VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 1 (2018)
Article No. 8
PP 1-8

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Methods for Sentiment Analysis: A Literature Study

Shiv Dhar[1], Suyog Pednekar[1], Kishan Borad[1], Ashwini Save[2]
[1](B.E. Computer Engineering, VIVA Institute of Technology, University of Mumbai, India)
[2](Head of Department, Computer Engineering, VIVA Institute of Technology, University of Mumbai, India)

Abstract : *Sentiment analysis is a trending topic, as everyone has an opinion on everything. The systematic study of these opinions can lead to information which can prove to be valuable for many companies and industries in future. A huge number of users are online, and they share their opinions and comments regularly, this information can be mined and used efficiently. Various companies can review their own product using sentiment analysis and make the necessary changes in future. The data is huge and thus it requires efficient processing to collect this data and analyze it to produce required result.*
In this paper, we will discuss the various methods used for sentiment analysis. It also covers various techniques used for sentiment analysis such as lexicon based approach, SVM [10], Convolution neural network, morphological sentence pattern model [1] and IML algorithm. This paper shows studies on various data sets such as Twitter API, Weibo, movie review, IMDb, Chinese micro-blog database [9] and more. The paper shows various accuracy results obtained by all the systems.
Keywords– *Machine Learning, Sentiment Analysis, CNN, analysis, AI, SVM, NLP.*

1. INTRODUCTION

Sentiment analysis intents to define the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual division or emotional response to a document, interaction, or event. It refers to the use of natural language processing, text analysis, computational semantics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is broadly applied to "voice of the customer" materials such as reviews and survey responses, as well as to online and social media. Sentiment analysis has claims in a variety of domains, ranging from marketing to customer service to clinical medicine.

Sentiment analysis stands at the intersection of natural language processing and large-scale data mining. Sentiment analysis has important applications in academia as well as commerce. The understanding of human language is a core problem in AI research. At the same time, with increasingly lowering barriers to the Internet, it is easier than ever for end-users to provide feedback on the products and services they use. This information is highly valuable to commercial organizations; however, the volume of such reviews is growing rapidly, necessitating an automated approach to extracting meaning from the huge volume of data. This automated approach is provided by sentiment analysis.

In this paper, various approaches to sentiment analysis have been examined and analysed. Techniques such as lexicon based approach, SVM [4], Convolution neural network [9], morphological sentence pattern model and IML algorithm are discussed. These techniques all have different strengths and weaknesses, and are have different use cases. Their advantages and disadvantages are explored and compared.

2. SENTIMENT ANALYSIS TECHNIQUES

2.1 Sentiment Analysis on Twitter using Streaming API [8]

The propose system focuses on analyzing what people thinks about various products and services. Many users share their opinions about various products and contents. Sentiment analysis helps in classifying the positive or negative data. In the proposed system, it uses Natural language processing - NLTK, where it helps in

tokenization, stemming, classification, tagging, parsing and sentiment reasoning. Its basic feature is to convert unstructured data into structured data.

It uses Naive Bayes for classification which requires number of linear parameters. The system uses Hadoop for extracting information and uses Twitter Application Programming Interface. The system gathers streaming tweets using Twitter API and assigns every tweet a positive or negative probability.

The future system mainly focuses on real time sentiment analysis such as evaluating tweets from twitter. It performs sentiment analysis, feature based classification and opinion summarization.

Advantages:

Classification is done in real time which makes it very efficient.

Sentiment analysis in this system uses Hadoop to load live data.

Several systems use stored tweets for classification, leading to high requirement of space, whereas here the storage required is less.

Disadvantages:

While classifying, the words are accepted individually rather than in a fix pattern or complete sentence.

The semantic meaning is neglected as single words are scanned.

2.2 Neural Networks for sentiment analysis on twitter [7]

The proposed system mainly focuses on providing polar views by dividing the opinions in social media and websites having customer reviews. It divides the reviews from websites and divide them into positive, negative and neutral reviews. The system used feed forward neural network. The neural network used is on MATLAB, using neural network toolbox. It reduces the input by removing the punctuations, single characters, stop words like and, to, the etc. and also mentions to other users using @ symbol.

The system uses Porter's stemming algorithm for stemming. Each tweet obtained is assigned a value and arranged linearly in 2D table. It uses pattern matching in neural networks for checking the data.

The proposed system performs sentiment analysis on twitter using neural networks. Sentiment analysis is performed by various methods, here it uses neural networks which helps in achieving more accuracy and efficiency. Preprocessing is also implemented by the proposed system which helps in obtaining better results.

