LETIZIA MOLINARI ID 959194

MASTER'S IN DATA SCIENCE AND ECONOMICS, UNIVERSITY OF MILAN

TEXT MINING AND SENTIMENT ANALYSIS PROJECT

SARCASM DETECTION ON REDDIT

[Final project submitted as exam requirement of the course of Text Mining and Sentiment Analysis]

Course coordinator: Professor Alfio Ferrara

ABSTRACT

Sarcasm in the expression of an implicit information which is usually the opposite of a message content and it is used by people in order to hurt someone emotionally or to criticise something in a humorous way. Its identification in textual data has always been a hard challenge in natural language processing (NLP) and it has recently become an interesting research area due to its importance in improving the sentiment analysis of social media data.

The following project analyses the Reddit dataset, which was generated by scraping comments from the Reddit website containing the *sarcasm* tag. This tag indicates the fact that a certain comment has not meant to be taken seriously and is generally an indicator of the fact that a certain comment is sarcastic. The goal of the project is to predict the probability a certain parent comment is sarcastic or not given only the *parent comment* and the category (*subreddit)* the comment belongs to.

1. RESEARCH QUESTION AND METHODOLOGY

The main goal of the following project is to determine whether or not a parent comment will receive a sarcastic comment, given only the parent comment and the subreddit. This is a challenging and innovative task since most studies in literature only focused on the detection of sarcasm on a corpus by analysing a certain comment and not the parent comment. The analysis is divided in two parts: in the first one, machine learning algorithms such as Logistic Regression, Random Forest Classifier and Support Vector Machine have been implemented only on parent comments, while in the second part the same algorithms with the addition of Multinomial Naïve Bayes classifier have been implemented, in order to perform a model comparison and try to see whether this addition leads to a model improvement. Moreover, TF-IDF (term frequency-inverse document frequency) has been used for extracting features from data.

2. EXPERIMENTAL RESULTS

2.1. Dataset Description

"Sarcasm on Reddit" dataset has been downloaded from Kaggle and it contains 1.3 million sarcastic comments from the Reddit website. The dataset was generated by scraping comments from Reddit website which contained the sarcasm \s tag. This tag is often an indicator of the fact that the comment of a certain redditor has not to be taken seriously and is generally a reliable indicator of the fact that a certain comment is sarcastic. The dataset is very huge: it contains 1010826 rows and about 10 columns, but the analysis only focuses on the three columns defined below:

- Label: a dummy variable which displays whether a certain comment is sarcastic (1) or not (0).
- Subreddit: a categorical variable which indicates the category a certain comment belongs to
- Parent Comment: the original comment posted on the website

First of all, it has been checked whether there were null comments or duplicates. It has been noticed that there were not null comments and that some comments had some duplicates, which have been removed. The final dataset used for the analysis contained 997883 rows and 3 columns. Moreover, all parent comments have been put in lower case.

