Proposal: To Create a Chatbot for Detecting Hate Speech

Wesley Aldridge
waldridge@knights.ucf.edu

Miles C. Crowe
miles.crowe@knights.ucf.edu

Mauricio De Abreu
mabreu@knights.ucf.edu

1. MOTIVATION & PROBLEM STATEMENT

Hate speech is defined broadly as aggressive language targeted at a person for attributes typically beyond their control. Traits such as nationality, gender, race, sexual preference or disability are popular targets of hate speech[*]. Aggression towards the victim may manifest as insults, personal attacks or even threats. As online communication between people is becoming more commonplace, exposure to hate speech is almost certain to reach more people. Unfortunate as it is, this allows the spread of underlying beliefs that contribute to such behavior.

This project aims to curtail the spread of such speech by detecting hateful dialog and providing an alert mechanism to raise events for handling the occurrence automatically as it happens. Clearly stated, the chat bot would be a silent participant in a chat, regardless of the number of participants, which would monitor the text exchanges. When hate speech is detected, the event would be passed to a host framework or external agent to handle the event.

https://www.dictionary.com/browse/hate-speech

Communication channels such as social media and communication software channels such as discord, whatsapp, slack, etc should be moderated in order to protect its users against harassment and also to prevent hate speech to disseminate. Human moderators are far from a ideal solution from the cost perspective. The current advances on NLP however allow building automated online hate speech detection. This is an introduction part, in which you should include a clear motivation on the problem being addressed in this project (why should I care?). You will need to also define the problem in broad terms so that you can outline the motivation.

As the motivation is made clear, you want to also state the problem more formally; in terms that an NLP expert will understand.

2. RELATED WORK

At this stage, you don't want to provide a comprehensive list of related work, but rather you want to consider this as a part in which you will provide a list of the resources that will be used to assist you conducting the project. Find out some of the related work that would be relevant to this project, and summarize how similar or different your work to them is. Even better, highlight in broad terms what would the δ you think you will achieve by this magnificent work be.

Challenge 1

https://homes.cs.washington.edu/ msap/pdfs/sap2019risk.pdf

3. HOW TO TRAIN YOUR DRAGON

We want to use the best/most robust Twitter hate speech dataset from these sources [1][2] to train a neural network to detect hate speech. We then want to use this neural network to create a server bot for Discord that reads user messages and detects hate speech in real time and alerts the mods of the Discord server. Here should be the actual proposal. What methods are you proposing? Describe the method you will use in designing your training method for your dragons. Relate to the problem statement. The description should be specific so that reproduction of the training method you used for your dragon are reproducible. Avoid copying others' work and others' style, and be creative.

4. EVALUATION

A subset of the data (which is over 43,000 data points) will be set aside to be used as the test dataset. The test dataset will not be used for training and will only be used for testing, to avoid overfitting to the test data. In this section, you want to describe two things: the data that you will use for evaluating the method against the problem stated above, and the results. As for data, describe your source of dragons, in as much details as needed, but not too much that I won't have the time to read. Be realistic.

Here, I know that you won't have results, so don't worry. Describe to me the evaluation metrics that you will use for evaluating the approach on the dataset above. Describe concisely the steps you propose to use for evaluation against those methods/metrics. The description should be intended for the non-expert so that she is able to reproduce the results of your work.

A bonus would be if you could propose to compare your work against a baseline.

5. EXPECTED OUTCOMES AND RISK MANAGEMENT

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Here is your chance to tell me what you expect of outcomes.

Also you want to tell me what are the risks associated with the project, and how you plan to deal with them.

6. PLAN AND ROLES OF COLLABORATORS

Wesley will work on creating the neural network and training it with the hate speech data. {Someone} will work on using the neural network to create the Discord bot. Divide your project into components, and tell me who is going to work on what, how much time each will work on each item. Have a timeline for the project. Tasks may include coding, testing, evaluation and analysis, write-up, presentation, etc.

7. REFERENCES

[1] "Hate speech." https://www.kaggle.com/mohit28rawat/hate-speech#labels.csv. Accessed: 2020-01-22.

[2] "Hate speech." http://hatespeechdata.com/. Accessed: 2020-01-22.

This space would be your chance to list the resources you used for the proposal.