Advantages:

Tweets are easily classified into positives and negatives.

Preprocessing helps in improving the time required.

Reducing redundant data helps in gaining better accuracy.

Disadvantages:

The input is still comparatively large and thus require more time.

The input on twitter has #, which are connected having no space. This requires dividends and thus needs to be implemented in future.'

2.3 Product related information sentiment-content analysis based on Convolution Neural Networks for the Chinese micro-blog [9]

Sentiment analysis is performed by the suggested system on various Chinese micro-blogs. It performs sentiment analysis to determine whether positive/negative or it is an advertisement. The system uses convolution neural networks for classification. And support vector machine algorithm (SVM algorithm) is used. It reduces the size of input data by breaking down major data set containing all the information into smaller data set by removing unwanted data like author's name, duplicated data and similar texts. It uses CNN, which has four layers namely input layer, convolution layer, pooling layer and fully connected layer. SVM and lexicon analysis is used as baseline.

In the proposed system sentiment analysis of Chinese micro-blogs is performed using CNN. There are many product advertisements and promotions in micro-blogs which can be detected using this process. The system is quite useful in removing the redundant data such as advertising and promotions, resulting in better results for sentiment analysis.

Advantages:

It provides better result than earlier lexicon analysis.

Along with positive and negative statement, it also determines advertisement.

It also provides better result than SVM.

Disadvantages:

It takes an entire sentence or concatenated sentence as input. Thus, more time is required for analysis.

Here entire data results in more time required to analyze than just predefined patterns.

It uses sentence embedding viz. less effective than character embedding.

2.4 Convolution Neural Networks based sentiment analysis using Adaboost combination [6]

The proposed system focuses on feedback analysis. This paper states various techniques for sentiment analysis classification. It includes SVM, Naive Bayes, recursive neural network, auto encoders. It also states various methods for identifying the different roles of specific N-grams. It uses Adaboost to combine different classifiers.

The system proposes preprocessing using sentence matrix input, i.e. the input matrix in classified and used as input in a matrix. Here N-grams are formed which are used for dividing into smaller segments according to the N. It uses Adaboost algorithm to combine week classifiers with the strong classifiers. Parameter overfitting is checked and reduced using regularization. It drops certain parameters which are not defined earlier. The data set used in this system are Movie Review and IMDB.

Advantages:

It uses boosted CNN to provide better results than general CNN.
The proposed model separates the features by passing the documents and then boost the classifiers trained on these representations.

Disadvantages:

Although it uses N-gram approach, it still must cover the total input and thus time required is high.
There can be more layers added in CNN for better result.

2.5 A Feature based approach for sentiment analysis using SVM and Co-Reference Resolution [4]

Online shopping is trending these days as it's found secure. People buy products online and post their reviews on it. These are in the form of tweets or product reviews. It is difficult to manually read these reviews and assign sentiment to them. So, for all these tweets an automated system can be created which will analysis the review and extract the user percepts. In this paper they have developed a producer for feature based sentiment analysis by a classifier called Support Vector Machine.

In this paper they have used machine learning approach called supervised classification which is more accurate than all other methods as the classifier is to be trained using the real-world data set. They used SentiWordNet which is created mainly for opinion mining. As every word in SentiWordNet have 3 polarities as Positive, Negative and Subjective. SVM is used because sentiment analysis is a binary classification and it has capability to work on huge datasets. Co-Reference Resolution is to remove the repetitions of words in a sentence and to develop the relation between two sentences for the sentiment analysis.

Advantages:

Combination of SVM and Co-Reference Resolution improves the accuracy of feature based sentiment analysis.
SentiWordNet helps to find the Polarities of the opinion words.

Disadvantages:

Sarcastic reviews are different to identify for computer as well as human.
Some reviewers may spam the reviews which is different to identify.
Reviewers may post advertisement which is to be detected and discarded.

2.6 SentiReviews: Sentiment analysis based on Text and Emoticons [5]

Sentiment using emoticons is increasing gradually on social networking. People comment, tweets post their opinions using test as well as emoticons. Which increase the difficulties in analysis the sentiment of the reviewer. Various factors that affect sentiment analysis are discussed here but the focus is on the emoticons and the role of emoticons in sentiment analysis also various issues like sarcasm detection, multilingualism handling acronyms and slang language, lexical variation and dynamic dictionary handling are discussed. Users these days use emoticons to express most of their emoticons, text communication erase the uses of emoticons.

