iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** X **Month of publication:** October 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Review Paper on Twitter Sentiment Analysis Techniques

Jatinder Kaur[1]
[1]M. Tech Student, CSE Department, Guru Nanak Dev Engineering College, Ludhiana

Abstract— Sentiment Analysis is growing exponentially due to the importance of the automation in mining, extracting and processing information in order to determine the general opinion of a person Hence the conventional sentiment analysis approaches fails to efficiently handle the vast amount of sentiment data available now a days.. Twitter sentiment analysis is difficult to compare with the general sentiment analysis because of the lots of slang words and incorrect spellings. The extreme limit of characters that are permitted in Twitter is 140. Lexicon based and Machine learning are the two methods used for analysis the sentiments from the content. In this paper we attempt to improve the accuracy using both techniques and Experimental results show that the proposed method dramatically improves the accuracy and outperforms the state-of-the-art baselines.

Keywords— Twitter Sentiment Analysis, Machine Learning, Symbolic Technique, Tweets

I. INTRODUCTION

Twitter has become very popular and has grown rapidly. An increasing number of people are willing to post their opinions on Twitter, as per the current report 313 million monthly active users per day and 500 million tweets per day that is considered as a valuable online source for opinions. But the main challenging task is extracting and analyzing the useful things from Twitter. The unstructured nature of the content and the natural language used to write these content added up the complexity more and it opened a new area of research called Opinion Mining and Sentiment Analysis. Sentiment Analysis is generally carried out in three steps. First, the subject towards which the sentiment is directed is found then, the polarity of the sentiment is calculated and finally the degree of the polarity is assigned with the help of a sentiment score which denotes the intensity of the sentiment. Sentiments can be classified at various levels: Aspects or feature level, sentence level and document level. Aspects or feature level sentiment classification classifies the sentiments based on the sentiments polarity of each aspects or feature about some target object and sentence level sentiment classification on the other hand classifies each sentence based on their sentiment polarity towards some topic. In document level sentiment classification the polarity of whole document is determined. It classifies the entire document into positive or negative or neutral class. Generally, two techniques are used for opinion mining and sentiment analysis: 1) Machine learning based techniques 2) Lexicon based techniques. In machine learning based techniques various machine learning algorithms are used for sentiment classification. Both supervised and unsupervised learning algorithm can be used to classify text. In Lexicon based techniques, a sentiment dictionary with sentiment words are used for sentiment classification. The dictionary contains polarity of each word whether they are positive, negative and objective words. Polarity of the opinion words can be determined by matching those words with dictionary words.

The aim for this paper is to propose a technique to achieve high sentiment analysis accuracy for tweets. In order to do this, we are proposing to combine two different approaches: the Lexicon-Based approach and the Machine Learning approach. We are going to compare the methods from each approach and then we will integrate them in an effort to achieve our goal.

II. RELATED WORK

In the past years, many works has been released in sentiment analysis. Implementation of sentiment analysis has been carried out for a variety of applications over a wide range of classification algorithms and for varying data size. There exist many possible variants; some of them are discussed in following section.

A. Machine Learning

M. S. Neethu and R. Rajasree,[1] Sentiment analysis deals with identifying and classifying opinions and grouping feelings or sentiment shown in source content. Online networking is creating an unlimited measure of notion rich information as tweets, announcements, blog entries and so on. Conclusion examination of this client produced information is exceptionally valuable in knowing the assessment of the group. Twitter notion examination is troublesome contrasted with general opinion investigation

because of the vicinity of slang words and incorrect spellings. The most extreme breaking point of characters that are permitted in Twitter is 140. Information base approach and Machine learning methodology are the two procedures utilized for investigating assessments from the content. In this paper, we attempt to investigate the twitter posts about electronic items such as mobiles, portable PCs and so on utilizing Machine Learning approach. By doing sentiment Analysis in a particular space, it is conceivable to distinguish the impact of area data in opinion arrangement. We show another component vector for grouping the tweets as constructive, contrary and concentrate people groups' sentiment about items.

G. Gautam and D. Yadav,[2] The broad of World Wide Web has brought another method for expressing the opinions of people. It is a medium with an enormous data where clients can see the views of different clients that are classified into various sentiment classes and are growing as a key factor in choice making. This paper adds to the sentiment analysis for clients' review classification which is useful to analyze the data in the form of tweets where conclusions are unstructured and are either positive or negative. For this we first pre-prepared the dataset, after that separated the modifier from the dataset that make them feature vector, then choose the component vector list and applied machine learning based classification calculations i.e., : Naive Bayes, Maximum entropy and SVM alongside the Semantic Orientation based Word Net which extracts equivalent words and similitude for the content feature.

