Aggression Identification

Abstract

In this project, we describe about aggression identification. Over the past few years, there has been significant growth in social media and user-generated content. Various internet platforms such as blogs, Q&A forums, online forums and more help the user to present their ideas and respond to another user's ideas and comment their opinion. These comments can vary from being pleasant or showing appreciation to aggressive, hate speech, offensive language. Since the number of people along with the amount of interaction/content on the web increases, it is impossible to segregate it and filter out the offensive content manually. Therefore, we implemented several Machine Learning and Deep Learning models to classify the aggression in the social media content into three categories: Overtly Aggressive, Covertly Aggressive and Non-Aggressive.

Introduction

Social media and user-generated content have grown at an exponential rate in recent years. Online platforms such as blogs, Q&A forums, online discussion forums, and so on enable users to publish comments and respond to those of other users. These remarks can take many forms, including adoring, hostile, hate speech, and insulting language. As the number of individuals using the internet has grown, so has the number of occurrences of aggressiveness and related behaviours such as trolling, cyberbullying, flame, and hate speech, among others. As a result, online aggressive behaviour has become a major cause of social friction, with the potential to turn criminal.

One of the most difficult aspects of detecting aggressiveness on social media is distinguishing it from rude or caustic language. Although some of the work has been completed in this area, it remains a hot topic among scholars. We create a technique to distinguish between Overtly Aggressive (OAG), Covertly Aggressive (CAG), and Non-aggressive (NAG) material in texts with this in mind.

We took advantage of the fact that emojis may be used to represent the emotional content of a text. Pre-training on the classification problem of determining which emoji were originally part of a text can thus help with the objective job. We used three different methods: TFIDF (Term Frequency Inverse Document Frequency), Word Embeddings, Word2Vec to train for classification.

Literature Review

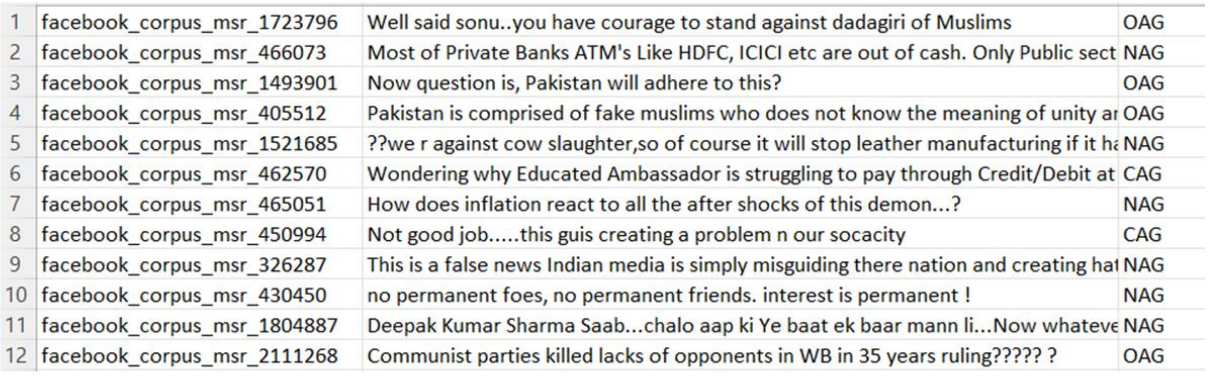
The paper [1] describes the work that their team did at Indian Institute of Technology (ISM) towards TRAC-1 Shared Task on Aggression Identification in social media for COLING 2018. In this paper, they labelled aggression identification into three categories: Overtly Aggressive, Covertly Aggressive and Non-aggressive. Then, they trained a model to differentiate between these categories and analyse the results in order to better understand how it can be distinguished between them. They participated in two different tasks named as English (Facebook) task and English (social media) task. They used LSTM Model with F1-score as their evaluation metric. For English (Facebook) task System 05 was their best run (i.e., 0.3572). For English (social media) task their system 02 got the value (i.e., 0.1960). Overall, their performance is not satisfactory. However, as new entrant to the field, their scores are encouraging enough to work for better results in future.

The paper [2] describes about their participation in the First Shared Task on Aggression Identification. The method proposed relies on machine learning to identify social media texts which contain aggression. The key features employed by their method are information extracted from word embeddings and the output of a sentiment analyser. The official evaluation showed that for texts similar to the ones in the training dataset Random Forests work best, whilst for texts which are different, SVMs are a better choice. The purpose of this shared task was to classify messages from social media into three categories Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-aggressive (NAG). The first one was a dataset which contained text from Facebook and therefore it was similar to the training set. The second dataset consisted of tweets and gave the opportunity to assess the systems on a dataset which was quite different than the training data. The obtained results showed that RF performed well when combined with sentiment analysis than other models.