Sentiment analysis can be done based on two approaches, Lexicon based approach and Machine Learning approach. The Lexicon is the vocabulary of person, language or branch of knowledge used for calculating polarities of sentiment, in the Lexicon based approach. In Machine Learning approach, approach the machine/ computer learns the sentiment on regular bases and the polarities are assign.

Advantages:

Various methods are available to find the sentiment in a tweet or review.

Various approaches can be used to detect the sentiment from the feature based sentiment analysis.

Disadvantages:

Emoticons in the sentences, tweets or review is the problem to define the sentiment.
Dealing with the variation of lexicon can be challenging task in sentiment.

2.7 A feature based approach by using Support Vector Machine for sentiment analysis [10]

As the modern era of globalization, e-commerce is growing in vast number so as their opinion also, but it is very difficult to identifying whether it is positive, negative or neutral and it would be tiresome job to study all the opinion manually. To find out the sentiment an automated system must be developed "Support Vector Machine" can be used for this method. SVM is machine that takes the input and store them in a vector then using SentiWordNet it scores it decides the sentiment. It also classifies the opinion in overall way by positive, negative or Neutral.

Advantages:

The accuracy rate is increased as each word in the opinion are scored and the overall sentiment is given.
It can work on large data a single time.

Disadvantages:

Sarcasm detection can be a problem.
Anaphora Resolution is most user ignores the pronouns.
Grammatical mistakes of user.

2.8 Localized twitter opinion mining using sentiment analysis [11]

As the public information from social media can get interesting result and the public opinion on any product. Service or personally is most effective and it is necessary to find this information from social media. Sentiment analysis mining using Natural Language Processing, Rapid miner, SentiWord, SNLP as mining of all the opinion on social media has become a necessity for the analysis of sentiment from the user. Stanford NLP is used to extract the sentiment from an opinion, Rapid Miner is used to mine all the opinion, tweets from the social using a keyword, SentiWord is used to assign the polarities to the keywords in the opinion.

Advantages:

Mining of opinion using keyword of product is done faster.
Polarities assignment helps to analysis the opinion.

Disadvantages:

Emoticons used in tweets can be difficult to result the sentiment.
Co-reference Resolution in opinion must be serious issue

2.9 A Method for Extracting Lexicon for Sentiment Analysis based on Morphological Sentence Patterns [1]

Aspect-based sentiment analysis is higher-quality and more in-depth than, the probability-based model. But building the aspect-expression pairs is a challenge (manually is slow, obviously). An unsupervised approach to building aspect-expression pairs is proposed. The natural morphological (i.e. grammatical) patterns in sentences are exploited to build aspect-expression pairs. It uses POS tagging to find expression candidates for aspects. Thus, it builds aspect-expression pairs which are then analyzed for sentiment.

Advantages:

The biggest advantage is that aspect-based sentiment analysis can be done automatically, in an unsupervised manner.
This helps us scale this in-depth approach to large datasets and new data, without human intervention.
It does so while maintaining or increasing classification accuracy.

Disadvantages:

The aspect-based approach is an all-or-nothing approach.
That is, if an aspect-expression pair is found, then results are usually quite accurate.
But if no aspect-expression pair is found, then that review or tweet cannot be processed further, rendering it effectively into useless noise.

2.10 Sentiment Analysis Using Machine Learning and Lexicon-based Approaches for Student Feedback [2]

Evaluation of instructors and courses at the end of the semester is becoming a standard practice. Along with scale-based feedback, students also provide textual feedback about the courses; this feedback can be analyzed for sentiment.

The paper recommends a hybrid methodology to sentiment analysis using both machine learning and lexicon-based approaches. This hybrid methodology yields an enhanced result than the lexicon-based approach or machine-learning approaches alone.

System Used:

The process methodology is as follows:

1. Dataset Description:

The dataset was manually labelled as positive, negative, neutral.
Thus, it is a supervised dataset.

2. Preprocessing:

The Python NLTK package was used to perform preprocessing: punctuation, tokenization, case conversion, stop words.

3. Data Partition:

The training test ratio of the dataset was 70:30.
TF-IDF (Term Frequency - Inverse Document Frequency).
The words that occur frequently in the dataset but not in a 'neutral' corpus are assigned a higher weight.
N-gram Features, Unigram (1-word) and bigram)2-word) features were extracted.
Lexicon Features, the semantic orientation was determined using an existing sentiment dictionary.

4. Model Training:

The hybrid model for sentiment analysis was trained using unigrams, bigrams, TF-IDF and lexicon-based features.
To train the model, random forest and support vector machine (SVM) algorithms were used.
This paper yielded a marginally better result than purely lexicon-based approaches.
It outperforms many commercial implementations such as Microsoft's API, Alchemy, and Aylien.

