ASIAN JOURNAL OF INFORMATION TECHNOLOGY
MEDWELL PUBLICATIONS

A Review on Sentiment Analysis: Approaches, Practices and Applications

[1]Okeke Ogochukwu and [2]Amaechi Chinedum
[1]*Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli Anambra State, Nigeria*
[2]*Department of Computer Science, Nnamdi Azikwe University, Awka, Anambra State, Nigeria*

Key words: Sentiment analysis, Naïve Bayes, SVM, opinion mining, unsupervised and supervised algorithms

Corresponding Author:
Okeke Ogochukwu
Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli Anambra State, Nigeria

Page No.: 92-98
Volume: 20, Issue 4, 2021
ISSN: 1682-3915
Asian Journal of Information Technology

Abstract: Sentiment Analysis (SA) has recently become the focus of many researchers because analysis of online text is useful and demanded in many different applications. Analysis of social sentiments is a trending topic in this era because users share their emotions in more suitable format with the help of micro blogging services like twitter. Twitter provides information about individual's real-time feelings through the data resources provided by persons. The essential task is to extract user's tweets and implement an analysis and survey. However, this extracted information can very helpful to make prediction about the user's opinion towards specific policies. The motive of this study is to perform a survey on sentiment analysis algorithms that shows the utilizing of different ML and Lexicon investigation methodologies and their accuracy. The study also focuses on the three kinds of machine learning algorithms for Sentiment analysis-supervised, unsupervised algorithms.

INTRODUCTION

Everyday millions of people use social media and return huge volume of digital data that can be effectively exploited to extract valuable information regarding human dynamics and behaviors. Such data, contains valuable information about user activities, interests and behaviors which makes it inherently appropriate to several applications[1]. In this day and age, social communication channels like Twitter, Facebook and YouTube have obtained, so much popularity. There is a massive amount of information related to distinct individual entities that are recorded every day in in these communication channels. With the increasing popularity of social networking, bloggingand micro-blogging websites, every day a huge amount ofcasual subjective text statements are made available online. The information captured from these texts, could be employedfor scientific surveys from a social or political perspective[2]. Sentiment analysis includes classification of data into various classes like optimistic i.e., good sense or negative i.e., bad sense or neutral i.e., non-effective. Sentiment analysis is the task of perceiving whether a given opinion is positive or negative[3]. For example, a movie review, a person, a political party or a policy or product feature review. Because of the free format of messages and easy accessibility of micro-blogging platforms, most of the data on social media are unstructured[3]. Sentiment analysis techniques have to start with people's data for the analysis of a different kind of area like politic, economy or biology, etc[2]. When it is necessary to make conclusion or final output, especially as regards government policies and programs towards its citizens, it is important to get Opinions of persons. The latest Arab

Spring sensation is an instance of how governments can crumpled when they disregard anxious voters' sentiments[4]. Appropriate training set is required for sentiment analysis for better performance and accurate dataset for improper analysis of the text.

SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is the process of identifying and categorizing the user's emotion or opinion or attitude for any services like movies, products, policies, events to be positive, negative or neutral. The sources of data for this analysis is social communication channels i.e., web site which include reviews, forum discussions, blogs, micro-blogs, Twitter etc. Sentiment Analysis is very popular nowadays because of it has opinioned data of feedback and critiques provided by internet users, showing attitudes and sentiments towards specific topics, products or services[4]. It helps to achieve an understanding of the attitudes, opinions and emotions voiced within an online comment[4]. The large amount of opinionated data is stored in digital forms. Sentiment analysis also known as Opinion mining which uses NLP-Natural Language Processing to follow the emotions, feelings of the public opinion about a particular topic for any product or services. As Sentiment analysis is very famous (no so famous in Nigeria), it can be also useful in many ways in surveys and advertisement campaign by getting the success rate of any product or services with people's opinion or suggestion. It also gives the information about people liking and disliking and company gets much clear idea regarding its product features. Sentiment analysis has increased a lot of acceptance among various zone like politics[5], Stock predictions[6] and marketing/selling and advertisement (to estimate sales of specific products). So, identifying type of sentence is the most important part of opinion mining. Recent or existing research is using both supervised and unsupervised learning technique to provide different techniques for several purpose of sentiment analysis.

MATERIALS AND APPROACHES

Numerous methodologies are available for opinion mining but two main groups are used. The problems of SA will be solved by the first group using by implementing the machine learning approach. The second group uses lexicon-based method which is a linguistically-inclined method. In both groups, many techniques exist. From the following way, we can extract the features of text or sentences.

