IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

JIGSAW MULTILINGUAL TOXIC COMMENT CLASSIFICATION

[1]Shanmughapriya M, [2]Gowri R R, [3]Abitha K E

[1]Master of Computer Applications, [2] Master of Computer Applications, [3] Master of Computer Applications
Department of Computer Applications,
Coimbatore Institute of Technology,
Coimbatore,
India

Abstract: We all know that internet is important tool in today's era. It has made our life at ease. It has both pros and cons, where the great growth of the internet has made wide selection of individuals to come back on-line. Every part of people's time, work, and training takes place in the digital world. All these different cultures have a lot of effects different from their aspect with this start to act irrationally and start to fight online. Communication is a transmission medium which leads to connect people around the world. When it comes to interaction with many other people it could lead to non-verbal communication. The new exploration suggested that people transfer behavior they learn in online setting to their day-to-day life. This issue may leads to on-line harassment and personal attacks. With this project, the idea is to help online user to identify the toxic comments and mark them as inappropriate.

Keywords: LSTM, Natural Language Processing, Text mining, Toxic text classification, Word embeddings, Word2vec.

1 INTRODUCTION

All platforms that serve a great deal of individuals can, in one purpose of their existence, have disagreements and harassment from individuals. To counter that flow of non-constructive comments, this competition was created to yield the most effective rule for drooping toxic comments. Many comments are shutdown in the comment section to enable the effectiveness in the online platform struggle. This project was made to focus machine learning models to identify toxicity in online conversations and mark them as rude, disrespectful using Natural Language Processing concepts and techniques. If these comments could be identified that would lead to safe and more security.

1.1 Related work

In 2018, a contest was persevered Kaggle known as "Toxic Comment Classification Challenge". In this competition, competitors were asked to create models not solely to acknowledge toxicity, however to conjointly find few styles of toxicity. The kinds of toxicity that had to be detected are: severe toxicity, obscene, threat, insult and identity hate. The goal was to modify users to select specific styles of toxicity and specialize in them, since some sites can be fine with one kind of toxicity (e.g., severe toxicity) and not others

1.2. Kaggle

Kaggle is one of biggest data science and machine learning communities where users are publishing datasets and kernels for everyone to see. This web-based system allows data scientists and machine learning engineers to enter competitions where they try to solve data science challenges. Jigsaw developed Perspective API, used to determine an impact a comment that affects the conversation with the help of machine learning models. The goal is to see it these results are applicable to toxicity classification. The result lies between 0 and 1.

1.3. Dataset

Provided training data is English-only. Data consists of columns such as id, comment text, toxic and types of toxicity each one in separate column.

Test data is consisted of comments from Wikipedia talk page in several different languages (Spanish, Italian, French etc.,). Test data consists of few columns: id, comment content and language of the content.

Other than that, competitors were provided validation data, which is just in non-English languages as a test data. It consists of columns: id, comment text, language and toxic column.

2 METHODOLOGY

The first step was to inspect the data. We work on training dataset; the number of provided comments exceeded in million. Comparing the toxic and non-toxic comments, we come to know only 6% comments are toxic. To lower the possibility of false negatives after the training, we decided to normalize the dataset. The ratio we opted for is 1:2. So, for each toxic comment there are 2 not-toxic ones.

After that, we did some feature extraction processes. In the end, we used a classifier of choice to train the model and predict the toxicity probabilities of test data.

2.1. Feature *Extraction*

Feature extraction is a process of dimensionality normalizing the set of raw data. Characteristic of datasets differ from one to another. In the situation with datasets provided for this competition, there are no available features which could help determining the toxicity of a comment based on the emotions e.g., sad, happy, angry. Extracted features and their significance will be explained in further sections.

2.2. Comment length

We take the length of each and every comment given by the user. For this reason, we have a tendency to check the length of comment, and located out that there's a distinction between average length of toxic and non-toxic comments. Non.-toxic comment is larger than toxic ones. The correlation could be a negative price as a result of the upper the comment length is, the lesser the toxicity is.

2.3. Count of bad words

Word representation algorithm like Word embeddings or Word2vec is used to have a similar representation of semantic in vector format. In this way it could be said that words such as "Man" and "Woman" is similar such as "King" and "Queen". One of the most famous examples are "Man" + "King" - "Woman" = "Queen" that simplifies understanding of these algorithms for novice developer. With this finding we implemented word2vec into this dataset to obtain the correlation between the toxic columns.

2.4. Sentiment

Sentiment analysis is classification of emotions from text data using text analysis techniques. Sentiment analysis allows data

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

Fig. 1. Word vectorization

scientist to identify user that has positive, negative or neutral thoughts about certain topic. In this case sentiment was used on comments to figure out which comments deviate from neutral in to negative spectrum and they would be marked as negative in nature. It is Python library allows us to perform basic NLP tasks in returns two properties polarity and subjectivity.