Akaichi, Z. Dhouioui and M. J. Lopez-Huertas Perez, [3]In recent years, text mining and sentiment analysis have received great consideration because of the abundance of opinion information that exist in social networking sites for example, Facebook, Twitter, and so on. Sentiments are projected on these media usingmessages for communicating emotions, for example, friendship, anger, happy, satisfaction, and so on. Existing sentiment analysis has a tendency to identify user behavior and stateof mind but still unsatisfied because of complexities in conveyed texts. In this paper, we concentrate on the use of data mining for sentiment classification .Our point is to extract usefuldata, about clients opinion and behavior during this significant period. For that reason, we propose a strategy i.e., Support Vector Machine (SVM) and Naïve Bayes. We also construct sentiment lexicon based on the emoticons. In addition, we perform some similar analysis between two machine learning calculations SVM and Naïve Bayes through a training model for sentiment classification.

N. Azam, Jahiruddin, M. Abulaish and N. A. H. Haldar, [4]The popularity of social sites like Twitter, which encourages users to exchange short messages, Twitter is being used by public for upgrade and feeling expression. Since tweets don't follow grammatical structure, parsing methods don't work properly because of incorrect parts of speech used to individual words. In this paper, we have proposed a n-gram based statistical approach which identify significant terms and use them for vector-space modeling of the tweets. After that, a social graph method is proposed, considering tweets as nodes and the level of similarity between a couple of tweets as a weighted edge between them. The Experiment used 3100 tweets identified with Delhi assembly election and union budget 2015. The results are empowering, demonstrating the adequacy of the proposed social graph and event classification technique.

M. Kanakaraj and R. M. R. Guddeti, [5]Mining conclusions and analyzing sentiments from social networking information help in different fields, for example, even prediction, analyse people state of mind on a specific social issue. This paper analyse the views of public on a specific news from Twitter posts. The key thought of the paper is to expand the accuracy of classification by including Natural Language Processing Techniques (NLP) particularly semantic. The mined content data is subjected to Ensemble classification to analyze the sentiment data. Tests demonstrated that ensemble classifier outperforms traditional machine learning classifiers by 3-5%"Performance analysis of Ensemble strategies on Twitter sentiment analysis using NLP Methods.

B. Lexicon Based

Turney(2002)[6] used unsupervised machine learning approach to classify the review dataset. The algorithm classified the review into recommended review and not recommended review. Point-wise mutual information (PMI) of the words are used to determine the polarity. Adjectives and adverbs are considered for feature space construction. An accuracy of between 66 - 74% is achieved while conducting experiment on datasets from different domain such as movie, bank and automobile.

Harb et al. [7] considered two set of seed words having positive and negative polarities and by using association rule, more seeds are collected from Google Search API. The sum of polarities of the sentiment words classified the document. For positive categorization of documents they yielded 71% while for negative categorization of documents they achieved only 62%.

A. Khan et al. [8] conducted a sentence level sentiment classification using rule based domain independent approach. Sentences are categorized first into subjective and objective sentences then sentiment score is calculated using SentiWordNet. By considering the sentence structure the final sentiment score is calculated. They achieved an accuracy of 86.6% at the sentence level.

Zhang et al. [9] used an aspects based sentiment analysis to develop a system that finds weakness of the product, so that the producers can improve quality of the product. For every aspect the system tries to find implicit and explicit features and then the

related sentiment words. For getting accurate sentiment words they used sentence based sentiment analysis. The system showed 85.26% recall, 82.62% precision and about 83.92% F1-measure.

Kamps [10] used a simple technique based on lexical relations to perform classification of text. Andrea [11] used word net to classify the text using an assumption that words with similar polarity have similar orientation.

Ting-Chun [12] used an algorithm based on pos (part of speech) patter. A text phrase was used as a query for a search engine and the results were used to classify the text.

Prabhu [13] which used a simple lexicon based technique to extract sentiments from twitter data Turney [14] used semantic orientation on user reviews to identify the underlying sentiments.