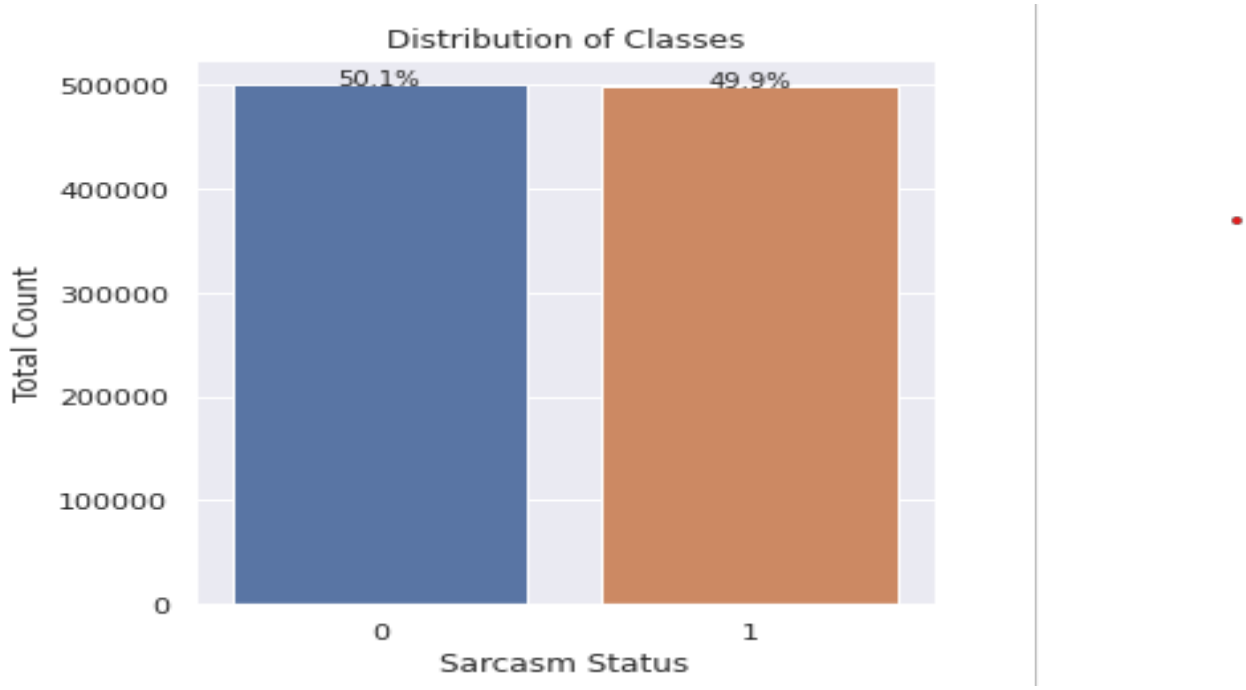

Figure 1, visualization of the balance of the labels

By looking at the bar plot in figure 1 it is possible to notice that the dataset is balanced: 499687 observations are labelled with 0, corresponding to the 50.1% of the total, while 498196 observations are labelled with 1, corresponding to the 49.9% of the total. So, the proportion of sarcastic and not sarcastic labels is the same.

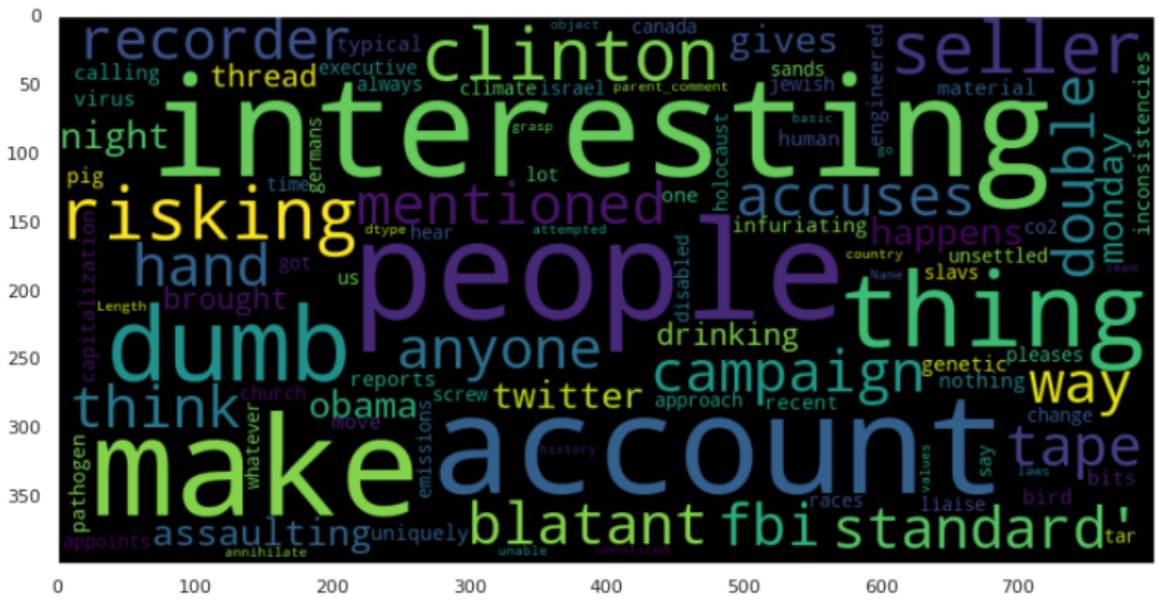

Figure 2, wordcloud of parent sarcastic comments

By looking at the wordcloud in figure 2, it is possible to observe the most popular words contained in a certain parent comment. In particular, it is immediately possible to notice that words ‘interesting’, ‘people’, ‘account’, ‘thing’ and ‘make’, are the most popular ones in parent comments, since they present the highest dimensions.

