Natural Language Processing Project Proposal: Automated Hate Speech Detection

Xiaoyan HU
x.hu@mpp.hertie-school.org

1. Introduction[1]

The popularity of social media provides more means of connection between humans. However, in the thousands of contents posted on them, each second, hate speech and hateful contents also travel like viruses. For example, video footage from the suspect of the 2019 terror attack in Christchurch, New Zealand, was broadcast live on Facebook.

While there is no formal consensus on the definition of hate speech and how it differs from other abusive or aggressive forms of language, this project adopts the definition given by European Commission against Racism and Intolerance (ECRI): Hate speech covers many forms of expressions which spread, incite, promote or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons. Derived from this definition, we can observe that there are many possible typologies of hate speech. Inspired by a project demonstration of Microsoft Research[1] on aggressive speech detection, we sort four types of hate speech based on the targets of aggression: 1) physical threat; 2) sexual threat/aggression; 3) identity threat/aggression; 4) non-threatening aggression. Some of those verbal aggression are overt while there also exists cover aggression which does not attack the victims directly and often include sarcasm and satirical attack. Understanding potential typologies of hate speech will help us to design or select appropriate labels for detecting them among unstructured corpus.

Due to the harmful social impact of hate speech content, leading social network companies have designed policies to remove those content. Currently, one conventional way of keeping the platform "clean" is to hire human moderators according to a feature from online media. Facebook is reported to have hired Cognizant to remove those tagged content manually[2]. There are several drawbacks in this practice. While the contractors are working under enormous pressure and invest many efforts, their efficiency cannot compare to the explosive increase of hateful content on Facebook. Furthermore, the moderators must cope with the blatant exposure to traumatic images, out-cry racism, violent xenophobia, and other aggressive content, which foster PTSD-like symptoms when they finish working. Thus, this manual labour is not sustainable and should be replaced by more efficient and less costly technological approach.

So, researchers decided to try an alternative: automatic intervention. Researchers run their data sets and intervening slogan databases through multiple machine learning and natural language processing systems to create AI prototypes that intervene in hate speech on the Internet. Ideally, this automated moderator that could automatically flag and categories aggressive content on social media. Ideally, it could also distinguish in between ratified and unratified aggression depending on the platform, user preferences and several other criteria.

However, at present, AI is not fully prepared. In theory, the system should be able to detect hate speech and immediately send a message to inform the publisher that its content contains visible hate speech. This requires not only keyword detection, but also AI's correct understanding of speech content. Thus, several possible errors are common in automated hate speech detection. For example, an algorithm could fail to detect hate speech which demands societal knowledge. Many learning models also cannot catch some context-based aggression. Imbalance sample can introduce biases in the automation.

The border between hate-speech and freedom of speech is very delicate and trespassing the line will not only harm the healthiness of democracy, and sometimes create an embarrassing scenario. Recently, due to epidemic control, many Chinese schools moved their classrooms online. One teacher of obstetrics and gynaecology was using an online chatting app, QQ, to deliver the lecture but received a prohibition notice no sooner than she started the class. Considering the online censorship against pornographic content in China, she reflected the reasons for this instant prohibition later that probably because she mentioned a few terminologies about external genitalia when she tried to explain the anatomy. This anecdote is indeed far from being an example of hate-speech. However, hate speech detection is fun-

[1] GitHub page link:https://github.com/Xiaoyan-Adele/Automated-Hate-Speech-Detection

damentally an activity of censoring/filtering for a different purpose. Thus, we can reflect on the potential over-reaction of algorithm, discuss possible ways to improve the robustness, or maybe acknowledge the limitation of the automation.

2. Motivation

Recent years, due to the collective reflection on cyberbullying and platform responsibility, there are ample research and funding into this field and thus make hate speech detection a very dynamic area. The biggest motivation of this project is to understand the mechanism behind automated hate speech detection by building a detector oneself. With a growing number of more accurate solutions and more precise labels, it is also meaningful to compare some of the existing models and understand the state-of-art.

3. Evaluation

Character-level bag-of-words has been proven effective in many types of research on hate speech detection. This paper decides to use it as the baseline of our research. The successful completion of this project involved the use of automatic evaluation metrics what yield a result that outperforms the baseline model in terms of accuracy precision, recall and F-Score.

- Accuracy: the percentage of texts that were predicted with the correct topic.
- Precision: the percentage of texts the model got right out of the total number of texts that it predicted for a given topic.
- Recall: the percentage of texts the model predicted for a given topic out of the total number of texts it should have predicted for that topic.
- F1 Score: the harmonic mean of precision and recall.

