Text Analytical Models for Data Collected from Micro-blogging Portal – A Review

[1]Jasandeep Kaur, [2]Dr. Rajeev Kumar
[1]Research Scholar, Department of Computer Science and Engineering,
[2]Assistant Professor, Department of Information Technology,
[1,2]DAV Institute of Engineering and Technology, Jalandhar, 144009, India.

***Abstract*:** The text analytics are being used in a number of applications ranging from trend detection to sentiment analysis. These models are applied over the different kinds of data, which involves the limited length text responses from micro-blogging portals (e.g. Twitter) and product reviews. Most of the text analytics services are used to determine the public sentiment for the target products, political issues, news, etc. But the major problem in the case of sentiment analysis is caused by the sarcastic content, which depicts the different context, but caries a different meaning. Hence, it's very important to accurately detect the sarcastic tweets is increasing every year and decreasing the accuracy of the existing text analytical engines. This paper presents a survey on phases for sarcasm detection and also discusses various approaches based upon the combination of multiple features for classifying the text. This paper targets to improve the overall sarcasm detection accuracy by using the maximum possible feature combinations.

***Keywords*: Sarcasm detection, Support Vector Machine, Twitter analytics, Maximum Entropy.**

I. INTRODUCTION

In order to convey suggestions and emotions on different areas like products and events, social networking websites act as a prominent source for the people who operate it. In the same way, another mode for the same purpose is Twitter. In these sites, users put 340 million and more or 1.6 billion search queries on daily basis. However, it is basically not so easy to find out sarcasm in those tweets as it covers 140 characters per tweet along with in formal language like, hashtags, emotions, slangs, etc. which make it arduous to judge varied masses opinions [2][3].

Perspective of the user in the particular areas is known as sentiment analysis. It helps to find out polarity which is to know whether it covers positive, negative or neutral emotions. To detect sarcasm is the most challenging task under sentiment analysis. To convey something opposite to one statement in order to hurt someone's feelings in a verbal manner on social media is basically a sarcasm device. Sarcasm can be clearly defined under three areas such as:

- Sarcasm is used as wit for the purpose of being funny. Often this is expressed when use with exclamation marks, question marks, capital letter words, sarcastic emoticons like-:P.
- Sarcasm is used as whimper to represent how bothered (annoyed) or irritate the individual is.
- Sarcasm is used as evasion in those situations when someone wants to hide or refrain presenting a fair respond and creates use of it by using unusual words, intricate sentences, etc. [6].

II. LITERATURE SURVEY

The purpose of [1] is to present sentiment analysis on film reviews by using hybrid approach which involves a machine learning algorithm namely support vector machine (SVM) and a semantic oriented approach namely sentiwordnet. The

features are extracted from sentiwordnet. Training of support vector machine classifier is done on these features. Film reviews are classified by support vector machine afterwards. To determine the sentiment orientation of the film reviews, counting of negative and positive term scores has been done.

In paper [5], the authors proposed a supervised sentiment classification framework in which four basic features types are utilized for sentiment classification i.e. single word features, n-gram features, pattern features and punctuation features.

In paper [12] textblob is used for preprocessing which includes tokenization, part of speech tagging, parsing and by using python programming stop words are also removed. For polarity and subjectivity of tweets RapidMiner is used and weka tool is used for calculating the accuracy of tweets with the help of two classifiers i.e. Naïve Bayes and SVM. At the end, naïve bayes provides more accuracy as compared to SVM.

In this paper Soujanya Poria et.al. [8], developed a model for sarcasm detection on the basis of pre-trained CNN for the extraction of sentiment, emotion and personality features. Two kinds of experiments have done: firstly, they used CNN for the classification and secondly features are extracted from the fully-connected layer of the CNN and fed them to an SVM for the final classification. The experimental results show that the baseline features outperform the pre-trained features for sarcasm detection and also show that the sentiment and emotion features are the most useful features besides baseline features.

In this paper different supervised classification technique is identified by Anandkumar D. Dave et.al. [10] for sarcasm detection and also train SVM classifier for 10X validation along simple Bag-of-words as features and use TF-IDF for frequency measure of the feature. Two datasets were collected (Amazon product reviews and Tweets) and pre-processing also done for the removal of noise (spelling mistakes, slang words, user defined label, etc.) present in the dataset.

III. PHASES OF SARCASM DETECTION

- Dataset Formation: It is the first step in which dataset can be collected from different sources e.g. twitter or posts from Facebook.
- Data Preprocessing: In this case, cleaning of data is performed such as removal of URLs, hashtags, tags in the form of @user and unnecessary symbols.
- Sarcasm Identification: It involves two different phases i.e. feature selection and feature extraction. Feature extraction involves Part of speech, Term presence, Term frequency, Inverse document frequency, negation and opinion expressions for extracting the features. On the other hand, lexicon method and statistical method are use in case of feature selection.

