DOI: 10.20472/IAC.2018.042.041

**RAJANI**
**Kalindi College, India**

**PANKAJ SAMBYAL**
**Kalindi College`, India**

**SHALINI SHARMA**
**Kalindi College, India**

# SENTIMENT ANALYSIS AND CLASSIFICATION OF TWEETS BASED ON IDEOLOGIES

## Abstract:

In this paper, we use data mining and sentiment analysis techniques to classify the tweets based on different ideologies i.e. Secularism, Liberalism, Communalism, Socialism and Casteism. To analyze our model, we used tweets from three sources namely generic Indian tweets, a specific user profile tweet and tweets of particular hashtags.
The tweets are fetched using Twitter API. The fetched data is preprocessed by analyzing structure of tweets to find interesting analysis like most retweeted tweet, most favorited tweets, trending hashtags etc. Then tweets are tokenized and POS (parts of speech) tagging is done on tokens to find nouns, verbs, adverbs and adjectives which are relevant for the analysis.
We apply various relevance models on the data, to find sentiment of each tweet and ideological stance of the user. The results are shown using spider graph. It was observed that the model worked with 73% accuracy.

## Keywords:

Multiclass, data mining, twitter, hashtag

## 1. Introduction

Today a lot of opiniated data is available on social media like facebook, twitter and quorae.t.c. Among all these social media platform twitter is one of its kind. It is widely popular and generally used to publish short opinions. With these limited words still people are able to express their feelings, emotions and ideologies. This paper is an attempt to build a model that can analyze a tweet and then classify the same in to pre-defined set of ideologies. A lot of work has been done on polarity classification and emotion classification of tweets. But they generally focus on two types of classes. One being a positive class and other being a negative class. Some may even introduce a third class like neutral one. We tried to go beyond a step forward. Instead of these traditional classes we build a model which does multi-classification of a tweet. These classes are basically ideologies. The ideologies which we considered in our model are following: Secularism, Liberalism, Communalism, Socialism and Casteism.

The ideologies that we considered are important pillars of nation building. We have not ruled out the ambiguous nature of English language as in English one sentence or word can be used for more than one context. That is why while classifying a tweet in any of the class a weighted score of each tweet was assigned for each of the five ideology classes that we considered. This study is first of its kind and we hope it will open up a new frontier of research and will prove to us more beneficial in decision making and opinion making from social media.

The paper is organized as follows: section 2 contains the related work. Section 3 explains our model. Section 4 contains are result of analysis and section 5 contains conclusion and future scope. Finally, in section 6 References are provided.

## 2. Related Work

In [1] authors proposed a model to analyze tweet of the user for predicting sentiments of online transportation services. They only predicted tweets as positive and negative polarity. In [2] authors classified tweets as positive, negative and neutral for predicting opinion of the people. They used feed forward neural network for the task. In [3] authors tried to develop opinion prediction of smart-phone user experience. The experiments have proved that the results improve by around 2 points on an average over the unigram baseline. In [4] Sentiment analysis was done on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program. In [5] techniques of NLP are used for sentiment analysis of tweets. Here sentiment classification is done by adding semantics in feature vectors and thereby using ensemble methods for classification. In [6] authors developed a tool for sentiment analysis.

## 3. Proposed Methodology

Fig. 1 shows proposed model diagrammatically.

Preprocessing of Tweets

First we retrieve the tweets of a particular user or some trending hashtag or generic twitter data about 2000-3000 tweets are retrieved during one access using the twitter API. Also using the twitter api at –max 14 accesses are provided at a time.

Following steps are taken while the pre-processing of the tweets:

1. Remove all URLs (e.g. www.example.com), hash tags (e.g. #topic), targets (@username), special Twitter words (“e.g. RT”).
2. We also remove the stop-words from the tweets using the using the nltk library of python.
3. Remove all the punctuation symbols from the tweets.
4. Perform the POS (Part-of-speech) tagging for each of the words present in the tweet to categorize them as noun, pronoun, verb, adverb etc.
5. POS tagging is done with the help of the nltk’slibrary’s pos_tag method which generates a pair consisting of a word and the corresponding part of the speech to which the word belongs to.

