Reddit Auto-Moderation by Evaluating Community Opinion

Ayush Baid, Ankur Bhardwaj, Tarun Pasumarthi

1 INTRODUCTION

People on social media are more divided along political beliefs and ideologies than ever before. Platforms like Reddit, Facebook and Twitter have experienced of influx of radical and extreme views, echo-chambers and negative commentary about the people who do not agree with their views. Moreover, these platforms are becoming sites of contentious political arguments and sometimes go extreme with death threats and radical opinions. Amidst vast growth of social media communities, it is becoming more and more challenging for companies to moderate discussions and keep forums a place for open and honest conversations.

Reddit is organised into more than a million communities called subreddits, where subscriber count reach as high as twenty million. Each subreddit is devoted to a different topic - for example: r/Soccer, r/DonaldTrump, r/News and many more. Some of these subreddits are severely political, polarized and become echo chambers of radical opinions: r/TheDonald, r/OurPresident, r/YangForPresidentHQ.

Our objective is to create a scalable system to generate concise report for human moderators and enhance deletion/flagging of toxic subreddits. Till recent times, fully automated AI based systems were not competent enough to identify misinformation and hate speech across millions of subreddits/communities due to high level syntactic variations and semantic concepts of a language. We hope that our project bridges this gap and provide a small amount of data to review instead of millions of comments per day.

Not only would this be beneficial to society, as it prevents the proliferation of radicalized factions, but it will also help social networks keep their sites advertiser friendly. Furthermore, by detecting hate speech and toxic content, that often leads to the mass exodus of users and deteriorates public image, this tool should help foster positive and constructive dialogues.

2 RELATED WORK

2.1 Text Generation

Since 2013, word embeddings pre-trained on algorithms like word2vec [4] and Glove [5] have been used for all NLP related tasks to initialize the first layer of a neural network. Though these pre-trained word embeddings have been immensely influential, they have a major limitation: they only incorporate previous knowledge in the first layer of the model—the rest of the network still needs to be trained from scratch which is difficult in sparsity of labeled data. These embeddings also struggle against the large variations of the language as a same view can be expressed in numerous different styles.

Since the last few years, there has been a paradigm shift in the NLP community. Similar to advances in transfer learning in image recognition, people have moved from pre-trained shallow embeddings to fully pre-trained hierarchical deep networks for NLP. Networks like GPT-2 [7], ULMFit [1], ELMO [6] demonstrate successful learning of sentence semantics and syntactic concepts to generate text which matches human performance in some cases.

2.2 Auto-Moderation

Currently, content regulation on Reddit is a socially distributed endeavor in which individual moderators coordinate with one another as well as with automated systems [2]. These automated systems, or automods, filter out individual comments by looking for user generated regex patterns. Not only is this approach reliant on human moderators to come up with these expressions, but it provides a very limited rule based approach that is expensive and not feasible at a massive scale.

Machine Learning based automoderator tools, like Washington Post's ModBot [3], bridge the scalability issue, but they are susceptible to false positives as they are not good at understanding the context of a given post.

Furthermore, There is no moderation tool which works on a community as a whole and can condense petabytes of user's posts. During recent years, Reddit has been slow to shut down subreddits even after numerous user complaints.

3 PLAN OF ACTION

3.1 Proposed Approach

We plan to create a deep learning based model which can condense the belief and opinion of different subreddits. Instead of using the existing approach of performing sentiment analysis/moderation on individual comments and aggregation, we think that condensation based approach will give us more insights into how the community thinks of different topics.

We will learn a model for each subreddit, using the user comments. After the training phase, we will use the model to complete our manually compiled corpus of phrases. These phrases will be charged and based on topics on which people have divided and strong opinions. Some examples are:

- Democrats are ...
- Donald Trump is ...
- Immigrants are ...
- <x race> people will ...

Responses from each subreddit's language model will serve as an indicator of the views of their people. This data can be very useful for human moderators, as they can easily parse around 100 responses per subreddit instead of going through millions of comments, and decide to shut down communities which break the rules and engage in toxic conversation. A flowchart with examples is presented in figure 1.

As this approach to condense information is novel, we will use different language models for sentence prediction. We will start with n-gram and word2vec embedding as base-line to complete the phrases. These methods try to predict the next words by using the probability distribution of the training data. These models employ just a few layers and can be trained from scratch.

