A
Project on
MALIGNANT COMMENTS CLASSIFICATION

By

Snehal Kumawat

Content

- Introduction
- Literature
- Data preparation & Cleaning
- Methodology
- Conclusion
- Limitations
- Scope for Future Work

Introduction

Business Problem Framing

- The proliferation of social media enables people to express their opinions widely online.
- This has resulted in the emergence of conflict and hate, making online environments uninviting for users.
- Our goal is to build a prototype of online hate and abuse comment classifier.

Conceptual Background of the Domain Problem

- The background for the problem originates from the multitude of online forums, where-in people participate actively and make comments.
- As the comments some times may be abusive, insulting or even hate-based,

Review of Literature

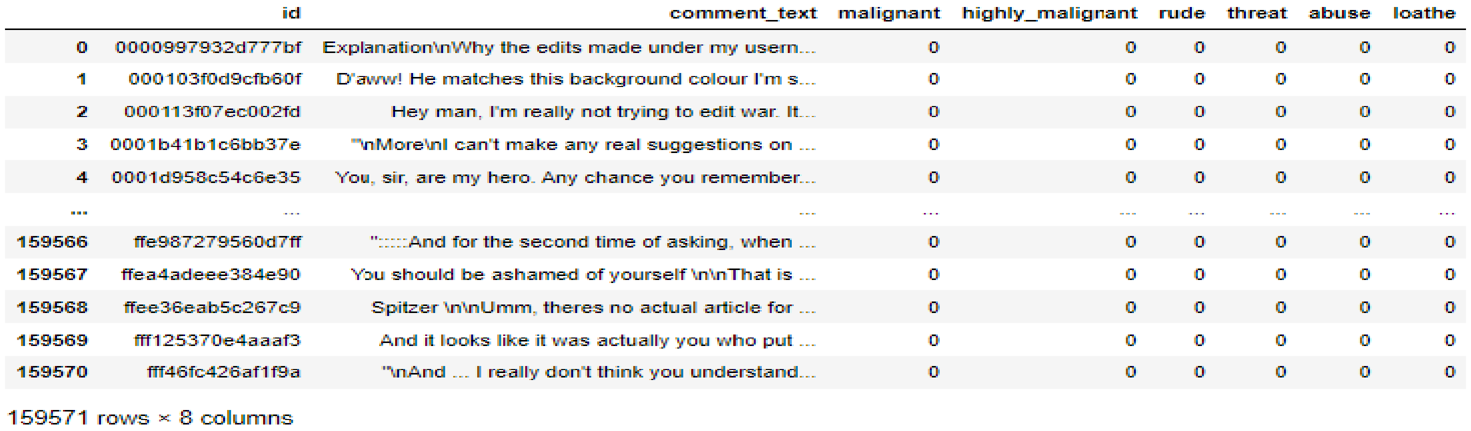
- There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments.
- Different techniques like multiple logistic regression analysis, k-nearest neighbours, Random Forest and decision trees, Gradient Boost, XGBoost classification techniques have been used to make the predictions. .

Motivation for the Problem Undertaken

- The problem we sought to solve was the tagging of internet comments that are aggressive towards other users.
- So based on existing data, the aim is to use machine learning algorithms to develop models for classification of malignant comments.

Data Sources and their formats

- The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples.



	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
159566	ffe987279560d7ff	":::::And for the second time of asking, when ...	0	0	0	0	0	0
159567	ffea4adeee384e90	You should be ashamed of yourself \n\nThat is ...	0	0	0	0	0	0
159568	ffee36eab5c267c9	Spitzer \n\nUmm, theres no actual article for ...	0	0	0	0	0	0
159569	fff125370e4aaaf3	And it looks like it was actually you who put ...	0	0	0	0	0	0
159570	fff46fc426af1f9a	"\nAnd ... I really don't think you understand...	0	0	0	0	0	0

159571 rows × 8 columns

fig 1.Sample data before data pre-processing

Data preparation and cleaning

Fig. 1 captures the research framework for the malignant comment classification problem. It includes five major blocks, namely data collection, data preparation, feature processing, mode training, classification and model evaluation.These blocks of the diagram are explained in detail in the next subsections.