Then an analysis of Subreddits has been done. First of all, it has been analysed whether there are subreddits more sarcastic than others, and all categories with more than 1000 entries have been taken into consideration.

subreddit	size	mean	sum
creepyPMs	5466	0.784303	4287
MensRights	3356	0.680870	2285
ShitRedditSays	1284	0.661994	850
worldnews	26377	0.642529	16948
Libertarian	2562	0.640125	1640
atheism	7377	0.639555	4718
Conservative	1881	0.639553	1203
TwoXChromosomes	1560	0.632692	987
fatlogic	2356	0.623090	1468
facepalm	1268	0.617508	783

Table 1, size, mean and sum of ten categories having more than 1000 entries.

The table above shows the size, mean and sum of ten categories having more than 1000 entries. Then, it has been decided to extract the top ten categories which are more likely to receive a sarcastic comment.

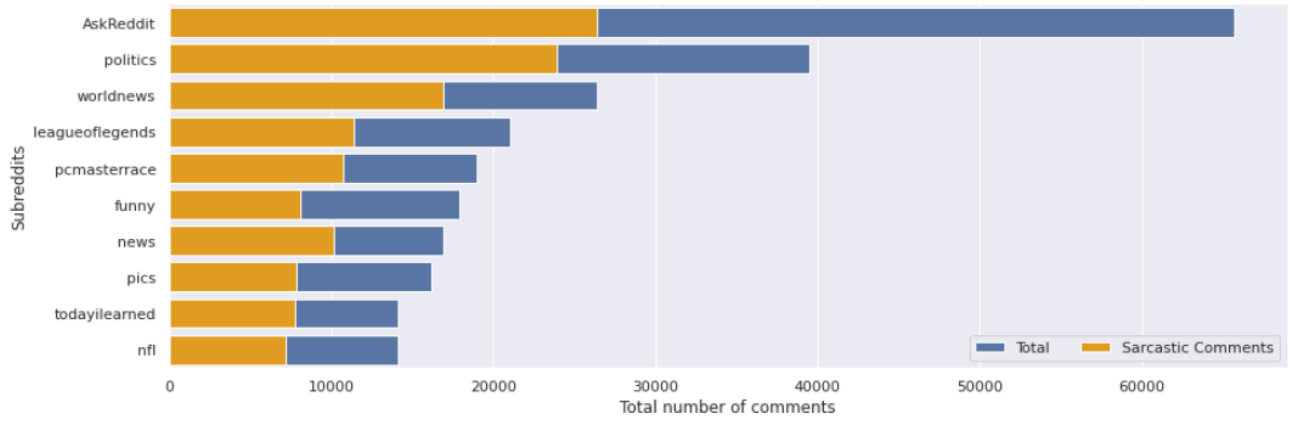

Figure 3, bar plot of the top 10 categories

By looking at the bar plot in figure 3, it is possible to observe the most popular categories and the number of Sarcastic Comments they received with respect to the total. It is possible to notice that AskReddit is the category which received the highest number of sarcastic comments, followed by politics and then worldnews.

	subreddit	sarcastic	natural	total
0	AskReddit	26367.0	39310.0	65677.0
1	politics	23910.0	15586.0	39496.0
2	worldnews	16948.0	9429.0	26377.0
3	leagueoflegends	11409.0	9628.0	21037.0
4	pcmasterrace	10760.0	8228.0	18988.0
5	funny	8099.0	9840.0	17939.0
6	news	10193.0	6698.0	16891.0
7	pics	7825.0	8329.0	16154.0
8	todayilearned	7754.0	6407.0	14161.0
9	nfl	7215.0	6935.0	14150.0

Table 2, quantity of sarcastic and natural comments of the top 10 popular subreddits with respect to the total

By looking at table 2, it is possible to observe the quantity of sarcastic and natural comments with respect to the total for the top 10 categories. For instance, the category AskReddit presents 26367 sarcastic comments on a total of 65777, so 40% of the comments belonging to this category are sarcastic.

2.2. EVALUATION STRATEGY

In order to evaluate the performances of the models implemented, Cross Validation has been used. In particular, five-folds cross validation has been implemented, where four folds are used for training the model and the fifth one is used for testing.
In order to perform the analysis, 60% of the dataset has been used for training and 40% of it has been used for testing.