N-gram: Only one word can be taken by one at a time (unigram) or two words (bigram) up to n words as a result. Unigram features cannot be captured by some opinions. For example, this book is fascinating. It is an optimistic comment if in only unigram model it is fascinating to take it together and negative.

POS tagging: It is the way of words to signify it in content (corpus) as it is linked to its parts of speech in the light of both its definition and its connotation with touching the words. Nouns, pronouns, adjectives, adverbs, etc. are examples of different parts of speech.

Stemming: In this eliminating prefixes and suffixes is the main approach. For example, "running", "sleeping", "ran" can be stemmed from "run" and "sleep" respectively. It basically helps in Cataloging but sometimes it also leads to decrease in cataloging accuracy.

Stop words: Stop words are Pronouns (he/she, it), articles (a, an, the), prepositions (above, in, near, under, besides). These words are nothing but offer no or little information about the emotions. On the internet you can access list of stop words. In the pre-processing step, it can be used to remove them.

Conjunction handling: In general, there is only one meaning of each sentence at a time. But there are certain available conjunction words like: But and while although, however, changes the whole denotation of its sentence. For example, even though the ride was good but it was not up to my hopes. By using these rules throughput can be amplified by 5%[7].

Negation handling: Negation words like "not" inverts the essence of the whole sentence. For example, the movie was not good as "good" in it which is optimistic but "not" upturns the schism to negative.

To identify emotions or opinion words is an important task in many applications in opinion mining. From the given feature, classifying the polarity is basic important task. Positive, Negative and Neutral are three classes where the polarity is categorized. From polarity identification, calculation of sentiment strength, sentiment score etc. can be done using Lexicon techniques. There are various ways and techniques are available for opinion mining, there are majorly two groups used:

- Uses lexicon methods
- Machine learning method which resolves the problems of SA

Lexicon based approach: In this approach,when using the available lexicon techniques for a text which is given will separate the words. In general it performed by aggregation of scores: for example subjective words scores as positive, negative and neutral etc. are summed up separately for same. It assigns a score to each word.

The one which gets the maximum score gives the overall split of the text[8]. It has mainly divided into two parts:

- Dictionary-based
- Corpus-based

Dictionary-based approach: In this system, the user collects a set of sentiments words and seed list is prepared by them. After that, the user start searching for phrasebooks and lexicon to find synonyms and antonyms of particular text. Once this is done, the newly created substitutes are added into the seed list. Up until there are no new words are found to users this process continues.

Disadvantage: There has to have struggle in finding context or domain-oriented emotion words.

Corpus-based approach: Corpus is a basically a term which is a cluster of writing like group of some writing which is often on a very precise matter. In this study, consumers uses the help of corpus text to drawn out the seed list which is in organized situation[2].

Machine learning approach: In this approach, initially classification is performed by taking two different assemblies of the document. Trained data and test data are part of these. This is termed as involuntary classification. Further text is extracted from the features and categorized into I) supervised and II) unsupervised.

Supervised system: Among various kind of datasets, branded training dataset is one of them which is used in supervised system. Each type of class has its own property and advantages and has its label related to its which can be used for this system. Each word, upon arriving is categorized under a label depending on its type and characteristics related with it.

Probabilistic classifier: Predicts or anticipates probability function related to input records among different modules.

Naive Bayes: In this study, to generate possibilities of a group to provide prediction that group of properties belongs to one particular label with help of Bayes theorem using merely a text document as an input. BOW - Bag of Words is a way to extract a text with using machine learning approaches which is simple and easy to implement. This existing model conduct that these all the features which are given autonomous[5]:

$$P(\text{label/features}) = \frac{P(\text{label}) \times P(\text{features/label})}{P(\text{features})}$$

Bayesian network: It is used to manifest relationships among different features. It can be compared with acyclic graph in which nodes represent random variable and edges represent dependencies This model is very pricey and hence it‘s hardly used.

Maximum entropy: By doing encoding, the labeled feature sets are converted into to vectors by using classifiers. This vector is changed and utilized to decide the weights of these features which can be able to use to suppose and predict the label for each of their feature set.

Linear classifier: The characteristics of the linear classification can be implementing by using this classifier which is used to shows predictor as result and can be divided in to two classes:

Support Vector Machine (SVM): This learning model is under supervision to utilize for classification. The most important purpose of this particular model is to assure that this is the best linear separator for classification. This will make a model that results in new information into one or two classes using SVM training.

Neural Network (NN): It is a neural structure of the brain having electronic networks of neurons. In this network, Neuron is the basic component. Neurons are categorized in to three parts- input, hidden and output.

Decision tree classifier: To make division of the data, there is a condition which is used. one class consist those data which mollify the condition and other class consist of the remaining of the data. This technique is called a recursive technique which has two parts: single attribute split and multi attribute split.

