CS 589 – Text Mining and Information Retrieval
Final Project Report
HateSet: A Dataset for Hate Speech Detection

Akshay Atam
CWID: 20016304
aatam@stevens.edu

Vikrant Gajare
CWID: 10459368
vgajare1@stevens.edu

Problem Statement

This project is based on generating a dataset based on Hate speech. When it comes to free speech online, it comes down to the user to not speak ill to a person on a certain group. We have been given the right to express our views however, we should not use this as a leverage to spread hate on social media. This project focusses on those texts that points to hate speech in the online world.

To put into more detail, the dataset consists of various different topics relating to hate speech. The different topics of hate speech are taken from the Facebook Community guidelines which is also the foundation of our dataset. The guidelines cover various topics of hate speech, making the dataset diverse. Furthermore, the data collection process also takes various datasets as well as GPT-3 to generate text relating to hate speech.

While there are a lot of datasets, there is a question of ambiguity to state one dataset as the benchmark for all hate speech detectors. Also, since there are a lot of models that predict hate speech, there is not a metric that tells us which model is best suited for some other hate speech dataset.

The entire report is as follows: we introduce the description of hate speech, what it is and the related background knowledge. Next, we showcase what research questions were answered. Moving on, we showcase the approach and topics that were covered on hate speech generation followed by some examples that are a part of our dataset. Finally, the report ends with a conclusion and future work.

Introduction

All that follows on this task's background we gained from Stevens Institute of Technology's CS 589 Text Mining and Information Retrieval guided by Professor Xueqing Liu. This project falls under the Data engineering part of Information Retrieval. This domain of Machine Learning deals with gathering information and converting them from unstructured form to structured form. Getting hate speech examples from various datasets would require compiling everything in one with different topics of hate speech to cover.

Hate speech is an act of communication that is intended to harm or intimidate a person or group based on their membership in a certain social category. It is often characterized by a message of intolerance or bigotry. Hate speech can take many forms, including but not limited to, name-calling, slurs, offensive jokes, and hate symbols. Social media has enabled users to express their views freely, at times, with anonymity. While the ability of free speech is a human right that should be cherished, spreading hate and violence towards a certain group is an abuse to this liberty.

Hate crimes are unluckily not anything new in society. However, social media and different means of on-line verbal exchange have all started playing a bigger role in hate crimes. For instance, suspects in numerous current hate-related terror assaults had an intensive social media history of hate related posts, suggesting that social media contributes to their radicalization [1].

Related Work

There are a number of datasets available for training and evaluating hate speech detection models. These datasets often consist of annotated text samples, where each sample has been labeled as containing hate speech or not. The annotations may also include information about the target of the hate speech (e.g., a specific racial or ethnic group), the type of hate speech (e.g., derogatory language, threats of violence), and the context in which the hate speech occurred (e.g., social media, news articles).

Using these datasets, researchers and developers have trained machine learning models to automatically identify and classify hate speech in text. Early state-of-the-art models used classifiers such as decision trees, random forests, SVMs, etc. However, these classifiers were outperformed by the use of deep neural networks due to the lack of word embeddings [4] in early machine learning classifiers. Word embeddings such as Word2Vec [5], GloVe (GV) [6], fastText (FT) [7], etc. are used by deep networks for hate speech detection.

Current deep learning architectures such as long short-term memory networks (LSTMs) [8], gated recurrent units (GRUs) [9], and bidirectional long-short-term-memory networks (BiLSTMs) [10] to name a few. These models can then be used to help identify and remove hateful content from online platforms, or to alert authorities to potential threats of violence. However, it is important to note that hate speech detection is a complex task, and there are a number of challenges and ethical considerations to be taken into account when building and deploying such systems.

Collecting and annotating hate speech data for training classifiers is hard and cumbersome. Specifically, identifying and agreeing whether or not a specific text should be labelled as hate speech is tough, as formerly mentioned, there is no generic definition of hate speech. One example of hate speech data set is "Dynahate" [11]. This dataset consists of hate speech text which is generated and labelled by trained annotators over four rounds of dynamic data creation. Another good example is data collected from twitter. This data can be used to detect hate speech in tweets, to check whether a tweet has a racist or sexist sentiment associated with it.

