

# Detect sarcastic comments on reddit

Muskan Gupta, Sheenu Mittal, Shreya Bansal, Tanya Sharma, Sejal Banta,  
Neeraj Rani

Indira Gandhi Delhi Technical University for Women, Delhi, India

**Abstract.** Sarcasm refers to the use of satirical or ironic language to convey a message. It is usually used in social networks such as Reddit, Twitter etc. The identification of sarcasm can improve the efficiency of sentiment analysis. Sentiment analysis refers to analyzing the attitude of people towards a particular topic or scenario. Our proposed method uses a supervised approach to learn the sarcastic patterns for classification. So our aim is to built model which helps in detecting sarcastic comments on Reddit using Logistic Regression. We'll be using the dataset from the paper "A Large Self-Annotated Corpus for Sarcasm" with 1.3 million comments from Reddit, labeled as either sarcastic or not.

## 1 Introduction

Sarcasm is employed to convey a meaning different than the literal one, usually in satirical context. It is complex and a bit difficult to comprehend since the actual message in the text has to be interpreted by the user. Sarcasm is utilized for humor as well as criticism of ideas, people or events. Sarcasm actually conveys a different meaning than the actual text. The presence of sarcasm can alter the polarity of the sentiment analysis. Hence, the detection of sarcasm can enhance and refine the existing system for sentiment analysis. We are working on the reddit dataset. Reddit is an online discussion platform, where the community members can post information regarding news, politics, hobbies etc and any other areas of interest. The areas of interests are categorized as subreddits (/news, /politics etc) .Our test data is from the subreddit sarcasm where the members post sarcastic comments which are denoted by appending a /s to the text. To detect the sarcastic comments, the frequently occurring sentence patterns are detected along with its syntactic, semantic and sentiment based features.

**Problem Statement.** *Sarcasm is the unconventional way of conveying a message which conflicts the context. It is both positively funny and negatively nasty, which plays an important part in human social interaction. So our aim is to built model which helps in detecting sarcastic comments on Reddit.*

## 2 Related Work

Mikhail Khodak, Nikunj Saunshi, Kiran Vodrahalli [1] proposed Self-Annotated Reddit Corpus (SARC), a large corpus for sarcasm research and for training and

evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements – 10 times more than any previous dataset – and many times more instances of non-sarcastic statements, allowing for learning in both balanced and unbalanced label regimes. They evaluated the corpus for accuracy, construct benchmarks for sarcasm detection, and evaluated baseline methods.

Soujanya Poria, Roger Zimmermann, Erik Cambria [2] proposed CASCADE (a Contextual Sarcasm Detector) that adopts a hybrid approach of both content and context-driven modeling for sarcasm detection in online social media discussions. For the latter, CASCADE aims at extracting contextual information from the discourse of a discussion thread. CASCADE utilizes user embeddings that encode stylometric and personality features of the users. When used along with content-based feature extractors such as Convolutional Neural Networks (CNNs), they see a significant boost in the classification performance on a large Reddit corpus.

In our project we will be using logistic regression and naive bayes to detect sarcastic comments on reddit.

### 3 Methodology

Given that project is data-driven, the data set is explained as: This dataset contains 1.3 million labeled comments from the Internet commentary website Reddit. The dataset was generated by scraping comments from Reddit containing the ( sarcasm) tag. This tag is often used by Redditors to indicate that their comment is in jest and not meant to be taken seriously, and is generally a reliable indicator of sarcastic comment content. Data has balanced and imbalanced (i.e true distribution) versions. (True ratio is about 1:100). The corpus has 1.3 million sarcastic statements, along with what they responded to as well as many non-sarcastic comments from the same source.

#### 3.1 Dataset Description

We have used data set from Kaggle website which is open source website for data sets <https://www.kaggle.com/danofer/sarcasm>

Table 1 is an example describing the dataset through counts of some key entities involved in the dataset.

Details	Count
Number of unique parent comments	984286
Number of unique comments	962294
Number of authors	256561

**Table 1.** Details of the dataset.

Every dataset also comprises of data attributes. Table 2 describes attributes of data. This project aims to find a solution using logistic regression, a form of supervised learning, hence the *label* in this project is determining if the comment is sarcastic or not. If the comment is sarcastic, a target value of 1 is assigned, else target value of 0 is assigned.