Rule based classifier: It is condition based classifier which makes usage of condition or rule like IF, THEN. It can be written as IF condition THEN decision. We can produce the rules based on our requirements at the time of training phase[3].

RESOURCES OF SENTIMENT ANALYSIS

To collect data is the main purpose of Sentiment analysis where social communication channels like Twitter, Facebook or any pre-existing resources.

Blogs and forums: It is source of opinions and emotions where we get information for research purpose and that all information can be used by researchers via Web forums and blogs.

Reviews: There are many available studies which dedicated only on reviews because of their usability with

the opinions and sentiment. During any research, Movie and product reviews were mostly studied by researcher where the main purpose is to get the feedback from the sentiment and opinions.

News articles: News articles, such as financial articles and political reviews are a popular source of sentiment analysis[6]. The main format of News articles texts is structured and formal.

Social networks: Many social networks sites are available from which we can take the opinions and reviews for sentiment analysis like Twitter, Facebook, etc.

Twitter: Tweets are the messages posted by different users, having restriction of 140 characters. Users can read message (called Tweets) of one another. The micro-blogging service which provides this facility is knows as Twitter. By using this tweets which can work as opinions and reviews for future patterns where we can generate the poll results.

Facebook: The provision of posting personal profile, photos, videos and other related information are provided by most famous social networking facility called Facebook which is popular right after it got launched in 2004.

Hence, these much ample amount of information available in form of user‘s message, computer technology which is dependent on sentiment behind this message is introduced known as sentiment analysis (Fig. 1, Table 1 and 2).

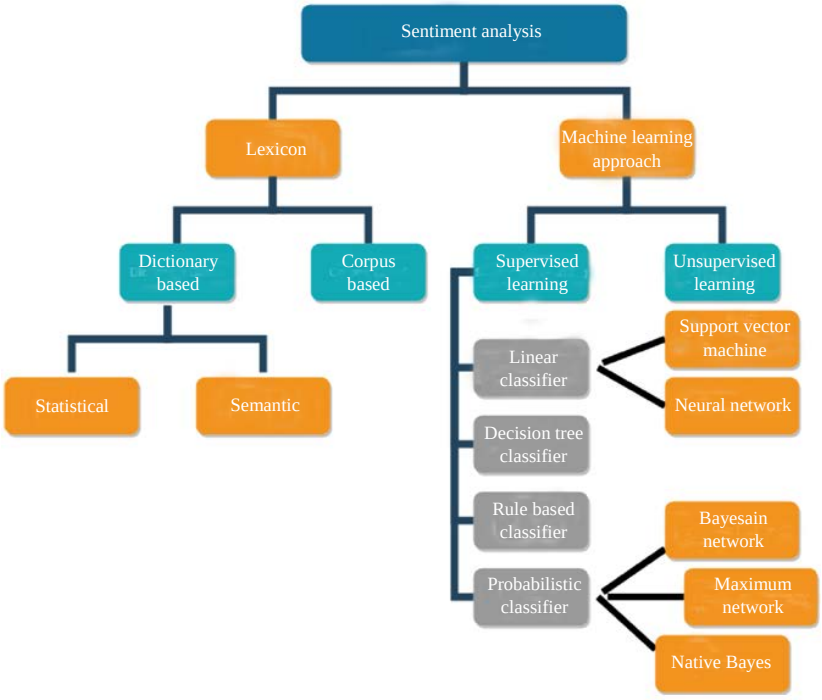

Fig. 1: Various approaches of sentiment

Table 1: Sentiment classification methods

Methods	Techniques involved	Advantages	Disadvantages
Machine learning	Naïve Bayes Decision tree classifiers Neural networks Support vector machine, SVM Logistic Regression Maximum entropy	For the particular application, models can be trained No dictionary requirement Classification accuracy is high	Only two sections. Either positive or negative Sometimes training new dataset cannot be applicable. In that situation, applying labelled training requires more cost
Lexicon-based approaches	Dictionary based approach Manual opinion approach Corpus-based approach	Broader term analysis It does not require a dataset for training. So, it requires less computation overhead	Cannot be implemented for content level classification There is fixed number of words Lexicon based approach always requires linguistic resources. Sometimes it may not be available