In the paper [3], Human annotated Twitter data was collected in the immediate aftermath of Rigby's murder in 2013, to train and evaluate a supervised machine learning text classifier that distinguishes between hateful and/or antagonistic responses with a focus on race, ethnicity, or religion, and more general responses. Classification features are extracted from the content of each tweet, including grammatical dependencies between words to recognize phrases, incitement to respond with antagonistic action, and claims of well-founded or justified discrimination against social groups. The results of the classifier were optimal using a combination of probabilistic, rule-based, and spatial-based classifiers with a voted ensemble meta-classifier. They demonstrate how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data.

Dataset Description

The dataset is a collection of posts from Facebook made by people. For training and validation, the Aggression Identification dataset contains 15,000 aggression-annotated Facebook Posts and Comments in English. It contains user id, text data in English and class it belongs to (Overtly Aggressive (OAG), Covertly Aggressive (CAG) or Non-aggressive (NAG)).



1	facebook_corpus_msr_1723796	Well said sonu..you have courage to stand against dadagiri of Muslims	OAG
2	facebook_corpus_msr_466073	Most of Private Banks ATM's Like HDFC, ICICI etc are out of cash. Only Public sect	NAG
3	facebook_corpus_msr_1493901	Now question is, Pakistan will adhere to this?	OAG
4	facebook_corpus_msr_405512	Pakistan is comprised of fake muslims who does not know the meaning of unity ar	OAG
5	facebook_corpus_msr_1521685	??we r against cow slaughter,so of course it will stop leather manufacturing if it ha	NAG
6	facebook_corpus_msr_462570	Wondering why Educated Ambassador is struggling to pay through Credit/Debit at	CAG
7	facebook_corpus_msr_465051	How does inflation react to all the after shocks of this demon...?	NAG
8	facebook_corpus_msr_450994	Not good job.....this guis creating a problem n our socacity	CAG
9	facebook_corpus_msr_326287	This is a false news Indian media is simply misguiding there nation and creating hat	NAG
10	facebook_corpus_msr_430450	no permanent foes, no permanent friends. interest is permanent !	NAG
11	facebook_corpus_msr_1804887	Deepak Kumar Sharma Saab...chalo aap ki Ye baat ek baar mann li...Now whateve	NAG
12	facebook_corpus_msr_2111268	Communist parties killed lacks of opponents in WB in 35 years ruling????? ?	OAG

Methodology

The approach used in this project relies on Machine learning and Deep learning to classify the posts into three respective classes: non-aggressive, covertly aggressive, and overtly aggressive. As we cannot directly use the raw corpus, we will clean and pre-process the text to improve the model's performance using 'ekphrasis' package in python.

The Pre-processing block contains Text Pre-processor to normalize, annotate and fix HTML tokens, Social Tokenizer to covert text into lower case and an Emoticon dictionary to change emoticons into text. This pre-processed text is then passed to Spell Corrector and WordNet Lemmatizer to obtain clean text.

To perform the task, we devised three strategies; Feature extraction using TF-IFD Vectorizer and feature selection using chi-squared method to reduce the extracted features. Several ML and DL models are employed in the pipeline and the model's performance is measured based on the classifications. The second strategy is based on word embeddings using Word2Vec model from Gensim package. The extracted embeddings are given as input to the LSTM models for predictions. The last strategy is to use BERT language model to consider the context of the corpus for better classification.

Results and Discussion

The results from the three strategies discussed earlier are as follow:

1) TF-IDF Features -

Model	Accuracy (in %)
Random Forest	56.3
Multinomial Naïve Bayes	55.4
Logistic Regression	56.8
RNN	51.4
LSTM	51.1
GRU	51.8

2) Word Embeddings + LSTM -

Evaluation Metrics	Value (out of 1)
Accuracy	0.423
Precision	0.426
Recall	0.522
F1 score	0.603

3) BERT Language Model -

Evaluation Metrics	Value (out of 1)
Accuracy	0.547
Precision	0.541
Recall	0.539
F1 score	0.534

Conclusion

We can observe that ML models achieved high accuracy by a close margin when compared to others. There are also fewer misclassifications in the CAG class in DL models, but OAG class misclassifications are more. The BERT language model, while compared to others, achieved a slightly lesser accuracy but the toughest class OAG is classified well. The performance of the best systems in the task shows that aggression identification is a hard problem to solve. Moreover, the performance of the Neural Networks-based systems as well as the other approaches does not differ much. If the proper features selection techniques are employed, then classifiers like SVM and even Random Forest and Logistic Regression performance is same as Deep Neural Networks. On the other hand, we find quite a few neural networks-based systems not performing quite well in the task. Hence, we may point out the apparent “inconsistencies” in the annotation and need to get the annotations validated by multiple human annotators.

References

[1] Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, M. R. Chennuru. TRAC-1 Shared Task on Aggression Identification: IIT(ISM)@COLING'18.

[2] Constantin Orasan. Aggressive language identification using word embeddings and sentiment features

[3] Pete Burnap, Matthew L. Williams. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making.

[4] Sepp Hochreiter and Jurgen¨ Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

[5] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a ConvolutionGRU Based Deep Neural Network.