Data Attributes	Brief Explanation
Comment	Reply to the parent comment that need to be stated as sarcastic/non sarcastic.
Author	User Id of the writer of comment
Subreddit	subreddit for the current comment
Score	Count of number of upvotes minus number of downvote
Ups	Count of upvotes
Downs	Count of downvotes
Date	Contains month and year of creation of comment
Created-utc	Timestamp for the comment
Parent-comment	Parent comment for reply comment
Label	Determines if the comment is sarcastic(1) or non sarcastic (0)

**Table 2.** Details of Data Attributes.

### 3.2 Data Pre-processing

The dataset is first tested to remove any null values, if existing. The sarcasm dataset taken from kaggle had null values in the comment column. The comment

Data Attributes	Number of null values
label	0
comment	53
author	0
subreddit	0
score	0
ups	0
downs	0
date	0
created-utc	0
parent-comment	0

**Table 3.** Null values in Data Attributes

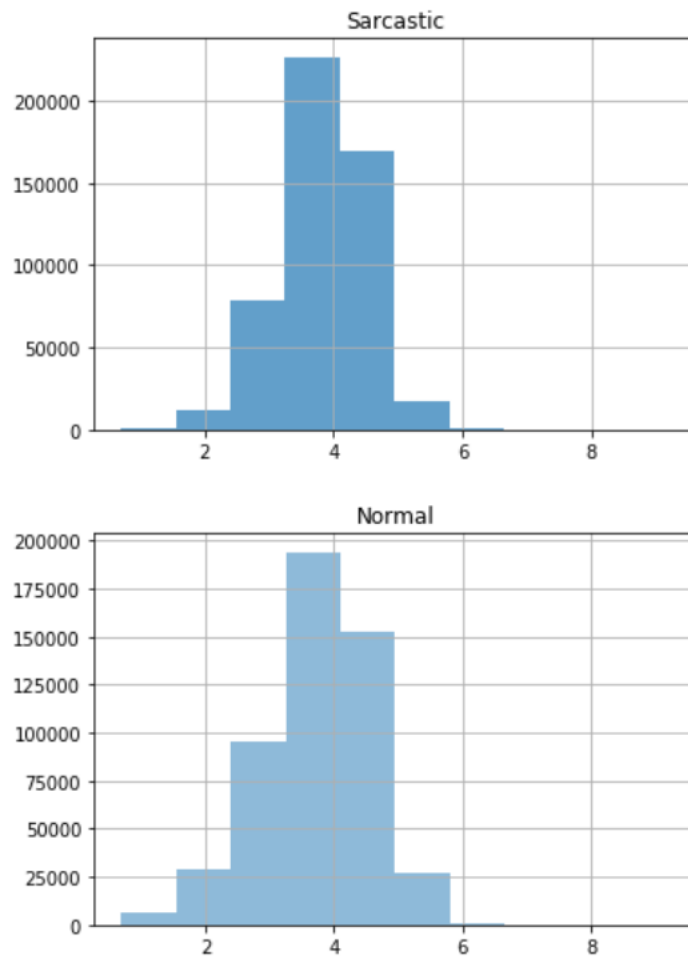
attribute had null values, which were removed from the dataset before processing the data. All the values are unique in the dataset. The dataset also contains outliers, which might be a consequence of some very popular or extremely unpopular comments.