Table 2: Literature of related work

Years	Paper title	Methodology used	Review dataset	Accuracy
(2017)	User Sentiments Analysis in Twitter with Social Context[9]	Naive Bayes	Twitter Dataset	73.6%
(2020)	Exerting 2D-Space of Sentiment Lexicons with Machine Learning Techniques: A Hybrid Approach for Sentiment Analysis[10]	Sentiment Lexicon, Hybrid Approach, Pure Machine Learning, Naïve Bayes, SVM Decision Tree	Social Media Dataset	67.02% 73.1% 74.9%
(2013)	Sentiment Analysis and Summarization of Twitter Data[11]	Hybrid Approach,	Twitter Dataset SVM,	89.78% 86.70%
(2020)	Learning Political Polarization on Social Media Using Neural Networks[12]]	Neural Network	Twitter Data	74.15% 64.69%
(2018)	Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants[13]	Valence Aware Dictionary and Sentiment Reasoner (VADER)	Reviews of Electronic product	87.4%
(2018)	A framework for sentiment analysis with opinion mining of hotel reviews[14]	Naive Bayes	Hotel Reviews from OpinRank	83.5%
(2018)	Aspect-Level Sentiment Analysis on E-Commerce Data[15]	Naïve Bayes SVM	Amazon Customer Review Data	90.423% 83.43%
(2019)	Sentiments Analysis for Governor of East Java 2018 in Twitter[16] include the LR in your second work	Naïve Bayes SVM	Twitter	The results of Khofifah's dataset is 77% For the results of Gus dataset, the accuracy is 76%
(2020)	Sentimental classification analysis of polarity multi-view textual data using data mining techniques[17]	Decision Trees Naïve Bayes	Social Media Datasets	98.5%
(2019)	Sentiment Analysis on Naija-Tweets[18]	Adapted Lesk algorithm to facilitate pre-processing	Twitter	99.17%
(2018)	Prediction and Analysis of Sentiments on Twitter Data using Machine Learning Approach[19]	Naïve Bayes SVM	Twitter	73% 83%
(2020)	Philippine Twitter Sentiments during Covid-19 Pandemicusing Multinomial Naïve-Bayes[20]	Multinomial Naïve-Bayes	Twitter Dataset	72%
(2020)	Predicting depression using deep learning and ensemble algorithms on raw twitter data[21]	Random Forest Bernoulli naïve bayes Multinomial naïve bayes	Twitter Data set	72% 73.95% 74.22%
(2020)	Social Media Analysis On Supply Chain Management in Food Industry[22]	SVM	Social media platform (Twitter)	
(2020)	Exploring Neural Network Approaches in automatic Personality Recognition of Filipino Twitter users[23]	Neural Network	Twiiter Data Set	90%

RELATED WORK/APPLICATIONS

Decision making support: Building a website that could perform decision making is a very crucial part. Analysis has its own advantage like; it can lead to different ideas which can help us to make decision in day to day life such as choosing a good restaurant to go for dinner or buying a new car or selecting a good movie to watch, etc.

Business related application: Every wants to create a innovative and newest product which can fully satisfy their customers. To achieve more valuation of their product, organization can assemble all the needs of their users and enhance the efficiency of product from feedback collected from their customers.

Predictions and trend analysis: Tracking views of public by sentiment examination which enable any person to predict the market scenario like the stock market which helps any person for trading and polls market. By using this all opinions user can predict the market trends.

Evaluation of policies of political office holders: With the recent ENDSARS protest across Nigeria, getting an evaluation of the policies of the government have become necessary. With the use of sentiment analysis, citizen's emotions can be evaluated.

Crime detection: Although this is new, with the use of sentiment analysis tools, flash points of crime can be ascertained by evaluating the emotion of comments in the social media.

CONCLUSION

In the current generation, the rapid growth of technology is affecting human lives significantly. People buying the products by entering the shops physically is getting replaced by people logging into the various online

shopping websites. Judgement of a polices being good or bad by the own experience of an individual is replaced by experience and views of other people over the policies.

The ratings on a product, number of reviews, number of positive and negative reviews, polarity of opinions towards positive or negative have become the major criteria for product sale. More reviews and its polarity towards positivity, more reliable it is for a new buyer. Thus, collecting tons of data(reviews/opinion/feedback), analyzing and processing it to categorize it into positive, negative or neutral becomes crucial.

Sentiment analysis is the current hot topic trying to provide a solution for this. Our Literature survey has presented several approaches for sentiment analysis by applying various algorithms, predominantly machine learning algorithms (supervised and unsupervised). Different algorithms in combination with different feature selection techniques have also been applied to extract the best feature to perform classification and identification of polarity. It is observed that unigrams give better presentation than any n-grams model. The results also show implementing feature selection using chi square will improve classification accuracy. Supervised machine learning algorithms provide higher accuracy and performance.

