Detecting threads on Reddit needing moderator attention using NLP

Kailhan Hokstam i6154002 and Ruben Bijl i6169804

May 27th 2019

1 Introduction

Reddit is one of the largest communities on the internet (Singer, Flöck, Meinhart, Zeitfogel Strohmaier, 2014). Users of Reddit, also called "Redditors" can submit, view and discuss information on subcommunities called "Subreddits" (Duggan Smith, 2013). With more and more people using the internet and online discourse hardening, it seems that phenomena like cyberbullying, hate speech, and online abuse, colloquially named "toxicity" are becoming more commonplace (Mohan et al., 2017). Because moderation can cost a lot of time, and specifically on Reddit moderators are not paid, and with the rise of machine learning related methods like deep learning, trying to apply machine learning to moderation is of interest (Reynolds, Kontostathis Edwards, 2011).

Therefore, the aim of this research project is to build a classifier that can detect threads that need moderator attention. More specifically, we want to look at a subreddit called "/r/BlackPeopleTwitter", and look how accurately we can predict if a thread is locked or not based on the comments it contains. The idea being that afterwards a bot could periodically assess if newly made threads are likely to be locked in the future and could then be lead in the right direction by extra moderation to nip the problem in the bud, while in the normal scenario moderators cannot check every thread and will lock the thread after the thread got out of hand.

This subreddit, which describes itself as: "Screenshots of Black people being hilarious or insightful on social media, it doesn't need to just be twitter but obviously that is best." is 59th largest subreddit and active according to redditlist.com as of 27th of May, 2019 and has a high ratio of locked threads, often because of racist, homophobic and transphobic, thus "toxic", comments according to the moderators. Because of the size of this subreddit and the apparent need for frequent moderator action, and because of the practicalities of getting our dataset only this subreddit has been analysed.

Table 1: Amount of threads locked from the top 1000 most upvoted posts all time

Subreddit	Locked	Not Locked
Blackpeopletwitter:	115	880
Politics	1	998
Cringepics	12	976
Cringe	10	976
News	29	967
IAmA	15	984
Gaming	13	970
Circlejerk	2	990
Askscience	36	964
Changemyview	2	977

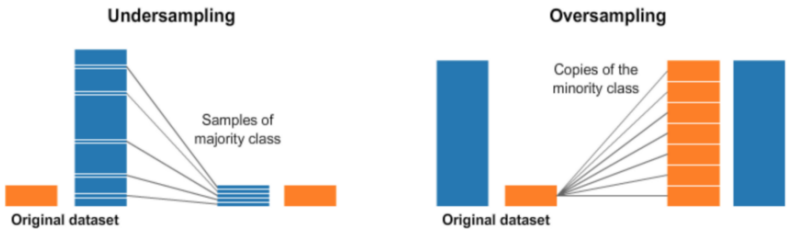
The data used to analyze '/r/BlackPeopleTwitter' has been obtained using Python and the Reddit API, but because of recently imposed limitations on this API only a 1000 posts can be retrieved from Reddit itself. For every thread the comments were flattened using a breadth-first traversal using a queue, the resulting representation of comments, the thread ID, if the thread is locked was then stored in a database. This traversal does mean that a significant amount of information with regards to the hierarchical structure of comments is lost.

For training the model with the data of the subreddit '/r/BlackPeopleTwitter', the data first has to balanced. To balance the data there were two options: one is undersampling the data and the other is oversampling. Since a dataset is used that isn't large at all, oversampling is chosen. Undersampling is a approach that is seen as way more effective. Since the dataset obtainable from Reddit is only as large as a 1000 threads, it would make the training data too small to create a working model.

For oversampling the data a method is made to divide the data equally. This is a important process for the Reddit threads, since the data consists of almost 1000 threads, all different amount of comments. This means that a few threads with the most amount of comments will be as large a dataset as a big set of threads with a small amount of comments. "Oversampling and averaging can be used to increase measurement resolution" (Silicon Laboratories, 2013). For oversampling the data of the locked threads, the data is divided in locked threads and non-locked threads, sorted by the amount of comments in those threads. The total amount of either locked or not locked comments is divided by 10. After this is done, 10 different arrays are created. While going through the sorted list: ID's of threads are being assigned to the 10 arrays, making sure that when you count the total comments of the threads with their given ID, all the 10 arrays will approximately have the same amount of comments. This is done both for Locked and not Locked threads. This way the first ID-arrays will have the highest amount of comments and the least amount of ID's (threads), and vica versa. This way the ratios for oversampling the Locked threads data can be calculated: the amount of ID's in the first array of the non-

locked threads, divided by the amount of ID's in the first array of the locked threads. And so forwards. Now for every locked ID-array there is a duplicate ratio. For completing the oversampling: every locked thread with a certain ID is being duplicated x times and added to the original data. Where x = The ratio (determined by the ID-array where its unique ID is in) added with a randomly generated number between 0.5 and -0.5 and and rounding this.