### 3.3 Data Visualization

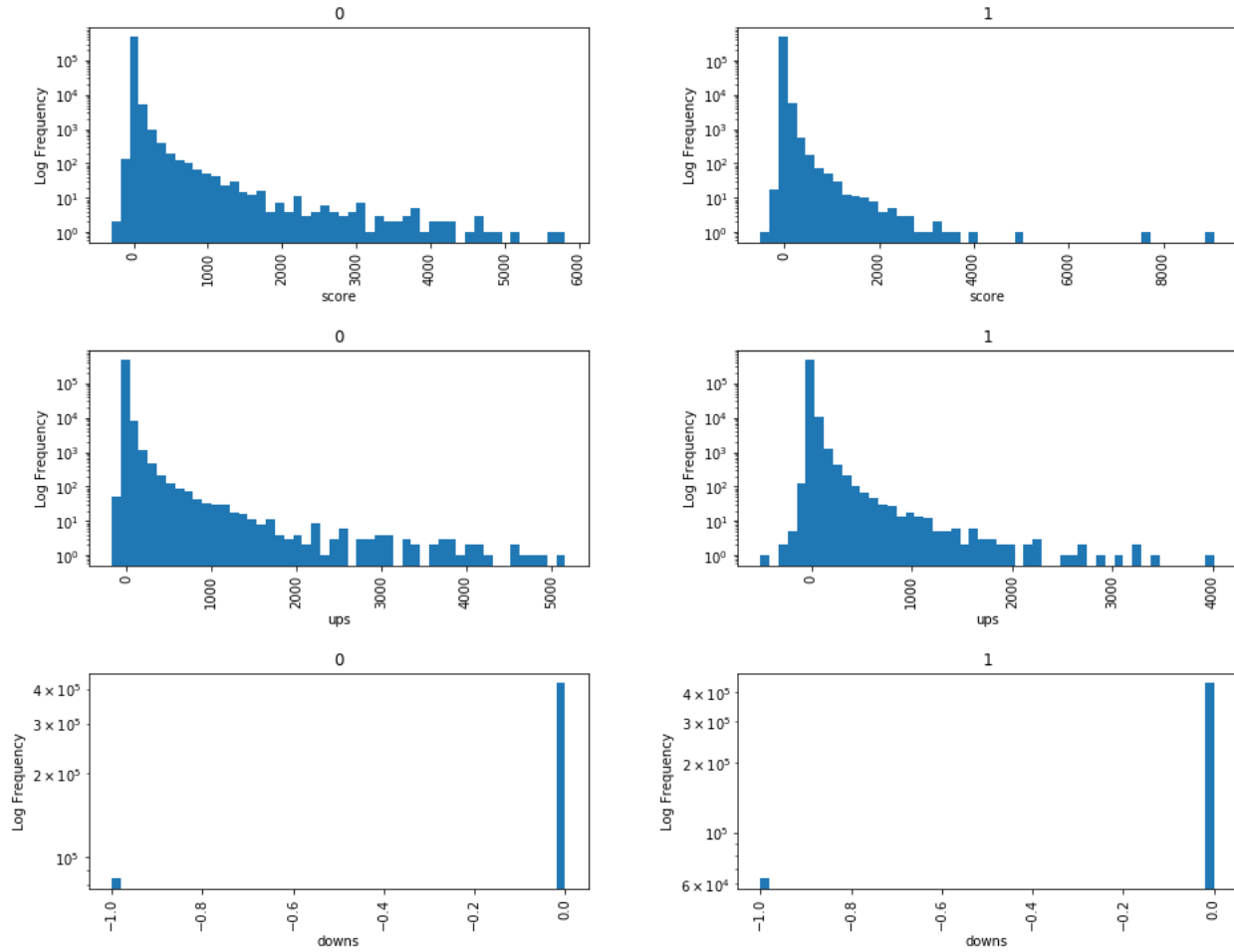
**Step 1** First we observe the number of sarcastic and non-sarcastic comments.