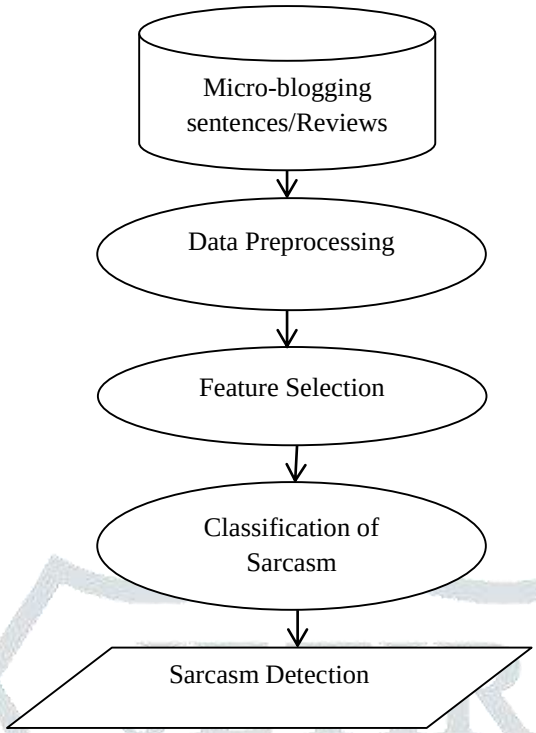

Fig 1.1: Phases of Sarcasm Detection

- Sarcasm Classification Approaches: Sarcasm analysis can be implemented using
  - Machine Learning Approach
  - Lexicon Based Approach
  - Hybrid Approach

i. Machine Learning: It is a field of artificial intelligence that trains the model from the current data in order to predict future outcomes, trends and behaviors with the new test data. Machine Learning is categorized into Supervised and Unsupervised Learning [13]. In machine learning technique following steps are involved: Data collecting, Pre-processing, Training Data, Classification and Results.

   a. Supervised Learning: Supervised Learning is used when there is a finite set of classes (positive and negative). In this method, labeled data is needed to train classifiers. Supervised Learning process involves two steps: Training and Testing.

   b. Unsupervised Learning: This method is used when it is hard to find labeled training documents. It does not depend upon prior training for mine the data. In document level, SA are based on deciding the semantic orientation (SO) of particular phrase within the document. If the average semantic orientation of these phrases is above some predefined threshold, then the document is classified as positive, otherwise it is deemed negative.

ii. Lexicon Based: One of the unsupervised techniques of sentiment analysis is lexicon based technique. There has been a lot of work done based on lexicon. In this classification is performed by comparing the features of a given text in the document against sentiment lexicons. Three methods to construct sentiment lexicon are: Manual Method, Dictionary Based Method and Corpus Based Method.

iii. Hybrid Based Techniques: It involves combination of other approaches namely machine learning and lexical approaches.

Table 3.1: Different approaches used for sarcasm detection

Author's Name	Dataset	Approaches Used	Results or Accuracy	Feature Extraction
Raghavan V M .et.al [1]	Facebook	Hybrid approach	82%	POS Tagger
Bruno Ohana, Bredan Tierney [4]	Film reviews	SVM and Semantic oriented approach	69.3%	SentiWordNet
Dmitry Davidov, Oren Tsur and Ari Rappoport [5]	Twitter	KNN (k- nearest neighbor)	64%	Single word, n-gram, pattern & punctuation
Edwin Lunando, Ayu Purwarianti [7]	Twitter	Naïve Bayes, Maximum Entropy, Support Vector Machine	53.1% Naïve Bayes, 53.8% MaxEnt, 54.1 SVM	No. of interjection words, negativity information, unigram, question word
Elisabetta Fersini, Enza Messina, Federico Alberto Pozzi [9]	Twitter	Ensemble approach (Bayesian Model Averaging and Majority Voting)	94.1%	Pragmatic Particles, POS lexical Components
Mondher Bouazizi, Tomoaki Ohtsuki [6]	Twitter	SVN, KNN, Maximum Entropy, Random Forest	83.1% random Forest, 60% SVM, 77.4% Maximum Entropy, 81.5% KNN	Sentiment related, Punctuation related, Syntactic & Semantic related, Patter related
Anandkumar D. Dave, Prof. Nitika P. Desai [10]	Amazon Product Reviews and Tweets	Supervised Classification	68%	Simple Bag-of-words, TF-IDF
Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya [11]	Twitter, Discussion Forum	Lexicon Based	88.7%	Lexical
Shubhodip Saha.et.al[12]	Twitter	Naïve Bayes, SVM	65.2% Naïve Bayes, 60% SVM	Textblob

Table 3.2: Accuracy calculated by different authors.