Wantanabe et al. collected hate speech examples using expressions on Twitter [2]. The work is focussed on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used as features to train a machine learning algorithm which classifies that data into three classes: Clean, Offensive or Hateful.

The authors of [3] use of Natural Language Processing to detect hate speech. Using Natural Language Processing the author focuses on certain key features to detect hate speech which are Surface-level features such as bag of words; Word Generalization; Sentiment analysis: usually negative sentiment pertains to a hate speech message; Lexical Resources as hate speech contains the negative words which can be found here; Linguistic Features; Knowledge-Based Features; Meta-Information and Multimodal Information as non-textual content is also regularly commented on and is therefore a part of the discourse of a hate speech utterance.

Research Question

To understand the research question, we have to understand the relationship between a model and data. With the advancement of technology, we have tremendous amount of data. Machine Learning algorithms use this data to predict future events and classify information accordingly. There is a healthy balance between the size of and performance of model. Although data available is plentiful, they vary in the attributes that they use to infer the same information. Moreover, the quality of data also matters when it comes to model's performance. The real question lies by combining the former with the latter. Below table briefly summarizes the questions that were answered:

Hate speech dataset	Model used	Evaluation Metric
Ethos Binary	BiLSTM + Static BERT embeddings	Accuracy: 80.15%
HateXplain	BERT-HateXplain[Attn]	Accuracy: 69.8%
Ethos Multilabel	MLARAM	Hamming Loss: 0.2948
AbusEval	HateBERT	Macro F1: 0.742
ToLD-Br	Multilingual BERT	F1-score: 0.75

Table 1: *Relationship between hate speech dataset and their associated model. It is hard to find a single dataset or a single model due to the balance of a good dataset and a good model. For instance, ToLD-Br is a hate speech dataset that contains English and Portuguese examples of hate speech. Using, say, HateBERT might give acceptable accuracy for hate speech containing English lexicon but would fail for Portuguese hate speech.*

A model might perform exceptionally well on one data, but will perform poorly on another data. This could be because instances used to train one model might vary with the other. In our case of Hate Speech detection, a model trained to predict hate speech based on derogatory statements only will perform poorly on a dataset that contains examples of threats of violence.

Another instance that needs to be answered revolves under the subjectivity of hate speech. Examples of hate speech will differ according to the society being asked. Although some examples of hate speech will be universal, it is important to note that those examples will not fully encapsulate the domain of hate speech. So, our questions are:

- How can we generate a dataset that encapsulates all/most of the domains for hate speech which a machine learning model, if trained, will give high accuracy and robust to other hate speech datasets as well? And,
- If we are to create a custom hate speech dataset, what policies must it fall under to ensure we are dealing with most of the topics?

Dataset

Our work proposes a formulation of a novel dataset, outsourced from existing datasets and leveraging the power of GPT-3 to construct the test cases. However, one cannot just generate test cases based on what he/she thinks is described as hate speech. We will be using the Facebook Community Guidelines* as a base to construct hate speech data. There are three tiers in the Facebook community standards, each describing in detail of what constitutes as hate speech.

The primary focus of Tier 1 is content that targets individuals or groups of people based on characteristics or immigration status, using violent or dehumanizing language. This can include comparing them to insects, animals, filth, bacteria, or virus, using terms that imply sub-humanity, referring to them as criminals, denying their existence, or discussing harmful stereotypes linked to intimidation, exclusion, or violence on the basis of a protected characteristic.

Tier 2 content targets individuals or groups based on characteristics that are perceived as inferior, such as physical deficiencies (e.g. hygiene or appearance), mental deficiencies (e.g. intelligence or education), or moral deficiencies (e.g. negative character traits or derogatory terms related to sexual activity). This type of content may also include expressions of inadequacy, comparisons to other protected characteristics, contempt for a protected characteristic, hate or dismissal, disgust, or profanity or sexual slurs. It may also include self-admission of intolerance towards a protected characteristic or statements that the characteristic should not exist.