Figure 1: oversamplin



https://www.kdnuggets.com/2019/05/fix-unbalanced-dataset.html

2 Our model

Our model is inspired by Tang, Qin Lio (2015) but as later explained in the discussion only a more simple model using Keras was implemented. This model consisted of Stanford's GloVe pre trained word embedding layer provided by Stanford GloVe's algorithm, specifically the 25 dimensional Twitter word vectors. This is connected with a LSTM layer of 128 units, which corresponds with the dimensionality of this layers output space. The last layer is a single standard densely-connected NN layer using a sigmoid activation function. The model is compiled with as objective function a standard binary cross-entropy loss, the adam optimizer and as metric accuracy.

3 Results

The models were trained on 80% of the available data and tested on the remaining 20% of data. The SVM also used Keras' StratifiedKFold with k = 5. The neural networks were trained for 20 epochs using a batch size of 64. The metrics were obtained using sklearn.

Table 2: The models were trained on 80 percent of the available data and tested on the remaining 20 percent of data. The SVM also used Keras' StratifiedKFold with k = 5.

Metrics	GloVe Twitter embeddings + LSTM on Reddit Dataset	GloVe Twitter embeddings + LSTM on Pang and Lee's Movie Review Dataset	Unigram + SVM on Reddit Dataset
Accuracy	0.937677	0.52000	0.963172
Precision	0.889447	0.615385	0.931578
Recall	1.000000	0.041026	1.000000
F1 score	0.941489	0.076923	0.964577
Cohen's kappa	0.875309	0.017023	
ROC AUC	0.962555	0.491532	1.0

Table 3: Confusion matrix for GloVe Twitter embeddings + LSTM on Reddit Dataset

N = 354	Predicted: yes	Predicted: no
True: yes	154	22
True: no	0	177

Table 4: Confusion matrix for GloVe Twitter embeddings + LSTM on Pang and Lee's Movie Review Dataset:

N = 400	Predicted: yes	Predicted: no
True: yes	200	5
True: no	187	8

4 Discussion and conclusion

The results indicate that it is possible to detect if a thread can be locked using our model using LSTM and therefore we can say with ¿90% according to our results if a thread is locked or not, but a livestream of threads would be needed to determine how accurately this can be determined when a thread is not actually locked yet.

The results obtained using the custom made model were very good, however these results could not be replicated on Pang and Lee's Movie Review Dataset strongly suggesting that the data scraped contains features that only appear after a thread has been locked for the locked threads or the dataset is in some other way strongly biased, because the model performs very weakly on a benchmark that is known to be valid. The fact that the SVM actually performs on

par and even slightly better than the model using LSTM also supports this fact, because according to the benchmarks in Tang, Qin Lio (2015), a LSTM based model should outperform a SVM on sentiment analysis related tasks, but this is not the case in these experiments.

In further research it would be interesting to not 'flatten' the comments of a thread, so that the structure of the thread could be taken into account, longer comment chains might imply discussions, which can lead to abusive behaviour. Due to limited resources it was not possible to implement a neural network as proposed by Tang, Qin Lio (2015), but such a method that tries to represent the structure of in their case a document with paragraphs and in this paper a thread with comments, seems promising.

For creating models that can recognize all threads on different subreddits, first thing to figure out is the reasons for the threads on differents subreddits that get locked. In the project of 'Hate me, hate me not: Hate speech detection on Facebook' this is mentioned clearly: We first propose a variety of hate categories to distinguish the kind of hate. (Del Vigna, F. Cimino, 2017).

5 Literature

- Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1422-1432).
- Manevitz, L. & Yousef, M (2001). One-Class SVMs for Document Classification (pp. 139-154).
- Del Vigna, F. & Cimino, A. & Dell'Orletta, F. & Petrocchi, M. & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook (pp. 86-95).
- Kotenko, I. & Chechulin, A. & Komashinsky, D. (2015). Evaluation of Text Classification Techniques for Inappropriate Web Content Blocking (pp. 412-417).
- Singer, P.& Flöck, F. & Meinhart, C. Zeitfogel, E. & Strohmaier, M. (2014, April). Evolution of reddit: from the front page of the internet to a self-referential community?. In Proceedings of the 23rd international conference on world wide web (pp. 517-522). ACM.
- Duggan, M. & Smith, A. (2013). 6% of online adults are reddit users. Pew Internet & American Life Project, 3, 1-10.
- Mohan, S. & Guha, A. & Harris, M. & Popowich, F. & Schuster, A. & Priebe, C. (2017, May). The impact of toxic language on the health of reddit communities. In Canadian Conference on Artificial Intelligence (pp. 51-56). Springer, Cham.

- Pavlopoulos, J. & Malakasiotis, P. & Androutsopoulos, I. (2017, September). Deeper attention to abusive user content moderation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1125-1135).
- Reynolds, K. & Kontostathis, A. & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine learning and applications and workshops (Vol. 2, pp. 241-244). IEEE.
- IMPROVING ADC RESOLUTION BY OVERSAMPLING AND AVERAGING, 2013, Silicon labs