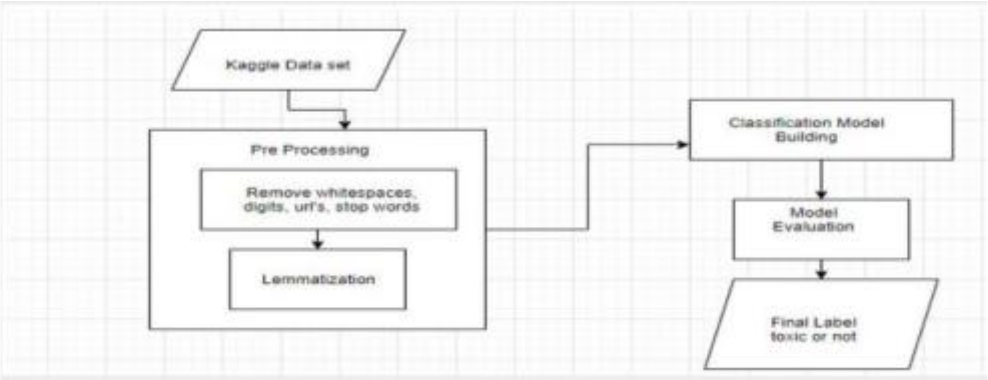


fig 2. Research framework

In this model, we need to feed the available independent variables to the model will classify whether the comments are malignant or not. For designing the model the machine learning method I used is multiple classification model and the tool I used for coding is jupyter notebook.

About Project

All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’. The project is related multilabel classification.
Packages used: Pandas, Numpy, Seaborn, Matplotlib, Scikit and Stats models, NLTK, Lemmeaizer.

- I counted comments belonging to each of the different categories.

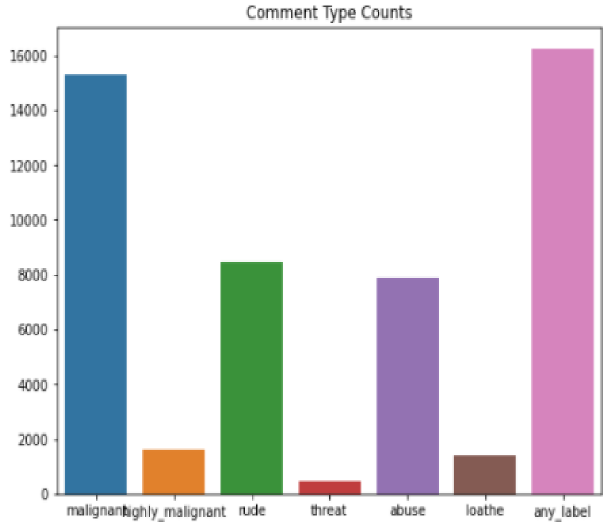
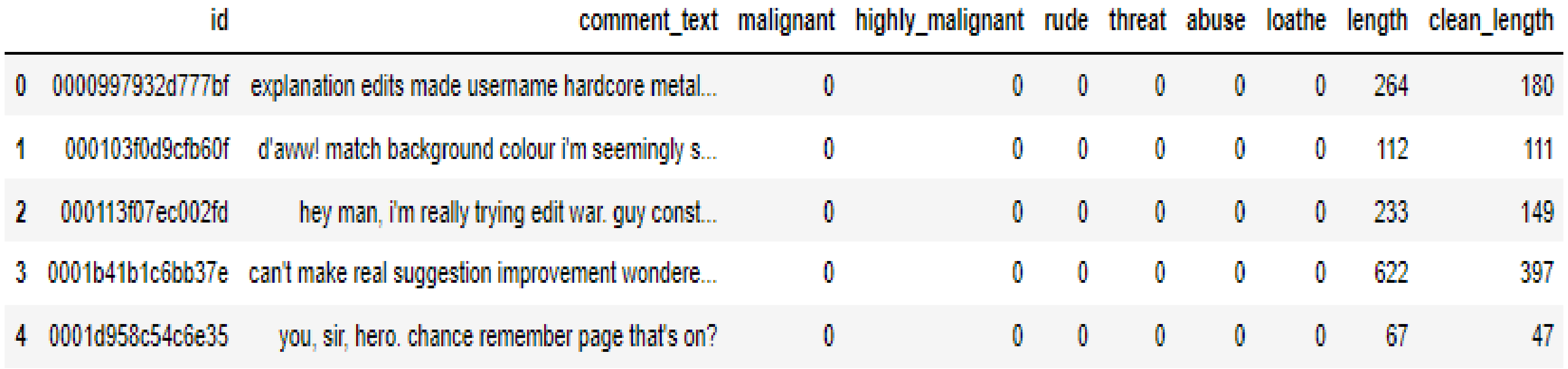


