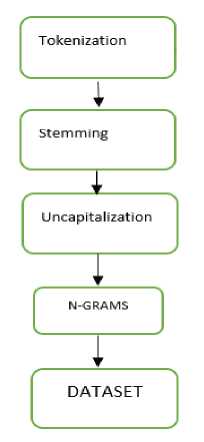
**Automatic Sarcasm Detection using feature selection**

The primary goal of this research paper is to understand different methods for Automatic sarcasm detection. Further, this paper describes techniques used for collectively studying the various approaches available in Automatic Sarcasm Detection Field.

IV. FEATURE ENGINEERING:

1. Tokenization: Tokenization is one of the main features of lexically analyzing the text. Here the aggregation of the sequence of characters takes place. Decluttering the words of the sentence into tokens is very helpful in individual identification these counts of tokens are very helpful. A token could be a complete group of words or a whole paragraph, but most frequently, we use words as tokens. E.g. Using Hashtag-Tokenizer to separate the hashtags that contain connected words, e.g., splitting #Sarcastic irony into #sarcastic and #irony.
2. . Stemming: It is a common method in Natural Language Processing (NLP). The main motive of this is to reduce the repetition of words by dropping the suffix of the words to arrive at a basic form of the word
3. . n-grams: n-grams is the group of co-occurring n-items in a sequence, which can be speech or text, used largely for NLP purpose
4. Sentiments: There is an emotion associated with each post or tweet we see. Some can be positive while the others are negative. The sarcastic posts we see are more negative as compared to the non-sarcastic posts. This makes it very important for analyzing each part of the text. Sentiment analyzers that are available analyze each token very carefully and thus make the categorization of text easy
5. Recognition of Pattern: 1. Exact Match: all the words in the patter nappear contiguously in the right order without any additional components.

A: Sparse Match: all the words in the pattern are found, but there are some additional words present.

Y.n/N: Incomplete match: only n>l of the total N patterns appear in the sentence while some non- matching words can be inserted in between. A special condition is that one word in this must be the high-frequency word.

0: No match: Nothing or only a single pattern component appears in the sentence.

FEATURE SELECTION:

Term frequency-inverse document frequency (TFIDF): it is a part of information retrieval and shows that how important a word is to a corpus. The TFIDF value for a word in a document increases with the increase in the number of times the word appears in the document. At the same time, the words which are available in all the documents tell a very little or no information about the document.

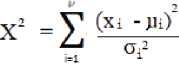
Chi-square: is a statistical method used to test the independence of two categorical variables. If two variables in a document are similar, then a low result is obtained, and the higher result would be obtained for words that differ .Definition: If v

idf(t,D) = log

|{de-D :te-d}|

Independent variables x are each normally distributed with mean p and variance o;2, then chi-squared is defined as:

Here if the chi-squared method will help in finding out the



Words which have identical distribution in documents and thus irrelevant.

1. CLASSIFICATION

Support Vector Machines (SVMs): These are the new machine learning technique. One feeds labeled training data set to the SVM to get an optimal hyperplane [4], which can help in classification and categorization of new data. On both sides of the hyper plane, we have instances (which consists of data points in space belonging to the two classes). After finding a hyper plane, the distance maximizes (distance between the nearest data point of each class and the hyper plane) between the two classes and the hyper plane

Logistic regression: Logistic regression is a statistical method used prominently in machine learning. [7]

It is a problem involving binary classification ,i.e., which can take two values. Measurement of the outcome is done with the help of dichotomous variables. E.g., Email: spam or not. In both these cases, two values for each case are possible. The First case can have two outcomes,i.e., spam and not spam (dichotomous).

Bamman and Smith used logistic regression for binary (sarcastic or non-sarcastic) classification of tweets. For example, while validating models bases on feature-type categorization takes place by capitalization feature of Twitter. Logistic regression for binary classification is dependent on

Naive Bayes: Naive Bayes Classification is very famous classifier because of the simplicity of the use. It is the most basic form of text categorization. Naive Bayes is a conditional probability model. The classifier makes the use of Bayes theorem and makes naive estimations. It exists for spam filtration purposes also. One downside of this is that it is not very trustworthy as probability outputs can be inaccurate sometimes.

FINAL EVALUATION AND ANALYZING

Testing of the model is essential. For this purpose, we divide the entire dataset into three parts, i.e., Training set, Validation set and Testing set. The training set consists of the data with the help of which the model is trained. This dataset consists of input as well as the correct output. The dataset is referred to as Gold Standard. The validation set is used to find the model which best fits the training set by using the algorithms that give the best results. It shows the extent to which the model has been trained. The third is the testing set. This consists of the real world data, and the model is tested against this data and results are found. [10]This is done by checking the accuracy of the proposed model.

|  |  |
| --- | --- |
| Feature | F-Score |
| n-grams | 0.56 |
| sentiments | 0.41 |
| topics | 0.35 |

Although these methods can help in sarcasm detection, they are not entirely efficient in detection of sarcasm and categorization of the texts.

Naive Bayes method also does not lead to higher precision, and the chances of poor results are quite high in this. Methods like, SVM need to be more precise in their analysis because these can provide more accuracy. However, the main issue with SVM is both the speed and the size (both in training and testing). Logistic regression also can provide excellent results but not as efficient as SVM, Logistic regression as discussed relies on the independence of the data points. Also, Logistic regressions cannot be used for continuous outcomes. Development of new techniques and methods can provide users with more than one feature and better results.