Tier 3 content targets individuals or groups based on characteristics and includes calls for action, statements of intent, or statements advocating or supporting segregation or exclusion. This may include statements that are aspirational or conditional, and may involve various forms of exclusion, such as explicit exclusion, political exclusion, economic exclusion, or social exclusion.

There are a lot of datasets available for hate speech detection, but we cannot find any single dataset that encapsulates a majority of what is considered to be hate speech. It is a very subjective topic. The use of Facebook community guidelines has been chosen because it encapsulates a lot of topics that relate to hate speech.

The dataset collected by us was divided into two and each work was done individually:

Part 1: Dataset generated by Akshay

Tier 1
- Dehumanizing speech: Disease
- Dehumanizing speech: Feces

Tier 2
- Content attacking a person or a group based on their hygiene
- Content attacking a person or a group based on derogatory terms related to sexual activity
- Profane terms or phrases with the intent to insult

Tier 3
- Call for action of social exclusion

Part 2: Dataset generated by Vikrant

Tier 1
- Dehumanizing speech: Sub humanity
- Harmful stereotypes historically linked to intimidation, exclusion, or violence on the basis of a protected characteristic

Tier 2
- Content attacking a person or a group based on their intellectual capacity
- Expressions about deviating from the norm
- Curse that referring to the target as genitalia

Tier 3
- Call for action of exclusion, e.g., explicit exclusion, which means things like expelling certain groups or saying they are not allowed.

Output

Dynahate was used as the main dataset for collecting hate speech. Some of the examples pertaining to above mentioned topics are:

1. Attack based on hygiene
   - women are filth? Absolutely yes!
   - if you r black u r a dirty dog

2. Attack based on derogatory terms (sexual activity)
   - wh*res are morons
   - why is she so sl*tty what the f*ck

3. Explicit exclusion

*https://transparency.fb.com/policies/community-standards/hate-speech/

- All Muslims should be banned from entering this country. Enough is enough!
- Black judges should be banned

4. Dehumanizing speech: sub humanity
   - All women are a monsters and should be killed
   - Primitive towel heads, they should be called, tend to be the worst workers around

Conclusion

With the knowledge acquired from CS 589 – Text Mining and Information Retrieval, we were able to construct a custom dataset that is used for the purpose of Hate Speech detection. The dataset acquired pertains to the Facebook Community guidelines, which covers various topics that consist of hate speech. Due to the vast areas covered by the guidelines, any model that is trained will have an advantage of being robust that solves the discrepancy of having a large dataset and an accurate hate speech detector.

Opportunities for Future Work

After completion of the dataset, the next step would be to see how the model performs using existing machine learning models. Since the data is diverse, the model trained will consist of various domains it could identify a given text as a form of hate speech. Using language models such as BERT could be used to identify patterns in the data which would work exceptionally well given the fact that many state-of-the-art models use BERT.

Another area that could be worked upon is to combine hate speech datasets in other language and train a unified deep neural network that can identify hate speech across all languages. This work might require collection of data from different languages however, we can use GPT-3 to convert existing hate speech into different languages or come up with new hate speech text that can be used.

The avenues that this work possess is only limited to the possibilities of our own intellect. We have the required data; all we need to come up with is how we can use the data we have into action that is used for the betterment of our society.

References

[1] MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, et al. (2019) Hate speech detection: Challenges and solutions. PLOS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152

[2] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394

[3] Anna Schmidt and Michael 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, Valencia, Spain. Association for Computational Linguistics.

[4] https://en.wikipedia.org/wiki/Word_embedding

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546 (2013)

[6] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532{1543 (2014)

[7] Joulin, A., Grave, E., Bojanowski, P., Douze, M., J egou, H., Mikolov, T.: Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)

[8] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735{1780 (1997)

[9] Cho, K., Van Merri• enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

[10] Schuster, M., Paliwal, K.: Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on 45, 2673 { 2681 (12 1997). https://doi.org/10.1109/78.650093

[11] Singh, A. (n.d.). APS/dynahate · datasets at hugging face. aps/dynahate · Datasets at Hugging Face. Retrieved November 25, 2022