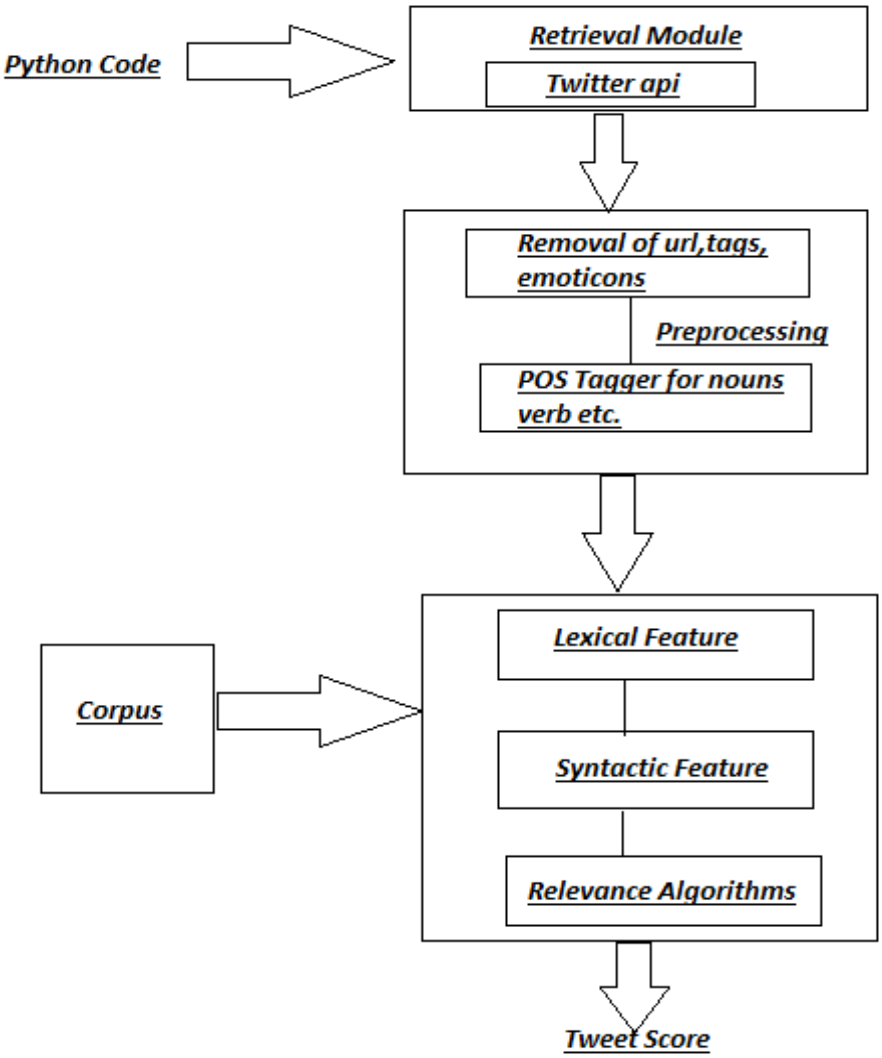
After the pre-processing of the tweets, we calculate for each of the tweets the corresponding features define above i.e. F1, F2, F3 using the models described below.

**Data Set**

We created the corpus of the different classes defined i.e Secularism, Socialism, Communalism, Liberalism, Casteism.

The corpus consisted of various papers on the different issues related to the defined classes along with the journals and research papers .We removed the stopwords from the corpus and each corpus consisted of the part-of-speech components which are relevant for assigning scores to the tweets on the basis of the various relevance algorithm.

**Figure: 1 Proposed Model**




### Types of Features

There have been many studies on measuring text quality and many features have been proposed to capture text quality.

### Lexical Feature:

Lexical features aim to capture the lexical usage of a piece of text compared to some reference corpus. Here we have used a lexical feature based on unigram language models, which provide a principled way to statistically model text.

