International Journal of Research in Engineering, Science and Management
Volume-2, Issue-4, April-2019
www.ijresm.com | ISSN (Online): 2581-5792

Prediction of Indian Election Sentiments on Twitter using Machine Learning

Rohith Nair[1], Sharan Rai[2], Vickson Rodrigues[3], Shivam Soni[4], Manasi Kulkarni[5], Shubhangi Rathod[6]
[1,2,3,4]*Student, Department of Computer Engineering, PCE, Navi Mumbai, India*
[5,6]*Professor, Department of Computer Engineering, PCE, Navi Mumbai, India*

***Abstract*: Sentiment analysis is considered to be a category of machine learning and natural language processing. It is used to extract and recognize opinions from different content structures, including news, audits and articles and categorizes them as positive, neutral and negative. The main objective here is to provide insights about the public opinion about different political parties and predict the polarity of opinions towards different political parties. In this proposed implementation we perform sentiment analysis of opinions and views on political parties and candidates posted on Twitter, a microblogging service. For this project we aim to explore deep learning techniques such as RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network) and do a comparative study with traditional machine learning algorithms used for sentiment analysis such as SVM.**

***Keywords*: Sentiment analysis, Natural Language processing, Machine Learning and RNN (Recurrent Neural Network) SVM (Support Vector Machine)**

1. Introduction

Sentiment analysis is considered to be a category of machine learning and natural language processing. It is used to extricate, recognize, or portray opinions from different content structures, including news, audits and articles and categorizes them as positive, neutral and negative. Our aim is to apply sentiment analysis on tweets gathered from Twitter. As Twitter is a popular micro-blogging social media platform, many people express their likes or dislikes for a political party. We use RNN (Recurrent Neural Network) which is traditional deep learning technique to calculate the sentiment of political tweets in the data corpus collected from twitter. The result of the analysis is displayed to the user using a graphical representation with the help of android application.

2. Proposed work

The proposed system performs sentiment analysis on the data collected to predict the general sentiment of the public towards each political parties over the period of time leading up to the elections. The system aims to perform aspect based analysis to understand the subjects the tweets are about, the subject could be any of the policies by the political parties which would help to analyze how the policies or ideas that the political parties propose to implement are received by the public. The system shows how the opinion of the voters change with each major event during the respective campaigns of the political parties, the events could be rallies organized by the political parties or debates participated by the party candidates. The tweets will be represented using multidimensional vectors generated using Word2Vec model and a comparative study on the different algorithms such as RNN CNN and SVM would be done to determine the best algorithm for sentiment analysis on tweets.

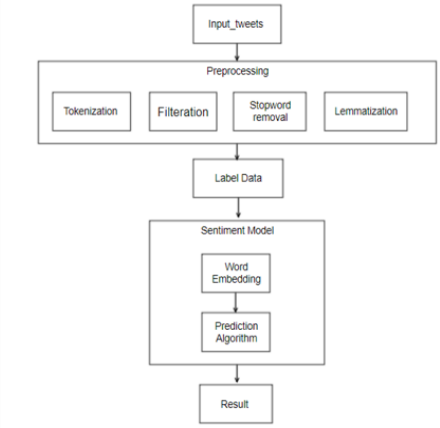

Fig. 1. Prediction of Indian elections

- *Input:* The input here are the tweets related to the 2014 Indian elections collected from twitter.
- *Preprocessing:* The data retrieved from twitter is in json form it needs to be converted into tabular form and the data needs to be cleaned to input for further processing. Some of the steps involved in preprocessing are detailed below.
- *Tokenization:* Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded.

Algorithm:
Input: A single sentence
Output: A list of tokens
Input:
No CM in Indias history has tried harder to bring his Govt

down Success at last Arvind Kejriwal can now get down to Lok Sabha campaign

Output:

"No" "CM" "in" "India's" "history" "has" "tried" "harder" "to" "bring" "his" "Govt" "down" "Success" "at" "last" "ArvindKejriwal" "can" "now" "get" "down" "to" "Lok" "Sabha" "campaign"

Filtration:

This is done to remove the special characters(@,!,&,$ etc) and numbers as they don't convey much information.

Algorithm:

1. Input
2. if words in sentence == Filtration list
   then goto step-4
3. else message ("No filtration is present") then goto step-4
4. output
5. Exit

Input:

RT @Sumit_Nagpal: The real worry of BJP & Congress is not what if @ArvindKejriwal becomes the CM, their worry is what if he delivers what h…

Output:

RT Sumit_Nagpal The real worry of BJP Congress is not what if ArvindKejriwal becomes the CM their worry is what if he delivers what h

Stop Words Removal:

Stop words are words which are filtered out before or after processing of natural language data. Though "stop words" usually refers to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search.

A. *Sample stop word list*

{ 'he', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'and', 'on', 'very', 'having', 'will', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'any', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below'}

B. *Algorithm*

1. Input
2. If words in sentence == stopwords list
   then goto step-4
3. else message ("No stopwords")
   then goto step-4
4. output
5. Exit

Input:

By far the best analysis on gas pricing and exposing kejriwal shoot and scoot wonder he will answer any

Output:

Such etadalal by far best analysis gas pricing exposing kejriwal shoot scoot wonder answer.

C. *Lemmatization*

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

English lemma list:
have -> had,has,'ve,having,'s,'d,of,d,ve
it -> its,they
he-> his,him,they
i -> my,me,we,is
they -> their,them,'em
you -> your,ya,ye
not -> n't
she -> her
do -> did,does,done,doing,du,d'

Algorithm

1. Input
2. If words word in sentence == lemma list then goto step-4
3. else message("already in lemma form") then goto step-4
4. output
5. Exit

Input:

BDUTT Aam Aadmi Party Corruption the main issue AAP is fighting for has been rampant in congress regime NOT BJP AK losing my trust respect.