Taboada [15] used lexicon based approach to extract sentiments from microblogs.Sentiment analysis for microblogs is more challenging because of problems like use of short length status message, informal words, word shortening, spelling variation and emoticons. Twitter data was used for sentiment analysis by.Negation word can reverse the polarity of any sentence. Taboada [15] performed sentiment analysis while handling negation and intensifying words. Role of negation was surveyed by [15].

Minquing [16] classified the text using a simple lexicon based approach with feature detection.It was observed that most of these existing techniques doesn't scale to big data sets efficiently. While various machine learning methodologies exhibits better accuracy than lexicon based techniques, they take more time in training the algorithm and hence are not suitable for big data sets. In this paper, lexicon based approach is used to classify the text according to polarity.

C. Hybrid Approach

Some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance.

Fang et al. [16] adopted entirely different approach. They considered both general purpose lexicon and domain specific lexicon for determining polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well. The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step1.Their system yielded around 66.8% accuracy.

Mudinas et al. [13] combined lexicon based and learning-based approaches to develop a concept-level sentiment analysis system, pSenti. It utilized advantages of both the approaches and attained stability and readability from semantic lexicon and high accuracy from a powerful supervised learning algorithm. They extracted sentiment words and considered it as features in machine learning algorithm. This hybrid approach pSenti achieved an accuracy of 82.30%.

Zhang et al. [16] carried out entity level sentiment analysis. They utilized both the supervised learning techniques and lexicon based techniques. By lexicon based method they extracted sentiment words. By using Chi-square test on the extracted seeds additional seeds are discovered. Sentiment polarities of newly discovered seed are determined through a classifier, which is being already, trained using initial seeds. There is no manual task in this proposed system and it achieved around 85.4% of accuracy.

III. SENTIMENT ANALYSIS TECHNIQUES

A. Maching Learning Technique(Supervised)

A dataset is created using twitter posts of electronic products.Tweets are short messages with full of slang words and misspellings. So we perform a sentence level sentiment analysis. This is done in three phases. In first phase preprocessing is done. Then a feature vector is created using relevant features. Finally using different classifiers, tweets are classified into positive and negative classes. Based on the number of tweets in each class, the final sentiment is derived.

1) *Creation of a Dataset:* Since standard twitter dataset is not available for electronic products domain, we created a new dataset by collecting tweets over a period of time. Tweets are collected automatically using Twitter API and they are manually annotated as positive or negative. A dataset is created by taking positive tweets and negative tweets.
2) *Pre-processing of Tweets:* Keyword extraction is difficult in twitter due to misspellings and slang words. So to avoid this, a preprocessing step is erformed before feature extraction. Preprocessing steps include removing url, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with two occurrences. Slang words contribute much to the emotion of a tweet. So they can't be simply removed. Therefore a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings. Domain information contributes much to the formation of slang word

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

dictionary.

3) *Creation of Feature Vector:* Feature extraction is done in two steps. In the first step, Twitter-specific features are extracted. Hashtags and emoticons are the relevant Twitter-specific features. Emoticons can be positive or negative. Therefore, they are assigned different weights. Positive emoticons are assigned a weight of '1' and negative emoticons are assigned a weight of '-1'. There may be positive and negative hashtags. Therefore, the count of positive hashtags and negative hashtags are added as two separate features in the feature vector. Twitter-specific features may not be present in all tweets. So, a further feature extraction is to be done to obtain other features. After extracting Twitter-specific features, they are removed from the tweets. A tweet can be then treated as simple text. Thereafter, using unigram approach, tweets are represented as a collection of words. In unigrams, a tweet is represented by its keywords. We maintain a negative keyword list, positive keyword list and a list of different words that represent negation. Counts of positive and negative keywords in tweets are used as two different features in the feature vector. Presence of negation contribute much to the sentiment. So their presence is also added as a relevant feature.