Specifically, it is assumed that there is a reference corpus that represents high quality text, e.g. a corpus of a well-known Journal articles.

A unigram language model, denoted as θr, can be estimated from this reference corpus. The lexical feature is defined as the log likelihood of the comment based on θr, calculated as:

$$\sum n(w,c)\log P(w|\theta r)$$

Where P(w|θr) is the probability of word type w according to θr, and n(w,c) is the number of times word type w appears in comment c.

In a unigram model we are mainly concerned with the occurrence of a single word in the corpus or the document used as the reference.

This is termed as feature F1.

**Syntactic Feature:**

According to the research by Pitler and Nenkova(2008) the average number of verb phrases per sentence is a useful feature with high correlation with text quality.

So, the second feature we use for our study is the average number of verbs per comment.

We also experimented with other syntactic features like average number of noun phrases and noun to verb ratio calculated from the user's comments. This is termed as the feature F2.

We found that the average number of verbs per comment had and the average no. of nouns per comment has the highest correlation with comment quality, and therefore we do not consider these other syntactic features while measuring the similarity with a particular class.

**Relevance Feature:**

One of the important differences between our problem and standard text quality assessment is that the quality of a comment also relies on its relevance to the target of the comment.

In our problem definition, the target is also a piece of text. For example, consider comments made to Obama's State of the Union speech. A comment such as "We are very lucky to live in the USA. I always have and always will support our president" is not directly related to any issue addressed by Obama in his speech, and therefore is not considered to be a thoughtful comment. Hence, for precise prediction of comment belonging to a particular class out of the 5 classes defined , we also consider a relevance feature in addition to text quality features.

The KL-divergence score between a comment c and a target document d is defined as the KL-divergence between the unigram language models θc and θd estimated from c and d, respectively:

$$Div(\theta c||\theta d) = \sum p(w|\theta c) \log (p(w|\theta c) / p(w|\theta d))$$

## 4. Results

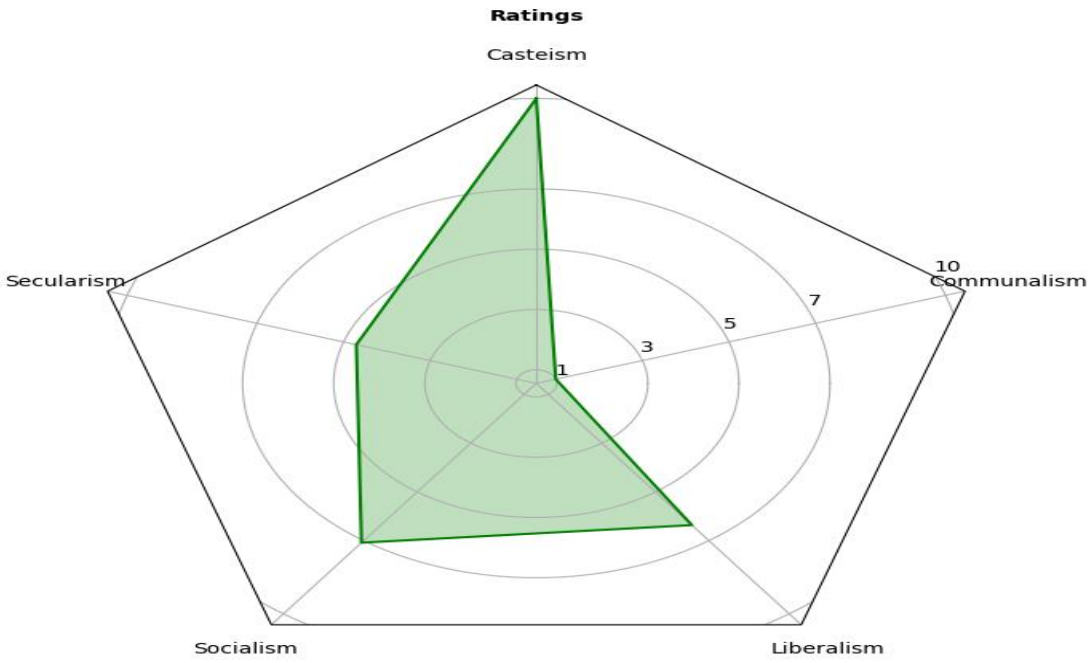
For the classification of the user tweets into the five defined classes, we fetch the user tweets using the Twitter API using the screen-name of the user and after fetching the tweets pre-processing is done for the tweets in order to remove the urls, hashtags, emoticons etc.