4. Resources

The field of hate-speech automatic detection and classification has evolved rapidly in the past years. As mentioned in the introduction, while the outspread of social media allows hate speech to travel like viruses, it also has completely changed the amount and form of information acquisition for today's natural language processing researchers. The samples scrapped from Twitter, Facebook, YouTube, blogs and forums can help to study the formation of hate speech and design appropriate models for automated detection. However, the abundance of raw data does automatically make it easier for supervised learning in hate speech detection. As the mechanism indicated below, hate speech detection model training requires annotated data which can be costly to obtain if this project wants to build from scratch.

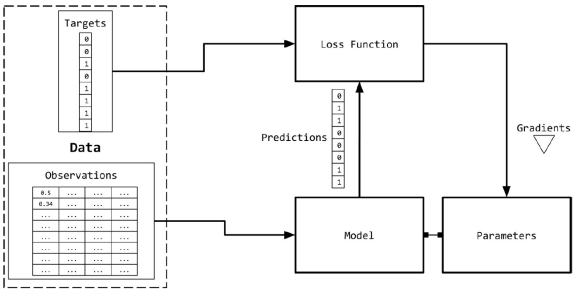

Figure 1-2. Observation and target encoding: The targets and observations from Figure 1-1 are represented numerically as vectors, or tensors. This is collectively known as input "encoding."

4.1. Open Dataset

Luckily, we already have many sufficiently sized and labelled English datasets from the website (http://hatespeechdata.com/). Among the available options, this project will request the dataset from the paper Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. It contains roughly 100k rows, where every row is consisted of a unique Tweet ID and its according to majority annotation and is created with the crowdsourcing methodology which should be the "golden standard" for testing and validating the model built. Comparing to other datasets, another advantage of this data is that it covers a broader range of labels, including offensive, abusive, hateful speech, aggressive, cyberbullying, spam, and normal with more annotators. However, one needs to contact the authors through email to obtain the dataset. If unfortunately, this project cannot get access to this dataset timely, we also found a ready-to-use alternative from the paper Automated Hate Speech Detection and the Problem of Offensive Language. The second dataset is also manually coded by CrowdFlower workers but has a much smaller size of 25k. The author of this project remains thankful to those scholars who decided to share their dataset on the open web.

4.2. Existing Models

Active research into the field of hate speech detection also offer us a variety of different models. MacAvaney et al (2019)[4] provided an overview of the mainstream approaches for automatic detection. One basic method is keyword-based approaches. Although it is fast and straightforward to understand, it has severe limitations. Key words sometimes cannot identify harmful content that is not included in its dictionary and even create false alarms when it encounters words that can be not harmful in certain scenarios. The misunderstood pornography alert anecdote mentioned in the introduction of this project proposal could be one of the negative example of key word approach. Machine learning classifiers (such as Naïve Bayes, Support

Vector Machine and Logistic Regression) are also popular models for topic labeling and it has a rather wide range of use, applying beyond hate speech detection. Open-source implementations of the these models exist, for instance in the well-known Python machine learning package sci-kit learn. Literature review also found some more advanced classification models such as Neural Ensemble, FastText, BERT, and C-GRU.

5. Contributions

As a one-person team project, instead of describing the division of work in this section, the author will present the planned steps to meet the requirements of the course. As a beginner in natural language processing, much of the early efforts to complete this project will be spent on understanding the state-of-art, mechanism, and some mainstream methods in hate speech automatic detection. During this stage of learning, we will pay more attention to identifying well-established metrics that have been proven to yield better results than our baseline model. Moreover, as the author is especially interested in avoiding overfitting, we shall also collect some cases in the correction methods. Then the author will choose a labelled English dataset from the website (http://hatespeechdata.com/) and split it into four sets: training, tuning, development and test. We will establish the baseline by applying the simple model and then compare its result to a more complicated model identified during the literature review. Finally, we shall seek ways to visualize the statistical summary. If capacity allows, the author would like to apply the trained model on scrapped data from Facebook to explore the validity of the Twitter model in predicting Facebook data. By comparing the result from both platforms, we can answer some automation transferability question on mainstream platforms.

References

[1] Microsoft Research (2017). *Detection of Aggressive Behaviour on Social Media.* Retrieved from https://www.youtube.com/watch?v=sO5afdXlhPg

[2] Casey Newton (2019 Feburary 25). *The Trauma Floor: Secret Lives of Facebook Moerators in America*. Retrieved from:https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona?fbclid=IwAR2C9cMNeH-L07mY0Yfb7Yb1uEjpHNirPRzq87J1lf1uZq4sbHbGkLu2MhI

[3] Founta,A.; Djouvas,C.; Chatzakou,D.; Leontiadis,I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; Kourtellis,N. (2018) *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*. 11th International Conference on Web and Social Media, ICWSM 2018

[4] MacAvaney S, Yao H-R, Yang. E, Russell. K, Goharian. N, Frieder. O (2019) *Hate Speech Detection: Challenges and Solutions*. PLoSONE 14(8):e0221152.https://doi.org/10.1371/journal.pone.0221152

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. Proceedings of the Eleventh International AAAI Conference on Web and Social Media