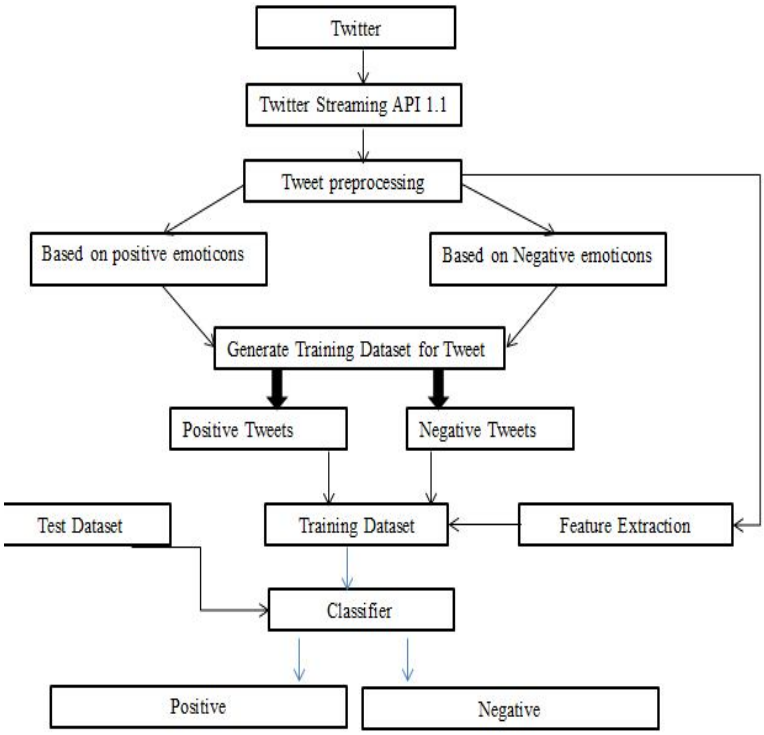


Fig 1. Sentiment Classification based on Emoticons

All keywords cannot be treated equally in the presence of multiple positive and negative keywords. Therefore, a special keyword is selected from all the tweets. In the case of tweets having only positive keywords or only negative keywords, a search is done to identify a keyword having relevant part of speech. A relevant part of speech is an adjective, an adverb or a verb. Such a relevant part of speech is defined, based on their relevance in determining sentiment. A keyword that is an adjective, adverb or a verb shows more emotion than others. If a relevant part of speech can be determined for a keyword, then it is taken as special keyword. Else, a keyword is selected randomly from the available keywords as special keyword. If both positive and negative keywords are present in a tweet, we select any keyword having relevant part of speech. If relevant part of speech is present for both positive and negative keywords, none of them is chosen. Special keyword feature is given a weight of '1' if it is positive and '-1' if it is negative and '0' in its absence. Part of speech feature is given a value of '1' if it is relevant and '0' otherwise. Thus, feature vector is composed of 8 relevant features. The 8 features used are part of speech (pos) tag,special keyword, presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hashtags and number of negative hashtags.

4) *Sentiment analysis:* After creating the feature vector,Classification procedure is begin. Naïve Bayes and Maximum entropy are used to classify the text. First data is train through the two classifier and then testing has to be done. Actual sentiments are determine through the best classifier in the form of negative, positive and neutral.

In this experiment we tested the impact of training set size on classifier performance. We took a random sample of 10% from the dataset and used it as a test set. Then we trained different classifiers on different sized portions of the remaining 90%. In the experiment, we compare the Naive Bayes (NB) and Maximum Entropy classifier, each trained with the full feature set.

a) *Naïve Bayes Classifier:* The Naïve Bayes (NB) classifier is based on Bayes rule, a practical Bayesian learning model that is easy to understand and implement. The Bayes rule allows us to determine this probability of any event. It is the probabilistic approach to the text classification and can learn the pattern of examining a set of documents that has been categorized . It compares the contents with the list of words to classify the documents to their right category or class. Let d be the tweet and c* be a class that is assigned to d, where

$$C^* = \arg mac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) = \frac{(P(c))\sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

From the above equation,'f' is a "feature", count of feature (f_i) is denoted with $n_i(d)$ and is present in d which represents a tweet.Here, m denotes no. of features.Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify using Naïve Bayes Machine Learning technique ,we can use the Python NLTK library .

b) *Maximum Entropy:* Maximum Entropy (MaxEnt) is a multinomial logistic regression model that allows for classification with more than two discrete classes. In Maximum Entropy Classifier, no assumptions are taken regarding the relationship in between the features extracted from dataset. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label.