Author's Name	Accuracy				
	Outstanding (>90 & <100)	Excellent (>80 & <90)	Very Good (>70 & <80)	Good (>60 & <70)	Average (>50 & <60)
Raghavan V M .et.al [1]	-	Excellent	-	-	-
Bruno Ohana, Bredan Tierney [4]	-	-	-	Good	-
Dmitry Davidov, Oren Tsur and Ari Rappoport [5]	-	-	-	Good	-
Edwin Lunando, Ayu Purwarianti [7]	-	-	-	-	Average
Elisabetta Fersini, Enza Messina, Federico Alberto Pozzi [9]	Outstanding	-	-	-	-
Mondher Bouazizi, Tomoaki Ohtsuki [6]	-	Excellent	-	-	-
Anandkumar D. Dave, Prof. Nitika P. Desai [10]	-	-	-	Good	-
Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya [11]	-	Excellent	-	-	-
Shubhodip Saha.et.al[12]	-	-	-	Good	-

In table 3.2, Elisabetta Fersini .et.al [9], achieved more accuracy by using ensemble approach on twitter data. Similarly, supervised machine learning classifiers also achieved more than 80% accuracy with the help of different feature extraction like: sentiment related, punctuation related, pattern related and syntactic and semantic related and Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya [11] also reached to 88.7% accuracy with the help of lexicon based approach. On the other hand, low accuracy found by Edwin Lunando, Ayu Purwarianti [7], by using No. of interjection words, negativity information, unigram, question word features which proves that these features were not detect the sarcasm properly.

IV. CONCLUSION

The sarcasm related features are studied in the literature, which are tested upon the variable length data collected from the Facebook, movie reviews, Twitter, Amazon and other discussion forums. Various classification algorithms are deployed for various text analytics systems, which are shortlisted on the basis of the feature engineering mechanism and type of data. For the data collected from Twitter, the Random Forest, SVM and KNN are used with the punctuation-related, syntax-based and other

features for the sarcasm detection. The Random forest classifier is found the best in comparison with other classification models, where it outperforms the other model by minimum margin of 1.6% from KNN.

REFERENCES

[1] Raghavan V M , Mohana Kumar P , Sundara Raman Rand Rajeswari Sridhar, "Emotion And Sarcasm Identification Of Posts From Facebook Data Using A Hybrid Approach", Ictact Journal On Soft Computing, 2017.

[2] D. Chaffey, Global Social Media Research Summary 2016. URL ⟨http://www.smartinsights.com/Socialmedia-marketing/social-media-strategy/new-globalsocial-media-research/⟩.

[3] W.Tan, M.B.Blake, I.saleh, S.Dustdar, Social-networksourced big data analytics, InternetComput.17 (5) (2013)62–69.

[4] Bruno Ohana and Brendan Tierney, "Sentiment Classification of Reviews using SentiWordNet", 2009.

[5] Davidov Dmitry, Oren Tsur and Ari Rappoport, "Enchanced Sentiment Learning Using Twitter Hashtags and Smileys", Proceedings of the 23rd International Conference on Computational Linguistics: posters, pp. 241-249, 2010.

[6] Mondher Bouazizi and Tomoaki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter", IEEE, pp. 5477 – 5488, 2016.

[7] Edwin Lunando and Ayu Purwarianti, "Indonesian Social Media Sentiment Analysis with Sarcasm Detection", Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, pp. 195-198, 2013.

[8] Soujanya Poria, Erik Cambria, Devamanyu Hazarika and Prateek Vij, "A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks", 2016.

[9] Elisabetta Fersini, Federico Alberto Pozzi and Enza Messina, "Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers", Proceedings of the IEEE Conference on Data Science and Advanced Analytics, IEEE, pp. 1-8, 2015.

[10] Anandkumar D. Dave and Prof. Nikita P. Desai, "A Comprehensive Study of Classification Techniques for Sarcasm Detection on Textual Data", Proceedings of the International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), pp. 19851991, 2016.

[11] Aditya Joshi1,2,3Vinita Sharma, Pushpak Bhattacharyya, "Harnessing Context Incongruity for Sarcasm Detection" Research Gate.

[12] Shubhodip Saha, Jainath Yadav and Prabhat Ranjan, "Proposed Approach for Sarcasm Detection in Twitter", Indian Journal of Science and Technology, 2017.