2.11 A Context-based Regularization Method for Short-Text Sentiment Analysis [3]

The authors suggest a fusion model that combines word-similarity knowledge and word-sentiment knowledge. They use the contextual knowledge obtained from the data to improve classification accuracy.

System:

To compute the sentiment polarity of a word, TRSR (TextRank Sentiment Ratio) is used. Word-embedding is used to compute the similarity between words.
This contextual knowledge obtained is not statistical but on the semantic level.
These two regularizations are combined as a classification model, which converts it to an optimization problem which can be solved computationally.

The parameter obtained from training the model applies into the logistic regression, and we get the final classification model. The hybrid model that combines word-similarity and word-sentiment performs better than either of the approaches used individually.

2.12 Aspect-based Feature Extraction and Sentiment Classification using Incremental Machine Learning Algorithm for Review Datasets [12]

The paper offers an approach for sentiment analysis using a planned iterative decision tree. Customer reviews are collected and from them, sentiments and aspects are identified; this is called aspect-based feature extraction. The authors compare the performance of their proposed system with baseline machine learning algorithms like SVM and naive Bayes.

There are 3 stages in this system:

1. Data preprocessing.

Many preprocessing stages are used to remove irrelevant and noisy data.

2. Aspect-based sentiment analysis.

The aspects and expressions are identified. Sentiment analysis will be performed on these aspects.

3. Opinion summarization using iterative learning tree algorithm.

It uses an iterative practice to categorize the given inputs for the assessment of sentiment. It starts comparison from the root node and then compares it with every instance of data. Labels are assigned to the leaf nodes. Every node in the tree represents an aspect.

Advantages:

This approach performs better than other algorithms like naive Bayes and SVM.
An incremental decision tree is much faster and better than a linear decision tree due to reduced memory and limited buffer requirements.

Disadvantages:

Classification, while better, is nevertheless supervised, because he class labels need to be well-defined.
This means that it cannot be used for new, unstructured datasets.

3. ANALYSIS

Following table is a summary of studied research papers on Sentiment analysis and various techniques used.

Sr. No.	Title	Technique Used	Dataset	Accuracy
1	Sentiment analysis of student feedback using machine learning and lexicon based approaches [2]	It uses a hybrid model. It integrates TF/IDF + lexicon with the machine learning algorithms like Random Forest and SVM.	1230 comments from the institute's education portal.	Accuracy: 0.93 F-measure: 0.92 Improved by 0.02
2	A context-based regularization method for short-text sentiment analysis [3]	It uses a classification model that combines two regularizations, word similarity and word sentiment. Introduces new word-sentiment calculating.	Movie comments from Cornell Univ. 2016 US election comments crawled from Facebook. S.C.D. from Weibo.	Accuracy is improved by over 4.5% baseline
3	A feature based approach for sentiment analysis using SVM and coreference resolution [4]	SVM for classification from huge dataset. Coreference resolution to extract the relation from two sentences.	Reviews from ecommerce sites	Combining coreference and SVM, it improves the accuracy of feature-based sentiment analysis
4	SentiReview: Sentiment analysis based on text and emoticons [5]	Lexicon-based approach for assigning polarities. Machine learning approach for constantly analyzing the polarities. Comparison for between different methods to analyses sentiment polarities.	Twitter API Weibo	Stating various methods
5	Convolutional Neural Network based sentiment analysis using	Uses boosted CNN Model.	Movie Review	Accuracy is 89.4%

	adaboost combination [6]	Adaboost algorithm is used to regularize.	IMDB	Increased by 0.2%
6	Neural Network for Sentiment analysis on Twitter [7]	Sentiment analysis using feed forward neural network. Reducing input sequence by removing &, @.	Tweets from Michigan's sentiment analysis. Twitter API	Accuracy achieved is 74.15%.
7	Sentiment Analysis on Twitter using streaming API [8]	NLTK for tokenization and convert unstructured data to structural data. Uses Naive Bayes for Classification.	Twitter API	It performs analysis on real time data.
8	Product based data sentiment content analysis based on Convolution Neural Network for the Chinese micro-blog [9]	Sentiment analysis using Convolution Neural Network	Chinese Micro blog database	Better accuracy than lexicon analysis.
9	A feature based approach for sentiment analysis by using Support Vector Machine (SVM) [10]	Support Vector Machine for classification from huge data.	Reviews of product from e-commerce site amazon, eBay	Accuracy increased total of 88.13%
10	Localized based Opinion mining using Sentiment Analysis [11]	Rapidminer is used for mining which extract information using keywords. Sentiword is used for assigning polarities.	Twitter API	Various processes for extraction of data.
11	Aspect based feature extraction & sentiment classification of reviews data sets using Incremental Machine Learning algorithm [12]	Identifies the sentiment, aspect & performs data classification. It uses incremental decision tree for classification. Opinion summarization using SVM & Naive Bayes.	General Data	The result shows that SVM method is much better than Bayes.
12	Sentiment Analysis on Social Media using Morphological Sentences Pattern model [1]	Extracts aspects & expression using sentiment pattern analyzer based on MSP model.	Movie reviews from IMDb Rotten Tomatoes Twitter YouTube	Accuracy increased to 91% Increased by 2.2%