fig2.comments belonging to different categories



	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe	length	clean_length
0	0000997932d777bf	explanation edits made username hardcore metal...	0	0	0	0	0	0	264	180
1	000103f0d9cfb60f	d'aww! match background colour i'm seemingly s...	0	0	0	0	0	0	112	111
2	000113f07ec002fd	hey man, i'm really trying edit war. guy const...	0	0	0	0	0	0	233	149
3	0001b41b1c6bb37e	can't make real suggestion improvement wondere...	0	0	0	0	0	0	622	397
4	0001d958c54c6e35	you, sir, hero. chance remember page that's on?	0	0	0	0	0	0	67	47

fig3. Sample data after data preprocessing

Data Preprocessing

1. Preparation for removal of punctuation marks:
2. Updating the list of stop words
3. Stemming and Lemmatising
4. Applying Count Vectorizer
5. WordCloud

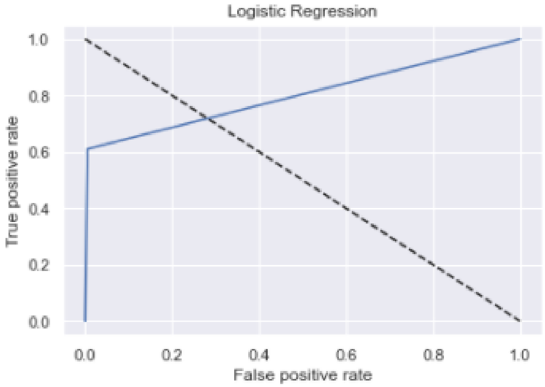
Methodology

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model. Using this we get

Training accuracy is 0.9595520103134316

Test accuracy is 0.9552974598930482

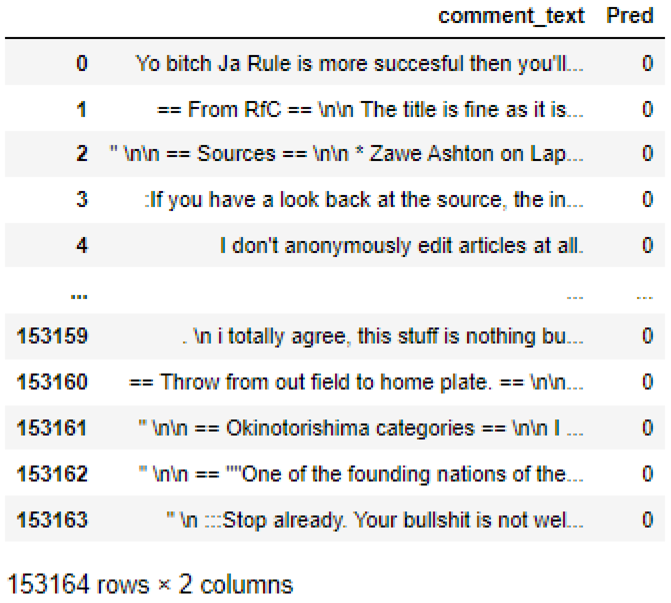


Auc ROC curve

Conclusion

It is seen that the most effective model in predicting malignant comments is Logistic Regression. .

And here is output for test dataset values.



	comment_text	Pred
0	Yo bitch Ja Rule is more succesful then you'll...	0
1	== From RfC == \n\n The title is fine as it is...	0
2	" \n\n == Sources == \n\n * Zawe Ashton on Lap...	0
3	:If you have a look back at the source, the in...	0
4	I don't anonymously edit articles at all.	0
...	...	...
153159	. \n i totally agree, this stuff is nothing bu...	0
153160	== Throw from out field to home plate. == \n\n...	0
153161	" \n\n == Okinotorishima categories == \n\n I ...	0
153162	" \n\n == ""One of the founding nations of the...	0
153163	" \n :::Stop already. Your bullshit is not wel...	0

153164 rows × 2 columns

Limitations of this work and Scope for Future Work

Our research has shown that harmful or toxic comments in the social media space have many negative impacts to society. The ability to readily and accurately identify comments as toxic could provide many benefits while mitigating the harm. Also, our research has shown the capability of readily available algorithms to be employed in such a way to address this challenge.