2.2.1. PART 1: USAGE OF ONLY PARENT COMMENTS AS INPUT

In the first part of the analysis only parent comments have been used as inputs. Parent comments have been vectorized by implementing TF-IDF with tri-grams and putting a limit of 1000 to the features to extract. The choice of putting a limit of 1000 on the maximum features to extract is driven by the fact that the dataset is very huge and is necessary to reduce the features in order not to have problems in running the algorithms. In order to perform this task, Logistic Regression, Random Forests and Support Vector Machine algorithms have been implemented.

2.2.2. PART 2: MODEL IMPROVEMENT ATTEMPT

In order to try to improve the models implemented, subreddits have been added in the analysis as inputs. In particular, they have been vectorized by implementing TF-IDF with bi-grams and putting again a limit of 1000 on the maximum features to extract. In order to concatenate the two inputs, a scikit learn pipeline has been implemented. This pipeline, called merged_features, has been applied to the X_train set of the models which have been implemented. Logistic Regression, Random Forests, Support Vector Machines and Multinomial Naïve Bayes Classifier have been implemented.

3. RESULTS

MODEL	CV-ACCURACY SCORE PART 1	CV-ACCURACY SCORE PART 2
Logistic Regression	0.56	0.58
Random Forest	0.56	0.57
Support Vector Machine	0.56	0.58
Multinomial Naïve Bayes	/	0.58

Table 3, cross-validation accuracy scores

By looking at table 3, it is possible to observe the cross-validation accuracy scores of the models implemented. In particular, it is possible to observe that in the first case, where only parent comments have been used as inputs, all the three models a score of 0.56. In the second case, where also subreddits were included as inputs, the accuracy score slightly increases, from 0.56 to 0.58 for Logistic Regression and Support Vector Machine, and from 0.56 to 0.57 for Random Forest. In the second part, where data present a dense distribution, also a multinomial Naïve Bayes Classifier has been implemented, with an accuracy score of 0.58. It is possible to say that the inclusion of subreddits has lead to a slight model improvement.

3.1. MODEL TESTING AND PREDICTION PROBABILITIES

Since the models used in the second phase of the analysis reported a slightly higher accuracy score, Logistic Regression, Random Forest and Multinomial Naïve Bayes implemented with a combination of the inputs have been used to predict the probability of the input array. In order to test the models, a random sample including ten observations has been extracted from the dataset, and between these ten observations, the following one has randomly been chosen: 'Participants of the movement for black lives conference harassed northeast ohio media group reporter brandon blackwell for recording their public demonstration'.

SUBREDDIT	LOGISTIC REGRESSION PROBABILITY	RANDOM FOREST PROBABILITY	NAÏVE BAYES PROBABILITY
Videos	0.6638	0.5944	0.6545

Table 4, prediction probability scores

By looking at table 4, which reports the prediction probability scores, it is possible to observe that the model having the best performance is Logistic Regression, followed by Naïve Bayes and then by Random Forest Classifier.

4. CONCLUSION

Sarcasm detection is a very interesting and challenging task in natural language processing. In this report the difficulty is increased by the fact that the input was not the comment itself but the parent comment. The choice of a very strict feature reduction is a consequence of the fact that the dataset is huge, and in order to make the algorithms run at acceptable times it is has been necessary to take that choice. It is possible to conclude that the model which also includes subreddits is the one which reports the best accuracy scores. It is possible to conclude that all the models have a 60% probability of predicting the fact that a comment is sarcastic or not. Among the four machine learning models implemented, the one which performs slightly worse is the random Forest Classifier, which has an accuracy score of 0.57 while the other models present an accuracy score of 0.58. After model testing, it is possible to conclude that Logistic Regression and Naïve Bayes, which are simpler models compared to the other two, have the highest probabilities.

Obviously, this paper does not cover a lot of interesting tasks, such as, for instance, the implementation of a Grid Search Cross Validation in order to tune hyperparameters, the evaluation of the model performance only on a random sample of the data and not on the whole dataset, the choice of a lower proportion of observations in the test set in order to compare the performances or the usage of a more powerful machine in order to put a higher limit on the feature extraction and see how the models perform in that case and so on.

REFERENCES

https://link.springer.com/article/10.1007/s10462-019-09791-8

https://dl.acm.org/doi/abs/10.1145/3124420

https://www.kaggle.com/danofer/sarcasm