Volume 1, Issue 1 (2018)
Article No. 8
PP 1-8

4. CONCLUSION

In this suggested approach, extensive study of numerous methods and practices used for sentiment analysis are considered. Methods such as lexicon based approach, SVM, SentiReview, CNN, IML and Morphological Sentence Pattern model are studied in this paper. Each method holds its own unique ability and provides different results. Lexicon based approach and SVM are the methods used in the past, but with the advancement in sentiment analysis various methods such as CNN and IML are being practiced more for better result. Sentiment analysis plays a vital part in reviewing any product, system, etc. The methods stated in paper have their advantages and disadvantages and can be used according to the system.

Acknowledgement

We would like to express a profound sense of gratitude towards Prof. Tatwadarshi P. N., Department of Computer Engineering for his constant encouragement and valuable suggestions. The work that we have been able to present is possible because of his timely guidance and support.

REFERENCES

[1] Y. Han and K. Kim, "Sentiment analysis on social media using morphological sentence pattern model," *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London*, 2017, pp. 79-84.

[2] Z. Nasim, Q. Rajput and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," *2017 International Conference on Research and Innovation in Information Systems (ICRIIS),* Langkawi, 2017, pp. 1-6.

[3] Z. Xiangyu, L. Hong and W. Lihong, "A context-based regularization method for short-text sentiment analysis," *2017 International Conference on Service Systems and Service Management, Dalian,* 2017, pp. 1-6.

[4] M. H. Krishna, K. Rahamathulla and A. Akbar, "A feature based approach for sentiment analysis using SVM and coreference resolution," *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore,* 2017, pp. 397-399.

[5] P. Yadav and D. Pandya, "SentiReview: Sentiment analysis based on text and emoticons," *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore,* 2017, pp. 467-472.

[6] Y. Gao, W. Rong, Y. Shen and Z. Xiong, "Convolutional Neural Network based sentiment analysis using Adaboost combination," *2016 International Joint Conference on Neural Networks (IJCNN),* Vancouver, BC, 2016, pp. 1333-1338.

[7] B. Duncan and Y. Zhang, "Neural networks for sentiment analysis on Twitter," 2015 *IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Beijing, 2015, pp. 275-278.

[8] M. Trupthi, S. Pabboju and G. Narasimha, "Sentiment Analysis on Twitter Using Streaming API," *2017 IEEE 7th International Advance Computing Conference (IACC)*, Hyderabad, 2017, pp. 915-919.

[9] K. Liu, Y. Niu, J. Yang, J. Wang and D. Zhang, "Product Related Information Sentiment-Content Analysis Based on Convolutional Neural Networks for the Chinese Micro-Blog," *2016 International Conference on Network and Information Systems for Computers (ICNISC)*, Wuhan, 2016, pp. 357-361

[10] D. V. N. Devi, C. K. Kumar and S. Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, 2016, pp. 3-8.

[11] A. Ekram, A. Ekram & T. Ekram, S. Islam, Mohammad & Ahmed, Faysal & Rahman, Mohammad. (2015). "Localized twitter opinion mining using sentiment analysis. Decision Analytics".

[12] R. Hegde and Seema S., "Aspect based feature extraction and sentiment classification of review data sets using Incremental machine learning algorithm," *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, 2017, pp. 122-125.

[13] B. Wang and L. Min, "Deep Learning for Aspect-Based Sentiment Analysis." (2015).

[14] Y. Han, K. Yanggon, and Jin-Hee Song, "Building Sentiment Lexicon for Social Media Analysis using Morphological Sentence Pattern Model." *Advanced Science and Technology Letters 136* (2016), pp. 103-106.

[15] S. Jebbara, and P. Cimiano, "Aspect-Based Sentiment Analysis Using a Two-Step Neural Network Architecture," *Communications in Computer and Information Science, vol 641*, 2016.