



SARCASM DETECTION ON REDDIT DATASET


Letizia Molinari 959194

[Final project for the course of text mining
and sentiment analysis]



INTRODUCTION

- *“Light travels faster than sound. This is why some people appear bright until they speak.”* – [Steven Wright](#)
- Sarcasm is 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.



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 category it belongs to.
- The difficulty in this project is increased by the fact that the input was not the comment itself but the parent comment.
- The analysis is performed on the Reddit dataset, which was generated by scraping comments from Reddit website which contained the sarcasm \s tag, which is often an indicator of the fact that a certain comment is sarcastic.



DATASET DESCRIPTION

- - Reddit 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
- After removing duplicates, a balanced dataset having 997883 rows and 3 columns has been obtained, where 50.1% observations are labelled 0 and 49.9% observations are labelled 1.



EVALUATION STRATEGY

- In order to evaluate the performances of the models implemented, five-folds cross validation has been implemented and the dataset has been split in two parts: 60% for training and 40% for testing.
- The analysis has then been divided in two parts and a comparison of the two has been performed:
 - 1. First part: implementation of Logistic Regression, Random Forest Classifier and Support Vector Machine using only parent comments as inputs.
 - 2. Second part: addition of subreddits as inputs in order to determine whether this could improve the performances of the models implemented.



FEATURE EXTRACTION AND REDUCTION

- Parent comments have been vectorized by implementing TF-IDF with tri-grams and putting a limit of 1000 to the maximum features to extract.
- Subreddits have been vectorized by implementing TF-IDF with bi-grams and putting a limit of 1000 to the maximum features to extract
- The two inputs have been concatenated in the second part in a pipeline called `merged_features`, which has been used as input.
- 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 has been necessary to take that choice.



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

As it is possible to observe by the table, the addition of subreddits as inputs to the model has lead to a slight increase in the model accuracy.



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 with a combination of the inputs have been used to predict the probability of the input array.
- In order to test the models, the following random sentence has been extracted :
'Participants of the movement for black lives conference harassed northeast ohio media group reporter brandon blackwell for recording their public demonstration'.



RESULTS OF PREDICTION PROBABILITIES

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

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.



CONCLUSION

- It is possible to conclude that the model which also includes subreddits is the one which reports the best accuracy scores. In particular, 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.