Following is the implementation for fetching the tweets of a specific user using the twitter API.

The Screen name used is “ VijayGoelBJP ”

Fig.2. shows results are obtained on applying the KL-divergence model to the user tweets obtained. Here approx. 2500 tweets have been processed to obtain the score of the tweets for each class.

**Figure: 2. Shows results of ScreenameVijayGoealBJP**




For analyzing the generic twitter data, we retrieved about 3000 tweets from INDIA and the results of applying the KL-Divergence relevance model are shown in figure 3.

Figure 4 shows results are obtained by the application of the KL-Divergence model to the retrieved trending hashtag tweets. The hash tag for which result are shown is #secular.

**Figure: 3 Results of generic twitter data**

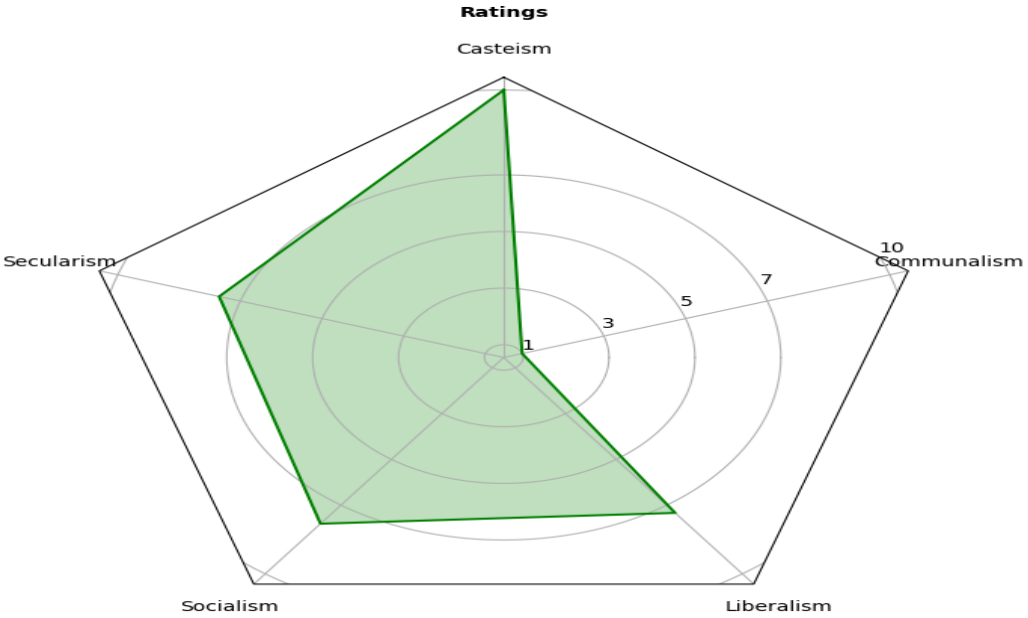
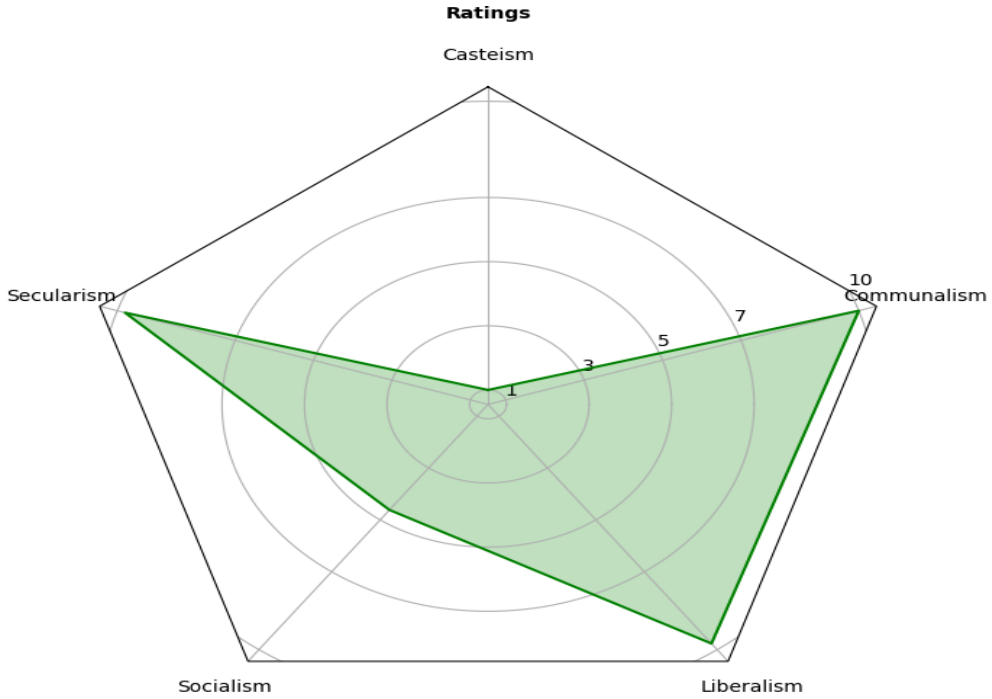