2.5. Subjectivity

Subjectivity is other part of sentiment analysis and besides of correlations it gives us information about in which way the sentences were written. For example, personal opinion comes under subjective sentences and objective refers to factual information. The correlation between toxicity and subjectivity of this dataset was found to be -1.5 percent. As the correlation was small we couldn't proceed further.

	id	subjectivity	polarity	toxic	comment_text	lang
0	0	0.403141	0.112854	0	Este usuario ni siquiera llega al rango de ...	es
1	1	0.368821	0.017503	0	Il testo di questa voce pare esser scopiazzato...	it
2	2	0.360789	0.271745	1	Vale. Sólo expongo mi pasado. Todo tiempo pasa...	es
3	3	0.421223	0.141766	0	Bu maddenin alt başlığı olarak uluslararası i...	tr
4	4	0.382030	0.040717	0	Belçika nın şehirlerinin yanında ilçe ve belde...	tr

Fig. 2. Correlations and Subjectivity dataset

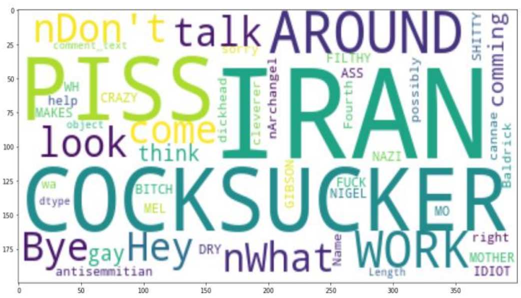

Fig. 3. Word cloud toxic comment

After extracting all the wanted features, our next move was to classify the comments. The classifier we chose for this task is Long Short-Term Memory (LSTM).

First, we dropped the columns which are not important for the prediction itself, which are: id, comment text, toxic and lang. We separated the features (all the calculated features) and labels (toxic), so we have the X and y for the classifier.

Try to fit the model after splitting the test and train data. One of the train datasets we were supposed to use had the probability of toxicity as a value in toxic column. That was giving us an error, since the model can't be trained and tested on labels (classes) which are decimal numbers. Our solution to this problem was to round up/down the toxicity probabilities in the training dataset. All the values which were below 0.5 were rounded down and other were rounded up. With this, changed, dataset we repeated the process and trained the model.

After training the model, we used the predict_proba method, which predicts the probability for each class. The output is an array of arrays, each one of them having two values.

One value is the probability that the comment is toxic, and other one the probability that it is not toxic.

3 RESULTS AND CONCLUSION

3.1 Result

In this work we fit the model using LSTM algorithm and obtain the accuracy score as of 0.9263 which is better score compare to other algorithm like random forest.

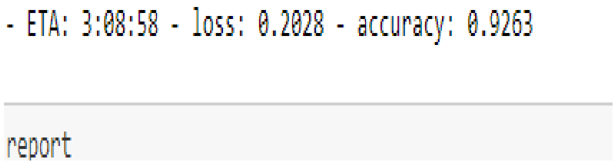

Fig. 4. Accuracy of the algorithm.

3.2 Conclusion

This result sounds better when it is put in the context. In conclusion with this approach, we have found a lot about the dataset with previously described ways and learned a lot about the NLP field. It broad spectrum of different ways of approaching the problem is sometimes overwhelming but really interesting to learn.

4 FUTURE SCOPE

In this work can be improved by the following things.

- The integer and float type is difficult to train in the machine learning model.
- Word embedding in optimization can be improved using genetic algorithm to maximize the semantic correctness

ACKNOWLEDGMENT

We take this opportunity to express our sincere thanks to our institution, Coimbatore Institute of Technology, Coimbatore, and to our faculty members for providing an opportunity to carry out the paper. We are privileged to have worked under the guidance of Dr.J.B.Jona Associate Professor in the Department of Computer Applications, Coimbatore Institute of Technology in Coimbatore for her motivation and encouragement to improve our skills in the successful completion of this paper.

REFERENCES

[1] Analytics Vidhya. 2020. Get Started Kaggle Cometitions. (2020).

[2]. DeepAI. 2020. Feature Extraction. (2020).

[3]. Kaggle. 2018. Toxic Comment Classification Challenge. (2018).

[4]. Kaggle. 2019. Jigsaw Unintended Bias in Toxicity Classification. (2019).

[5]. Kaggle. 2020. Jigsaw Multilingual Toxic Comment Classification. (2020).

[6]. Kawin Ethayarajh. 2019. Word Embedding Analogies: Understanding King - Man + Woman = Queen. (2019).