Output:

BDUTT Aam Aadmi Parti Corrupt the main issue AAP is fight for ha been rampant in congress regime NOT BJP AK lose my trust respect.

D. *Word Embeddings*

A word embedding is an approach to provide a dense vector representation of words that capture something about their meaning. Word embeddings are an improvement over simpler bag-of-word model word encoding schemes like word counts and frequencies that result in large and sparse vectors (mostly 0 values) that describe documents but not the meaning of the words. Word embeddings work by using an algorithm to train a set of fixed-length dense and continuous-valued vectors based on a large corpus of text. Each word is represented by a point in the embedding space and these points are learned and moved around based on the words that surround the target word. It is defining a word by the company that it keeps that allows the word embedding to learn something about the meaning of words. The vector space representation of the words provides a projection where words with similar meanings are locally clustered within the space. The use of word embeddings over other text representations is one of the key methods that has led to breakthrough performance with deep neural networks on

problems like machine translation. Here we will use word2vec embedding method to convert the textual data into multidimensional vectors which is created by google using millions of wikipedia documents.

E. Prediction Algorithm

1) RNN

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

2) LSTM

Long short-term memory (LSTM) networks were discovered by Hochreiter and Schmidhuber in 1997 and set accuracy records in multiple applications domains. Around 2007, LSTM started to revolutionize speech recognition, outperforming traditional models in certain speech applications. In 2009, a Connectionist Temporal Classification (CTC)-trained LSTM network was the first RNN to win pattern recognition contests when it won several competitions in connected handwriting recognition. In 2014, the Chinese search giant Baidu used CTC-trained RNNs to break the Switchboard Hub5'00 speech recognition benchmark without using any traditional speech processing methods.

3) SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

4) CNN

In machine learning, a network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

5) Result

The result is shown for the user in the form of an android app that lets the user know about the different trends about the election and the prediction about the election.

3. Implementation details

The tweets are collected using a pre-existing collection of tweet id's. The tweet id's are passed to a tweet hydrator app which fetches the tweet corresponding to each tweet id. The hydrated tweets are returned in json form containing the tweet text and many other details. The required details are extracted from the json tweet output. The data is then segregated according to the respective political party and the required cleaning is done. The dataset is then labelled. A time series split of data is done into training and validation data. The training data is fed into the classifier model in the form of multidimensional vectors created using word2vec model. The model then is used for the prediction.

Table 1
Sample Dataset

Created_at	Tweet_id	Full_text
Sat Mar 15 18:10:12 +0000 2014		@SushmaSwarajbjp sushma ji jehan per samman naa mile , vehan brahmano ko nahi rehna chahiye
Tue May 27 02:50:30 +0000 2014	61126132	RT @SriSri: Blessings & Best Wishes to Narendra Modi & his team of Ministers.May God give them the strength & wisdom to fulfil the high hop…
Sat May 10 16:13:48 +0000 2014	18839785	RT @narendramodi: I am overwhelmed by people's response! I assure them we will repay their affection with unprecedented development.
Wed Jul 02 03:02:05 +0000 2014	405427035	@ArvindKejriwal right action wud b 2 demand action again thieves of previous regime n keep note of present ones for next regime.
Sun May 18 20:06:24 +0000 2014	24705126	@ShashiTharoor if u wer so knowledgeable abt foreign policies, today ur govt wudnt hv faced such defeat @BDUTT

4. Requirement analysis

A. Software Requirements

- *Python* - python is an interpreted, object-oriented high-level programming language. It emphasizes code readability and reduces code maintenance making it suitable for Rapid Application Development. *TextBlob, Spacy* - Python libraries for dealing with textual data.
- *Scikit learn* - Python library for analysis of data
- *Keras* - Python wrapper for using Tensor flow.
- *Android, flask* - For creating android dash board.

B. Hardware requirements

- Amazon Web Services - For working with large amounts of data

5. Conclusion

A comparative study on different algorithms used for sentiment analysis on tweets is done. Most of the current implementations have seldom explored deep learning algorithms like RNN (Recurrent Neural Networks) and CNN (Convolutional Neural Networks) here we look to see how neural networks compare with the traditional algorithms like SVM (Support Vector Machine). The dataset contains about 22 million tweets which are used for analysis and prediction of the sentiment of the general twitter users towards each political party and how this reflects the sentiment of the general population. Volume analysis is performed to recognize the different trends found with the frequency of the tweets. The results are shown in the form of an interactive android application.

References

[1] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, 2017, pp. 1-4.

[2] P. Sharma and T. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 1966-1971.

[3] Yoon Kim, “Convolutional neural networks for sentence classification,” 2014.

[4] Abhishek Bhola "Twitter and Polls: Analyzing and estimating political orientation of Twitter users in India General Elections 2014," 2014.

[5] A. Timmaraju, V. Khanna, "Sentiment analysis on movie reviews using recursive and recurrent neural network architectures", Semantic Scholar, 2015.

[6] Socher, R & Perelygin, A & Wu, J.Y. & Chuang, J & Manning, C.D. & Ng, A.Y. & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP. 1631. 1631-1642.

[7] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle,” in Proceedings of the ACL 2012 System Demonstrations, ACL’12, (Stroudsburg, PA, USA), pp. 115–120, Association for Computational Linguistics, 2012.

[8] Tumasjan A, Sprenger T. O, Sandner P. G,Welpe I. M, “Election forecasts with Twitter: How 140 characters reflect the political landscape,” Soc Sci Comput Rev. 2011;29:402–418.

[9] Pak, A. and Paroubek, P, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” Valetta, pp. 1320, 2010.