REFERENCES

01. Belcastro, L., F. Marozzo and D. Talia, 2019. Programming models and systems for big data analysis. Int. J. Parallel Emergent Distrib. Syst., 34: 632-652.
02. Saberi, B. and S. Saad, 2017. Sentiment analysis or opinion mining: A review. Int. J. Adv. Sci. Eng. Inf. Technol., 7: 1660-1667.
03. Nakov, P., A. Ritter, S. Rosenthal, F. Sebastiani and V. Stoyanov, 2019. SemEval-2016 Task 4: Sentiment analysis in Twitter. SemEval-2016 task 4: Sentiment analysis in Twitter December 3, 2019, Association for Computational Linguistics pp: 1-18.
04. Arunachalam, R. and S. Sarkar, 2013. The new eye of government: Citizen sentiment analysis in social media. Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP), October 14, 2013, Nagoya, Japan, pp: 23-28.
05. Kang, H., S.J. Yoo and D. Han, 2012. Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews. Exp. Syst. Appl., 39: 6000-6010.
06. Abdul-Mageed, M., M. Diab and S. Kubler, 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang., 28: 20-37.
07. Kathuria, A. and S. Upadhyay, 2017. A novel review of various sentimental analysis techniques. Int. J. Comput. Sci. Mobile Comput., 6: 17-22.
08. Bollen, J., H. Mao and X. Zeng, 2011. Twitter mood predicts the stock market. J. Comput. Sci., 2: 1-8.
09. Nanaware, S.S., 2016. User sentiments analysis in twitter with Social context. Int. Educ. Res. J., 2: 39-41.
10. Khan, M.Y. and K.N. Junejo, 2020. Exerting 2D-space of sentiment lexicons with machine learning techniques: A hybrid approach for sentiment analysis. Evaluation, Vol. 11.
11. Bahrainian, S.A. and A. Dengel, 2014. Sentiment analysis and summarization of twitter data. 16th International Conference on Computational Science and Engineering, March 6, 2014, IEEE, pp: 227-234.
12. Belcastro, L., R. Cantini, F. Marozzo, D. Talia and P. Trunfio, 2020. Learning political polarization on social media using neural networks. IEEE Access, 8: 47177-47187.
13. Park, C.W. and D.R. Seo, 2018. Sentiment analysis of Twitter corpus related to artificial intelligence assistants. 5th International Conference on Industrial Engineering and Applications (ICIEA), April 26-28, 2018, IEEE, pp: 495-498.
14. Zvarevashe, K. and O.O. Olugbara, 2018. A framework for sentiment analysis with opinion mining of hotel reviews. Conference on Information Communications Technology and Society (ICTAS), March 8-9, 2018, IEEE, pp: 1-4.
15. Vanaja, S. and M. Belwal, 2018. Aspect-level sentiment analysis on E-commerce data. International conference on inventive research in computing applications, July 11-12, 2018, IEEE pp: 1275-1279.
16. Buntoro, G.A., 2019. Sentiments analysis for governor of East Java 2018 in Twitter. Sinkron: Jurnal dan Penelitian Teknik Informatika, 3: 49-55.
17. Yassir, A.H., A.A. Mohammed, A.A.J. Alkhazraji, M.E. Hameed, M.S. Talib and M.F. Ali, 2020. Sentimental classification analysis of polarity multi-view textual data using data mining techniques. Int. J. Electr. Comput. Eng., 10: 5526-5533.
18. ajo, T., O. Daramola and A. Adebiyi, 2019. Sentiment analysis on naija-tweets. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 28-August 2, 2019, Student Research Workshop, pp: 338-343.
19. Rao, C.S., S. Prasad and V.V. Rao, 2018. Prediction and analysis of sentiments on Twitter data using machine learning approach. Int. J. Comp. Sci. Inf. Secur., 16: 33-42.

20. Delizo, J.P.D., M.B. Abisado and M.I.P. De Los Trinos, 2020. Philippine Twitter sentiments during Covid-19 pandemic using multinomial naïve-bayes. Int. J. Adv. Trends Comput. Sci. Eng., 9: 408-412.
21. Shetty, N.P., B. Muniyal, A. Anand, S. Kumar and S. Prabhu, 2020. Predicting depression using deep learning and ensemble algorithms on raw twitter data. Int. J. Elec. Comput. Eng., 10: 3751-3756.
22. Nale, K.A., 2020. Social media analysis on supply chain management in food industry. Int. Res. J. Eng. Technol., 7: 3077-3087.
23. Tighe, E., O. Aran and C. Cheng, 2020. Exploring neural network approaches in automatic personality recognition of Filipino Twitter users. Proceedings of Philippine Computing Science Congress, March, 2020 Baguio City, Philippines, pp: 1-9.