Target Value	Number of rows
0	505405
1	505368

**Step 2** We see the distribution of lengths for sarcastic and normal comments which is almost the same.

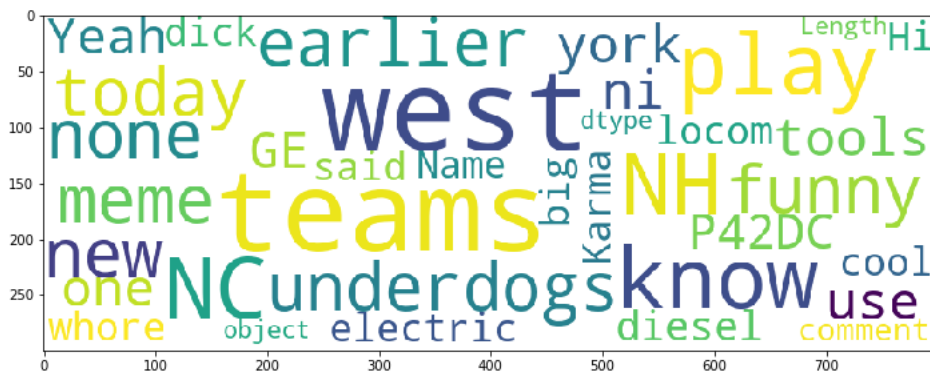
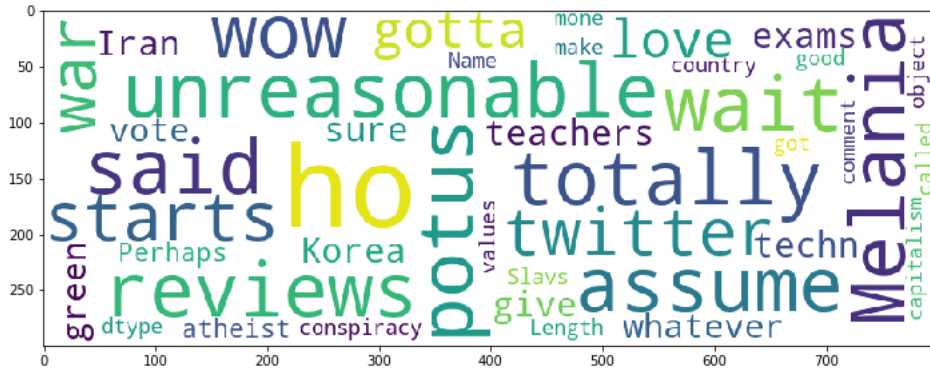


**Step 3** Visualize data by plots that show frequencies of score, upvotes and downvotes for both the classes 0 and 1.



These graphs are frequency vs score, upvotes and downvotes.

**Step 4** Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. We now visualize the word clouds for both sarcastic and non-sarcastic words.



### 3.4 Proposed Approach

## LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

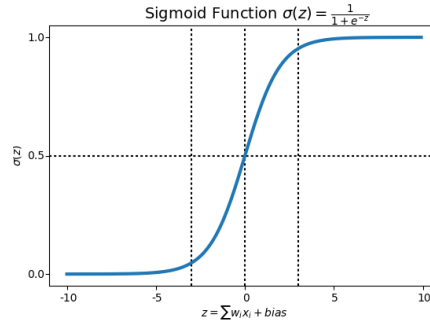
Types of logistic regression:

1. Binary (eg. Tumor Malignant or Benign)
2. Multi-linear functions failsClass (eg. Cats, dogs or Sheep's)

The hypothesis of logistic regression tends to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a

value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



**Fig. 1.** Sigmoid function of graph

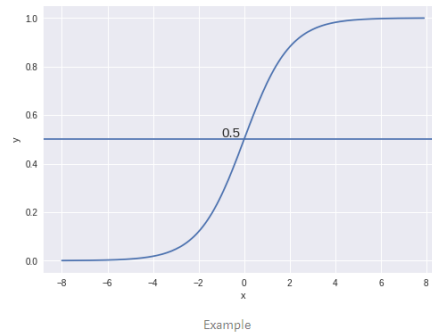
$$f(x) = \frac{1}{1 + e^{-(x)}}$$

**Fig. 2.** Formula of a sigmoid function

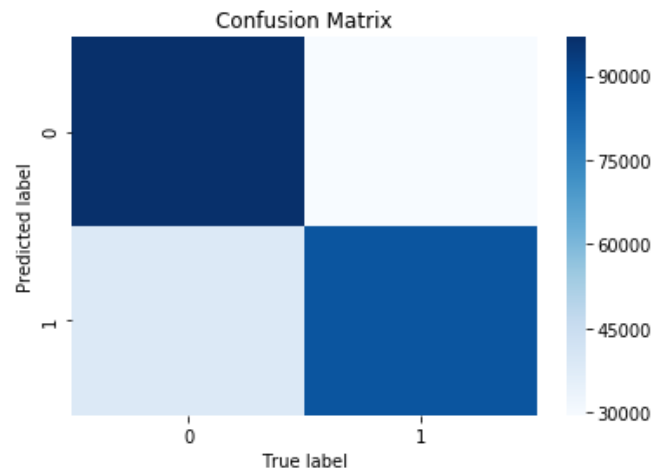
**Example:** We expect our classifier to give us a set of outputs or classes based on probability when we pass the inputs through a prediction function and returns a probability score between 0 and 1. For Example, We have 2 classes, let's take them like cats and dogs(1 — dog , 0 — cats). We basically decide with a threshold value above which we classify values into Class 1 and of the value goes below the threshold then we classify it in Class 2.