The baseline models do not have a solid understanding of the language, and hence are not up to the task of condensing the main ideas of the sentences. We will then move to use state-of-the-art

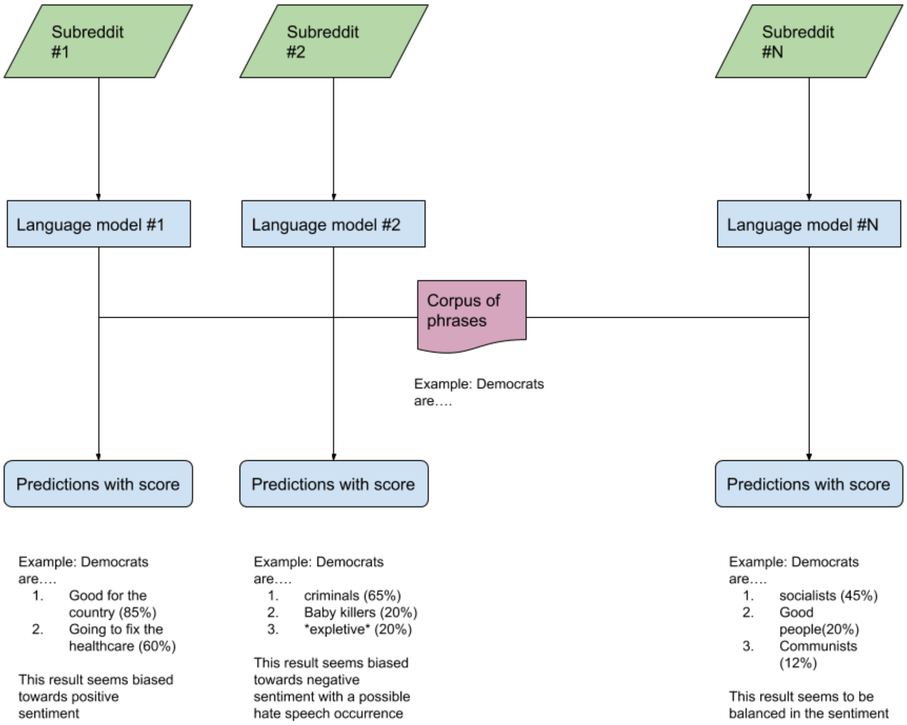

Figure 1: Flowchart for the proposed method

language models like GPT-2 [7], ULMFit [1], ELMO [6] which have proven to be better at understanding the language semantics and are good at text generation. As these methods have gigantic number of parameters. Hence, we will use transfer learning and use pre-trained models on large datasets and fine-tune the last couple of layers on comments from each subreddit.

We will then use the different language models to complete the corpus of around 100 phrases. The complete phrases can then be collated and compared between different subreddits. We will manually annotate the results for bias and potential toxicity.

3.2 Improvements over existing approach

Right now, there is no automated moderation tool which can perform comparable to humans. All the big social media platforms employ humans to moderate content. Most AI based tools analyse each individual post and classify it as problematic or otherwise.

Our approach will understand the beliefs of a large group of people, and how it changes over time. Moderation decisions like quarantining/banning is heavily contested, and we our language models will be able to provide granular reasoning for the decisions (homophobia/inciting violence etc.) The distilled opinions are few in number and are quick to analyse by a human supervisor.

3.3 Risk and Anticipated Challenges

Language models are notoriously difficult to train. It might require large amounts of data and training time. We plan to use transfer learning to lower the requirements, but it still remains to be seen if we have enough resources. Using sentence completion as a proxy for the beliefs of a community and enforce automoderation/manual-moderation on it is a new avenue we are exploring, and it may happen that the results are not as we expect.

3.4 Dataset and Existing Code

We will use the dataset of comments on Reddit provided by pushfit. Word embedding models like n-grams and word2vec are available from scikit-learn and ntlk, and we plan to use them directly for the baseline version.

For the advanced language models, we will use pretrained models from Open-AI and Huggingface. We will fine-tune these models on the comments obtained from each subreddit. We will also explore fast.ai's pipelines for handling large amount of text data.

REFERENCES

[1] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 328–339.

[2] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[3] Ling Jiang and Eui-Hong Han. 2019. ModBot: Automatic comments moderation. In *Proceedings of the Computation+ Journalism Symposium*.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[6] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.