Maximum entropy even handles overlap feature and is same as logistic regression method which finds the distribution over classes. The conditional distribution is defined as MaxEnt makes no independence assumptions for its features, unlike Naive Bayes.
The model is represented by the following:

$$P_{ME}(c \mid d, \lambda) = \frac{\exp\left[\sum_i \lambda_i f_i (c,d)\right]}{\sum_c \exp\left[\sum_i \lambda_i f_i (c,d)\right]}$$

Where c is the class,d is the tweet and λ_i is the weight vector.The weight vectors decide the importance of a feature in classification.

c) *Support Vector Machine:*Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data which represented as a vector is classified into a class. Nextly we find a margin between the two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

IV. NATURAL LANGUAGE PROCESSING/SYMBOLIC TECHNIQUE (UNSUPERVISED)

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. In Symbolic techniques that focuses on the force and direction of individual words (the so-called "bag of words" approach). In that approach, relationships between the individual words are not considered and a document is represented as a mere collection of words. We can divide the symbolic technique into two subparts:

A. Dictionary Based Technique

The idea of dictionary-based approach is first to collect some small set of seed words with their known opinion polarities and then to iteratively extend them with the help of online dictionary e.g. WordNet.WordNet database consists of words connected by synonym

relations. WordNet is the popular lexical resource used to determine the overall sentiment, sentiments of every word is determined and those values are combined with some aggregation functions.

1) Drawback : Can't deal with domain and context specific orientations.

B. Corpus Based Technique

This approach uses as its bases the syntactic or co-occurrence patterns and the predefined set of seed words with its polarities. SentiWordNet is a widely-used English sentiment lexical resource that was generated by automatically annotating the synsets of the WordNet, where each synset received the three scores indicating to which extend the respective synset is to be regarded positive, negative or objective. The sum of all three scores always equals one.The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques.

Methods based on statistics: Latent Semantic Analysis (LSA).

Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus.

V. HYBRID APPROACH

Hybrid Technique combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance.first approaches was used to classify each tweet through both classifiers and then

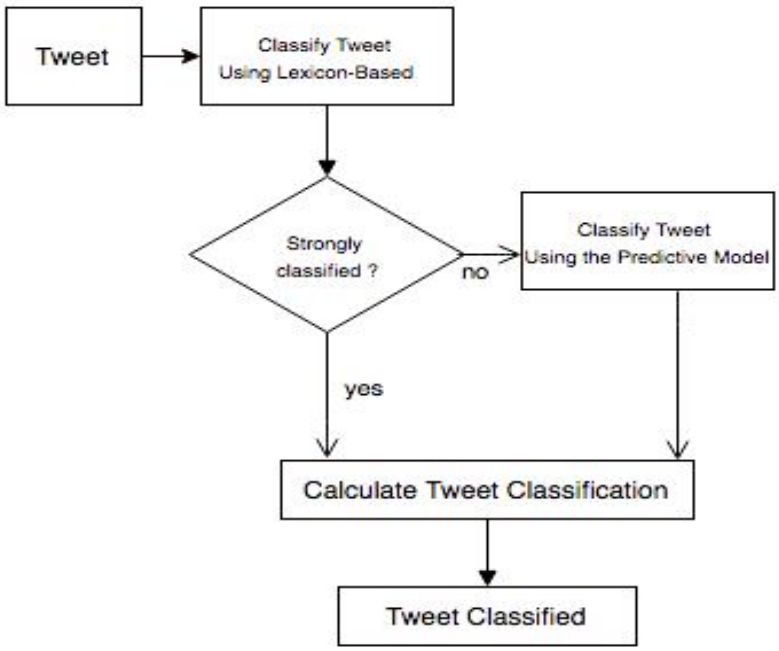


Fig2 : SA Methods-order is predetermined

calculate the average polarity, but after measuring the accuracy and speed, the results were not favorable.

After more testing, we found a pattern: a larger amount of the tweets that were incorrectly classified by both methods, had a polarity between 0.4 and 0.6 where 0 is negative and 1 is positive. In other words, only a small portion of the incorrectly classified tweets were classified with a value greater than 0.6 and less than 0.4, tweets that we called "strongly classified". Based on this, we saw an opportunity to limit the scope of tweets to be classified twice to those tweets with a polarity classification between 0.4 and 0.6. This was really useful because we could reduce the overhead of the process by a considerable amount.

In Figure2,where the tweet would be first analyzed by the Lexicon-Based technique and if it was not "strongly classified" then it would be analyzed by the Naïve-Bayes classifer, and an average of both polarities would determine the tweet classification.Continuous tests with similar and previously unseen tweets helped us to determine that even when the Lexicon-Based classification was more reliable and consistent with completely unseen data and across different domains, the Naïve Bayes Classifier would perform better for tweets similar to the ones used to train its predictive model.

