms or denying their human nature, the object and animal pronouns are also included.

The above features are fed into mutual information and logistic regression models as either binary or relative frequency features, which is done with a Sklearn feature section model. For logistic regression, I used the entire dataset for training and testing the model due to small size of the dataset for which it would be difficult to split into a training and a development/test set.

Features	Score
Similarity	0.039
Profanity Words	0.020
Sentiment	0.018
Out of Vocabulary	0.013
Sentence Length	0.012
Profanity Proximity	0.010
Othering Language	0.0097
Punctuation	0.0054
Particles	0.0041
Strong Tone Words	0.0019

Table 1: Feature Mutual Information Score

5 Results

Table 1 shows the mutual information scores of the 10 features in a descending order. While the non-zero scores indicate they features and the classes are not completely independent, they all have a very low value.

The similarity and profanity features have the best performance among other, scoring 0.039 and 0.02 respectively. The feature that indicates the frequency of strong tone words shows the least dependency with the class.

The baseline model I use is the model from Davidson et al. (2017) because they also reuse a dataset for training and testing, and the features used share some similarity with the features of interests in my research. Their model yields a precision of 0.91, recall of 0.90 and F1 score of 0.90. However, only 5% of the data is labelled as hate speech, which is very different from the distribution of my dataset. The HS Vs non HS ratio is 1:2.5. Therefore, it might be more telling to look at the scores of their H.S class: 0.44 precision score , and 0.61 recall score. My logistic regression model show that the precision score is 0.569, the recall score is 0.135 and the f1 score is 0.219. If only compared with the hate speech classification from the baseline model, my model performs better in terms of precision, but lower in recall.

6 Discussion

Despite all the features have low mutual information scores, it does not mean they are not useful for hate speech classification. Since mutual information compares the dependency of each individual feature with the classes, but linguistic features are often more indicative as a combination for features for classification. Therefore, future work should considering feature elimination methods to weigh the importance of each feature.

The low scoring results indicate that surface level features might not be sufficient to capture hate speech characteristics. For example, while sentiment could serve as a good indicator hate speech, simply looking at the count of words that are associated to negative/positive sentiments is too simplistic. Besides, the sentiment lexicon is unable to capture many words that exhibits strong emotions due to domain discrepancy. For example, the phrase '*傻叼，港灿扑街吧*' is highly negative with 3 words being offensive out of the total 4 words.

It is not surprising that the similarity and the profanity features are the strongest indicators for hate speech since offensive or aggressive words are often used in hate speech. However, this approach might give rise to generalizability problem particularly for this project. The tweets were all retrieved by searching certain hate-group terms, such as the Chinese equivalents of *black people* or *n*g***, as a way to narrow down topics, so a combination of hate-group term + offensive words is more indicative than just offensive words alone. Future work expanding the dataset which includes tweets that do not contain hate groups terms might capture more relevant, generalizable keywords for hate speech.

Looking at the results from the out-of-vocabulary feature, the reason for its low performance could partly contribute to domain and segmentation problems. Although the Chinese word embedding corpus contains millions of words and they were collected from a wide variety of sources, it still fails to capture many informal, relatively new words. Therefore, oov might serve to identify tweets which often contains novel expressions than tweets that contain hate speech. Besides, since Chinese words are not separated by space as in a

lot of Romanized languages, segmentation errors could also cause a cascading effect to affect the out of vocabulary rates.

Future work: Although the results are not satisfactory, they do give me some insights and understanding of Chinese hate speech, and challenges working with this language. I have planned to expand my dataset and extract deeper features that might detect hate speech better.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on the Web and Social Media, ICWSM '17, pages 512-515.*

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, San Diego, California. Association for Computational Linguistics.

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 *AAAI International Conference on Web and Social Media (ICWSM).*

Valerio Basile, Cristina Bosco, Elisabeth Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chikashi Nobata, Joe Tretrault, Achint Thomas, Yasher Mehdad, Yi Chang. 2016. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web, pages 145 – 153. https://doi.org/10.1145/2872427.2883062.

MacAvaney S, Yao H-R, E, Russell K, Goharian N, Frieder O. 2019 Hate speech detection: Challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152

Isobelle Clarke, Jack Grieve. 2017. Dimensions of Abusive Language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1 – 10. https://www.aclweb.org/anthology/W17-3001

Tommi Gröndahl, Pajola Luca, Mika Juuti, Mauro Conti, Asokan N. 2018. All you need is 'love' : Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2 - 12. https://doi.org/10.1145/3270101.3270103

Rohan Kshirsagar, Tyus Cukuvac, Kathleen Mckeown, Susan Mcgregor. 2018. Predictive embeddings for hate speech detection on Twitter. in Abusive Language Online Workshop, EMNLP 2018. https://arxiv.org/pdf/1809.10644.pdf

Nina Bauwelinck, Els Lefever. 2018. Measuring the impact for hate speech detection on Twitter. In The Fifth Conference on Human and Social Analytics, pages 17 - 22.

Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, Thamar Solorio. 2017. Detecting nastiness in Social Media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72. https://www.aclweb.org/anthology/W17-3010.pdf

Anna Schmidt and M. 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*, pages 1–10. https://www.aclweb.org/anthology/W17-1101.pdf