As shown in the above graph we have chosen the threshold as 0.5, if the prediction function returned a value of 0.7 then we would classify this observation as Class 1(DOG). If our prediction returned a value of 0.2 then we would classify the observation as Class 2(CAT).

In our proposed project logistic regression is used to classify the comments as sarcastic or non sarcastic. It is selected as sentiment analysis is a binary classification. With the help of this classifier, massive datasets can be executed. In order to train classifier, a manually produce training set is used. An X:Y relation is provided in training set. The variable X represent the reddit comments whereas variable Y represents whether a comment is sarcastic(1) or non- saracas-

**Fig. 3.**

tic(0).Confusion matrix is drawn to represent the accuracy of the model which comes out to be 72.6 percent.

**Fig. 4.** Confusion Matrix

### Naive Bayes

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or



not depending on his/her income, previous loan and transaction history, age, and location.

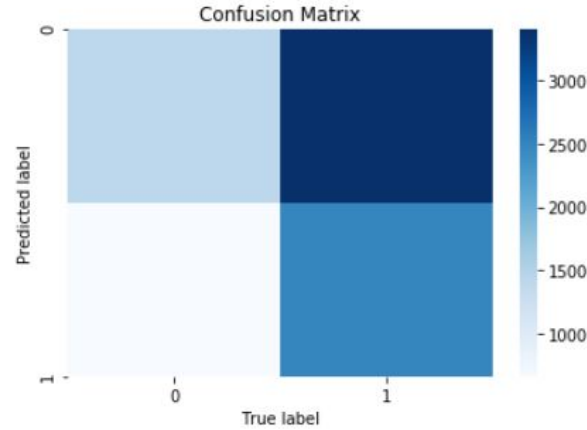


Fig. 5. Confusion Matrix

Accuracy of the test data is: 49.2

## 4 Result

Give details of the format of input and output. Explain the evaluation metrics or parameters.

Provide details of the experiment setup.

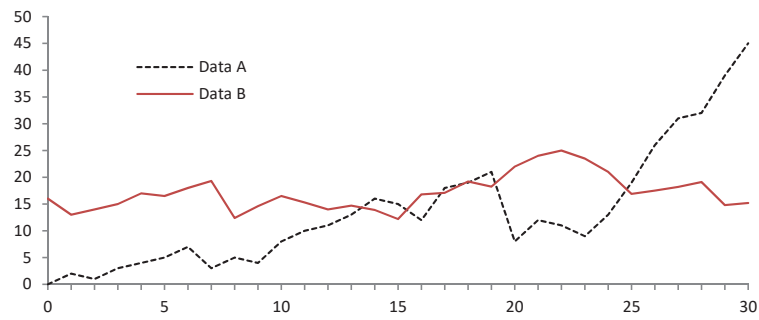
Results should be in form of plots, tables, etc.

If there is any information that you can present in tabular manner, then create a table and refer it in text as Table 4 gives a details of bla bla bla.

Heading level	Example	Font size and style
Title (centered)	<b>Lecture Notes</b>	14 point, bold
1st-level heading	<b>1 Introduction</b>	12 point, bold
2nd-level heading	<b>2.1 Printing Area</b>	10 point, bold
3rd-level heading	<b>Run-in Heading in Bold.</b> Text follows	10 point, bold
4th-level heading	<i>Lowest Level Heading.</i> Text follows	10 point, italic

Table 4. Table captions should be placed above the tables.

Please try to avoid rasterized images (snapshots are to be avoided) for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 6).



**Fig. 6.** A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

## 5 Future Work and Conclusion

Finally write the future work and conclusion.

## References

1. Kaushal, R., Singh, A.: Automated evaluation of programming assignments. In: 2012 IEEE International Conference on Engineering Education: Innovative Practices and Future Trends (AICERA). pp. 1–5. iee (2012)
2. Sinha, P., Zoltners, A.A.: The multiple-choice knapsack problem. Operations Research **27**(3), 503–515 (1979)