Volume 4 Issue X, October 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

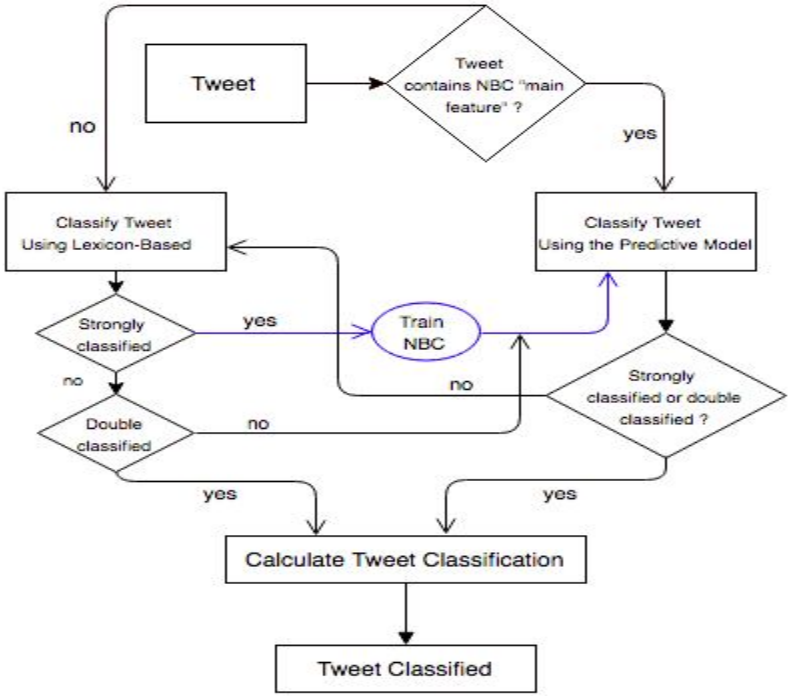


Fig3: SA Methods - order changes dynamically depending on the tweet's content, and Lexicon-Based trains NBC.

In our tests, using the "strong classified" tweets from the Lexicon-based to train the Naïve Bayes Classifier made the system perform slightly better. This is because even though there are tweets training the NBC incorrectly, the number of tweets training it correctly is still enough to slighty increase the overall accuracy. Depending on how performs the lexicon-based classifier, using the tweets that are classified as "strong classified" to train the Naïve Bayes Classifier could slightly outperform the previous implementation. The hybrid approach combines the advantages of both the techniques. It is inheriting high accuracy from supervised machine learning algorithm and achieving stability for lexicon based approach.This approach fulfilled the hypothesis of combining different methods to achieve a higher accuracy, however, there are a lot of situations with room for improvement, such as increasing the accuracy for those questions that contain opinion words, but do not express any sentiment, or those sarcastic sentences, it is difficult to deal with opinion words that implies the inverse of its usual meaning.

VI. COMPARATIVE ANALYSIS

Paper	Approach	Dataset	Technique	Accuracy
Turney[6]	Unsupervised	Movie, bank and automobile	PMI	66%
Pang et al.[18]	Supervised	Movie Review	SVM	82.9%
			Naïve Bayes	81.5%
			Maximum Entropy	81.0%
Hu and Liu[17]	Unsupervised	Customer reviews	Lexicon	84%
Abbasi et al.[20]	Supervised	Movie review	SVM	95.5%
Harb et al[7]	Unsupervised	Movie Review	Lexicon	71%
a. Khan et al.[8]	Unsupervised	Customer reviews	Lexicon	86.6%
Zhang et al.[9]	Unsupervised	Product reviews	Lexicon	82.62%
Zhang et al.[10]	Hybrid	Twitter tweets	ML and Lexicon	85.4%
Mudinas et al.[20]	Hybrid	Customer reviews	ML and Lexicon	82.3%
Fang et al.[21]	Hybrid	Multi Domain	ML and Lexicon	66.8%

TABLE 1: ACCURACY OF SENTIMENT ANALYSIS USING DIFFERENT TECHNIQUES.

VII. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

A. *Identifying subjective parts of text*

Subjective parts represent sentiment-bearing content. The same word can be treated as subjective in one case, or an objective in some other. This makes it difficult to identify the subjective portions of text.

For example:

The language of the Mr.Dennis was very crude.

Crude oil is obtained by extraction from the sea beds.

The word „crude‟ is used as an opinion in first example, while it is completely objective inthe second example.

B. *Domain dependence*

The same sentence or phrase can have different meanings in different domains. For Example, the word „unpredictable‟ is positive in the domain of movies, dramas ,etc, but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.