**Figure: 4 Results of #Secular**




## 5. Conclusion

In our Project, we intend to classify tweets on the basis of various ideologies i.e. Secularism, Communalism, Liberalism, Socialism and Casteism by using three different

methods. In the First method, we fetched a large amount of tweets from the twitter and then classified those tweets on the basis of different ideologies. Then in the second method, tweets related to trending hashtags in the last 24 hours were fetched and then they were classified on the basis of same ideologies as described above. In the third and the last method, same classification was performed on the tweets fetched from a particular user profile.

The twitter data was fetched with the help of an API named “tweepy”. The API enabled us to fetch the twitter data based on different hashtags. After fetching the data redundant words were removed from it like URLs, tags, emoticons, etc.

After preprocessing, the data obtained consists of only noun, verbs, adjectives, etc. known as Parts of Speech. This was the main feature used for the organization of data. Ultimately, a lexically and syntactically correct corpus was obtained. Then various models namely kl-divergence, Itakura-Saito and Hellinger distance were used to obtain the twitter score from the retrieved tweets. The result was represented in the form of Bar Graph and Spider Graph. The results obtained from these models were then checked manually for verification. We assigned scores belonging to a particular class and then compared that with the results obtained from the bar chart. From the calculations the accuracy comes out to be 73% for kl-divergence.

## References

D. G. Manju Venugopalan, "Exploring sentiment analysis on twitter data," in *IEEE*, Noida, India, Eighth International Conference on Contemporary Computing (IC3) .

F. N. U. K. I. S. Ike Pertiwi Windasari, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," in 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) , Semarang, Indonesia , 2017.

M. K. S. V. E. Tiara, "Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program," in *3rd International Conference on Information and Communication Technology (ICoICT )* , Nusa Dua, Bali , 2015.

M. A. M. O. F. W. Onifade, "SASM: A tool for sentiment analysis on Twitter," in *2nd World Symposium on Web Applications and Networking (WSWAN)* , Sousse, Tunisia , 2015.

R. M. R. G. Monisha Kanakaraj, "NLP based sentiment analysis on Twitter data using ensemble classifiers," in *3rd International Conference on Signal Processing, Communication and Networking (ICSCN)* , Chennai, India , 2015.

Y. Z. Brett Duncan, "Neural networks for sentiment analysis on Twitter," in *IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)* , Beijing, China , 2015.