C. *Sarcasm Detection*

Sarcastic sentences express negative opinion about a target using positive words in unique way..

“Nice perfume. You must shower in it.”

The sentence contains only positive words but actually it expresses a negative sentiment.

D. *Thwarted expressions*

There are some sentences in which only some part of text determines the overall polarity of the document.

Example: “This Movie should be amazing. It sounds like a great plot, the popular actors , and the supporting cast is talented as well. “

In this case,a simple bag-of-words approaches will term it as positive sentiment, but the ultimate sentiment is negative.

E. *Explicit Negation of sentiment*

Sentiment can be negated in many ways as opposed to using simple no, not, never, etc. It is difficult to identify such negations .

Example: “It avoids all suspense and predictability found in Hollywood movies.”

Here the words suspense and predictable bear a negative sentiment, the usage of „avoids‟ negates their respective sentiments.

F. *Order dependence*

Discourse Structure analysis is essential for Sentiment Analysis/Opinion Mining.

Example: A is better than B, conveys the exact opposite opinion from, B is better than A.

G. *Entity Recognition*

There is a need to separate out the text about a specific entity and then analyze sentiment towards it.

Example:“I hate Microsoft, but I like Linux”.

A simple bag-of-words approach will label it as neutral, however, it carries a specific sentiment for both the entities present in the statement.

H. *Building a classifier for subjective vs. objective tweets*

Current research work focuses mostly on classifying positive vs. negative correctly. There is need to look at classifying tweets with sentiment vs. no sentiment closely.

I. *Handling comparisons*

Bag of words model doesn't handle comparisons very well.

Example:"IIT‟s are better than most of the private colleges", the tweet would be considered positive for both IIT‟s and private colleges using bag of words model because it doesn't take into account the relation towards "better".

J. Applying sentiment analysis to Facebook messages

There has been less work on sentiment analysis on Facebook data mainly due to various restrictions by Facebook graph api and security policies in accessing data.

K. Internationalization

Current Research work focus mainly on English content, but Twitter has many varied users from across.

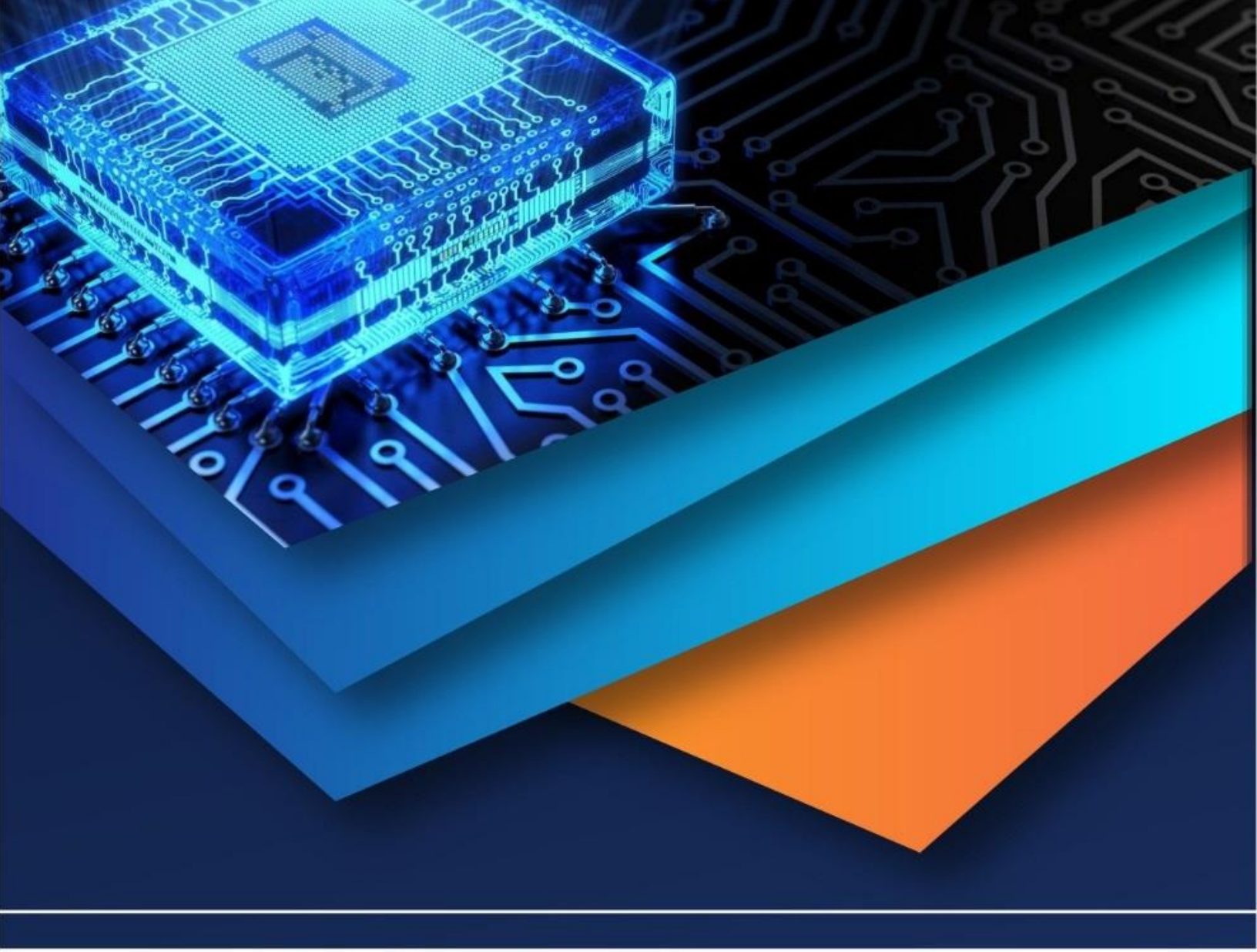
VIII. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document .We also studied the effects of various features on classifier. We also studied about hybrid technique to combine two different sentiment analysis methods that yielded an accurate sentiment analysis classification in a real-time environment using Twitter as the source of content. Such a method woud be very useful for systems that need to classify tweets in real-time mainly from a specific domain, but not limited to that. Depending on the tweet's content, would be able to delegate, to one of the methods, the responsibility of classifying and would use the second method for double validation and in some cases for training.Hence we can conclude that more the cleaner data, more accurate results can be obtained.We discussed many challenges for sentiment analysis that can visualize as future research for this study.

REFERENCES

[1] R.Rajshree and M S. Neethu," machine learning techniques used in sentiment analysis in twitter," (ICCCNT)Computing, Communications and Networking Technologies, Fourth International Conference 2013 on, Tiruchengode, 2013, pp. 1-5.
[2] G. Gautam and D. Yadav, ."Sentiment analysis in twitter using semantic analysis and machine learningapproaches," Contemporary Computing (IC3), 2014 Seventh International Conference on, Noida, 2014, pp. 437-442.
[3] J. Akaichi, Z. Dhouioui and M. J. Lopez-Huertas Perez, "For sentiment classification text mining face book updates on," (ICSTCC) System Theory, Control and Computing, 2013 17th International Conference, Sinaia, 2013, pp. 640-645.
[4] N. Azam, Jahiruddin, M. Abulaish and N. A. H. Haldar ,"Twitter Data Mining of Events Analysis and Classification," 2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI), Hong Kong, 2015, pp. 79-83.
[5] M. Kanakaraj and R. M. R. Guddeti," Semantic Computing (ICSC), 2015 IEEE International Conference on, Anaheim, CA, 2015, pp. 169-170.
[6] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424.
[7] A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, Cergy-Pontoise, France, 2008.
[8] A. Khan, B. Baharudin, K. Khan; "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp.317-331, 2011.
[9] W. Zhang, H. Xu, W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," Expert Systems with Applications, Elsevier, vol. 39, 2012,pp. 10283-10291.
[10] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011.
[11] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.
[12] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.
[13] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL OF COMPUTING, VOLUME 2, ISSUE 8, AUGUST 2010, ISSN 2151-9617 .
[14] Prabu Palanisamy, Vineet Yadav, Harsha Elchuri, "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis", Serendio Software Pvt Ltd, 2013.
[15] Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4):315–346.
[16] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexiconbased methods for sentiment analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press.
[17] Minqing Hu, Bing Liu. Mining and Summarizing Customer Reviews, Department of Computer Science, University of Illinois at Chicago, Research Track Paper.
[18] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
[19] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," In ACM Transactions on Information Systems, vol. 26 Issue 3, pp. 1-34, 2008.

[20] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for conceptlevel sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York,NY, USA, Article 5, pp. 1-8, 2012.
[21] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classication", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY