IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

Review Paper on Sentiment Analysis Technique by Different Machine Learning Approach

Sakshi Koli[1*], Ram Narayan[2]

[1,2]Dept. of Computer Science and Engineering, Tula's Institute, Dehradun, India

**Corresponding Author: Sakshi.koli@tulas.edu.in,*

DOI: https://doi.org/10.26438/ijcse/v7i11.125129 | Available online at: www.ijcseonline.org

Accepted: 12/Nov/2019, Published: 30/Nov/2019

Abstract—The growing popularity of social media, E-commerce, blogs and any social field created a new platform where anyone can discuss and exchange his/her views, ideas, suggestions and experiences about any product or services in market. This state of affairs open a new area of research called Opinion Mining and Sentiment Analysis. Opinion Mining and Sentiment Analysis is an extension of Data Mining that extracts and analyzes the unstructured data automatically. In this review paper our aim is to present the details study over Opinion Mining and Sentiment Analysis, its different techniques, methods etc.

Keywords— Introduction, Sentiment analysis techniques, Literature review, Comparative analysis, Conclusion

I. INTRODUCTIONS

Sentiment analysis is the process of determining the opinion feeling, emotions in the piece of text. With the ever-increasing popularity of social networking, micro-blogging and blogging websites, a large amount of data is engendered every day. These social websites depend largely on the user content that are generated rapidly. Typically, when people destine to purchase a product, they browse online sites to gain some information about the products before they make their final purchase. They take into consideration the available reviews and ratings of these products on these websites before making purchases. Thus, in order to make this process efficient and to automate it, several sentiment analysis techniques are used. Sentiment analysis is usually conducted at different levels varying from coarse-level to fine-level. Coarse-level analysis is mainly concerned with finding the sentiment score of the entire document whereas fine-level deals with attribute level. Sentence-level sentiment analysis comes in between these two [1]. There are many researches in the area of sentiment analysis of user reviews. The performance of sentiment analyzer is largely dependent on the topic. As a result, we cannot determine which classifier is the best.

II.SENTIMENT ANALYSIS TECHNIQUES

Sentiment Analysis can be implemented in three ways: 1) Sentiment Analysis based on Supervised Machine learning technique, 2) Sentiment Analysis by using Lexicon based Technique and 3) Sentiment Analysis By combining the above two approaches.

2.1 MACHINE LEARNING TECHNIQUE

Machine learning is the key part of data mining , natural language processing ,Pattern recognition, image recognition, and expert system, topic spotting , medical diagnosis . In machine learning approach we developed a computer programs that can teach themselves to change and grow when new data set exposed. Machine learning approach works by using different computer algorithm. Learning that is being done in machine learning is always relies on the past behaviour , some sort of past observation on the different field , past experiences .Machine learning based Sentiment Analysis or classification can be done in two ways: 1) Sentiment Analysis by using supervised machine learning techniques and 2) Sentiment Analysis by using unsupervised machine learning techniques.

2.1.1Supervised machine learning: In Supervised Machine learning techniques, two types of data sets are required training data and test data . An classifier learns the classification factors of the document from the training set and the accuracy in classification can be evaluated using the test set. Various machine learning algorithms are available that can be used very well to classify the documents. The machine learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used successfully in many research and they performed well in the sentiment classification . The first step in Supervised Machine learning technique is to collect the training set and then select the appropriate classifier. Once the classifier is selected, the classifier gets trained using the collected training set. The most important feature in the Supervised

© 2019, IJCSE All Rights Reserved

Machine learning technique is feature selection. The classifier selection and feature selection determines the classification performance. The most frequent techniques used for feature selection are:

Terms and their frequency: The data has been changed to vector form by using the TF-1DF function and find out the unique words from all the documents has been provided. It provides a single text file of containing all the data need to require for the document classification. Vector space model is the most generally used method in tweet representation. This model utilize feature entries and their weights to express the document information.

***Part of speech (POS) tagging* information:** Part of speech is the part of feature extraction in data mining various field. The process of part-of-speech tagging allows to automatically categorize each word of text in terms of which part of speech it belongs to: noun, pronoun, adverb, adjective, verb, interjection, intensifier, etc. Prabowo and Thelwall(2009)[6] used this approach in their studies and they utilized feature set easily by identifying adjectives *and adverbs.*

Negation Handling: Negation handling implements the process of conversion of the sentiment of the text from positive to negative or from negative to positive by using special words: "no", "not", "don't" etc.

Tokenization into N-grams: Tokenization is a process of breaking a bag-of-words from the text. The incoming string from the web gets broken into comprising words and other elements. The universal separator for identifying individual words is whitespace. Tokenization of social-media data is noticeably more complicated than tokenization of the general text since it contains numerous emoticons, URL links, abbreviations that cannot be easily separated as whole entities. It is a general practice to combine accompanying words into phrases or n-grams, which can be unigrams, bigrams, trigrams, etc. The second step of supervised learning is to classifying the unseen data based on the trained model in first step.

2.1.1 Unsupervised learning: In unsupervised learning method, no training set is available to learn the parameters. Unsupervised learning find the hidden pattern in hidden data in various field . Clustering algorithm uses unsupervised learning method approach. There are various clustering algorithms like K-mean clustering algorithm, K- Medoids algorithm, DBSCAN and OPTICS, hidden markov algorithm. Unsupervised learning provides the capability to learn more larger and complex models. In unsupervised learning strategy, the learning can be preceded in hierarchical fashion from the observations resulting into ever more deeper and abstract levels of representation. unsupervised learning uses unlabeled dataset.

2.2 LEXICON BASED TECHNIQUE

Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation approach to opinion mining in which sentiment polarity of features present in the given document are determined by comparing these features with semantic lexicons. Semantic lexicon contains list of words whose sentiment orientation is determined already. It classifies the document by aggregating the sentiment orientation of all opinion words present in the document, documents with more positive word lexicons is classified as positive document and the documents with supplementary negative word lexicons is classified as negative document. The processing steps of lexicon based sentiment analysis are the following:

Pre-processing: This step clean the document by removing HTML tags and noisy characters present in the document, by correcting spelling mistakes, grammar mistakes, punctuation errors and incorrect capitalization and replacing non dictionary words such as abbreviations or acronyms of common terms with their actual term.

Feature Selection: This process Extract the feature present in the document by using techniques like POS tagging. Pos tagging find the adverb, noun, adjectives, etc in the documents by predefined entities.

Sentiment score calculation: For each extracted sentiment word, check whether it is present in the sentiment dictionary or not. If present with negative polarity, w then s = s-w or If present with positive polarity, w then s =s + w. Where we 1^{st} initialize s with 0 If s is below a particular threshold value then classifying the document as negative otherwise classify it as positive.

Sentiment Lexicon Construction: Sentiment lexicon can be constructed in three ways: 1) manual lexicon construction, 2) dictionary-based lexicon construction and corpus-based lexicon construction. In manual lexicon construction, the lexicons are constructed manually. It is very difficult and time consuming task. In dictionary-based lexicon construction, a miniature set of sentiment words and their polarity are determined manually and then this set is widened by adding more words into it using WordNet dictionary or SentiWordNet dictionary and their synonyms and antonyms. In corpus-based lexicon construction, it considers syntactic patterns of the words in the document. It needs annotated training data to produces accurate semantic words.

© 2019, IJCSE All Rights Reserved

2.3HYBRID TECHNIQUES

Some researchers joint the supervised machine learning and lexicon based approaches together to get better sentiment classification performance. Fang et al. [2] 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 discovered that general purpose lexicon performed very poor while domain specific lexicon performed very well. Their system yielded around 66.8% accuracy. Mudinas et al. [3] combined lexicon based and learning based approaches to develop a concept-level sentiment analysis system, pSenti.

III. LITERATURE REVIEW

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. Turney [4] used bag-of-words approach for sentiment analysis. In this approach, associations between the individual words are not considered and a document is represented as a mere collection of words. To conclude the overall sentiment, sentiment of every word is determined and this value is combined with some aggregation functions. He described the polarity of a review based on the average semantic orientation of tuples extracted from the review where tuples are phrases having adjectives or adverbs. He discovered the semantic orientation of tuples using the search engine Altavista Kamps et al. [5] used the lexical database WordNet to determine the emotional content of a word along different dimensions. They implemented a distance metric on WordNet and determined the semantic orientation of adjectives. WordNet database stores of words connected by synonym relations. Baroni et al. [6] developed a system using word space model formalism that overcomes the difficulty in lexical substitution task. It shows the local context of a word along with its overall distribution. Balahur et al. [7] introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. EmotiNet used the theory of Finite State Automata to identify the emotional responses triggered by actions.

Domingos et al. [8] found that Naive Bayes works well for certain problems with highly dependent features. This is shocking as the basic assumption of Naive Bayes is that the features are independent .One of the participant of SemEval 2007 Task No. 14 [9] used coarse-grained and fine-grained approaches to identify sentiments in news headlines. In coarse-grained approach, they performed binary classification of emotions whereas in fine-grained approach, they categorized emotions into different levels. Knowledge-based approach is found to be difficult due to the requirement of a huge lexical database. Social network generates huge amount of data every second, which is significantly larger than the size of available lexical databases. Therefore, sentiment analysis often becomes arduous and erroneous.

Zhen Niu et al. [9] introduced a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is based on Bayesian algorithm. Here, weights of the classifier are attuned by making use of representative feature and unique feature. Representative feature is the series of data that classify class and "Unique feature" is the Sequence of data that helps in distinguishing classes. Using those weights, they evaluated the probability of each classification and thus improved the Bayesian algorithm.

Barbosa et al. [10] designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to trim down the labelling effort in developing classifiers. To begin with they classified tweets into subjective and objective tweets. Later, subjective tweets are classified as positive and negative tweets.

Celikyilmaz et al. [11] developed a pronunciation based word clustering method for normalizing noisy tweets. In pronunciation based word clustering, words having parallel pronunciation are clustered and assigned common tokens. They also used text processing techniques like passing on similar tokens for numbers, HTML links, user identifiers and target organization names for normalization. After performing normalization, they used probabilistic models to identify polarity lexicons. They classification using the Boos-Texter classifier with these polarity lexicons as features and obtained a reduced error rate.

Wu et al. [12] proposed an influence probability model for Twitter sentiment analysis. If @username is found in the body of a tweet, it is influencing action and it contributes to influencing probability. Any tweet that begins with @username is a re-tweet that represents an influenced action and it contributes to an influenced probability. They observed that there is a strong correlation between these probabilities.

Pak et al. [13] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that process, there is a possibility of error since emotions of tweets in training set are labelled solely based on the polarity of emoticons. The training set is also fewer efficient since it contains only tweets having emoticons.

Xia et al. [14] used an ensemble framework for sentiment classification. Ensemble framework is discovered by combining various feature sets and classification techniques.

© 2019, IJCSE All Rights Reserved

In their respective work, they used two types of feature sets and three base classifiers to form the ensemble framework.

Two types of feature sets are implemented using Part of speech information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are elected as base classifiers. They processed ensemble discrete methods like Fixed combination, Weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy. Number of attempts are made by some researches to identify the public opinion about movies, news etc. from Twitter posts. V.M. Kiran et al. [17] utilized the information from other publicly available databases like IMDB and Blippr after proper modifications to aid Twitter sentiment analysis in movie domain. Melville [15], Rui [16], Ziqiong [17], Songho [18], Qiang [19]and Smeureanu [20] used naïve bayes for classification of text. Naïve bayes is popular method in text classification. It is measured as one of the most simple and efficient approaches in NLP. It implement by calculating the probability of an element being in a category. First the former probability is calculated which is afterwards multiplied with the likelihood to calculate the final probability . The method consider every word in the sentence to be independent which makes it easier to implement but less accurate. This the reason why the method is given the name „naïve". Rui [16], Ziqiong [17], Songho [18], and Rudy [21] used SVM (Support Vector Machines) in their approach. The sentiment classification was done using discriminative classifier. This process is depend on structural risk minimization in which support vectors are utilized to classify the training data sets into two different classes based on a predefined sections. This multiclass SVM can also be utilized for text classification. Songho [18] implemented centroid classifier which allocate a centroid vector to different training classes and then again consider this vector to determine the comparison values of each element with a class. If the similarity value exceed a Known threshold then the element is allocated to that class (polarity in this case).Songho [18] also used K-Nearest Neighbour (KNN) approach which finds K nearest neighbours of a text document among the documents in the training data set. Then classification is evaluated on the basis of similarity score of a class with respect to a neighbour. Long-Sheng [22] used a neural network based approach to combine the advantages of machine learning and information retrieval techniques.

IV. COMPARATIVE ANALYSIS

In most of the cases the supervised machine learning approaches outperformed the unsupervised lexicon based approaches. But, the requirement of big labelled training data set for supervised machine learning approaches; compel the researchers to adopt the unsupervised methods, as it is very easy to collect unlabelled dataset.

Table 1

Paper	Approach	Dataset	Technique	Accuracy
P.D Turney [4]	unsupervised	Epinion (Review automobile, Movie)	PMI	84 % Automobile 66.6% Movie
Pang et al. [23]	Supervised	Movie Review	SVM	unigram results- 82.9
			Naïve Bayes	unigrams+ bigrams- 82.7 unigrams+ POS- 81.9
			maxent	Unigrams+position- 81.6
Hu and Liu [22]	Unsupervised	Customer Review (blogspare)	Lexicon(com bined BPN and SO indexes)	84%
Abbasi et al. [11]	Supervised	Movie Review	lexicon	71%
A. Khan et al. [5]	Supervised	Customer Review	lexicon	86.6%
Zhang et al.[2]	Unsupervised	Product	lexicon	82.62%
Zhang et al.[24]	Hybrid	Twitter tweets	ML and Lexicon	85.4%
Mudinas et al. [3]	Hybrid	Customer	ML and Lexicon	82.3%
Fang et al. [14]	Hybrid	Domain	ML and Lexicon	66.8%

V. CONCLUSION

The main contribution of this review is to discuss the various sensitive analysis Techniques employed on different machine learning techniques with different feature selection .The paper also gives a qualified comparison of all the techniques based on their accuracy , approach and dataset selection. We also extensively show the results of various supervised and unsupervised sentiment analysis techniques to efficiently detect the sentiments from the different frame-work .

REFERENCES

[1] Mejova, Y. (2009). Sentiment Analysis: An Overview. Comprehensive exam paper, available http:// www. cs. uiowa. edu/~ ymejova/ publications/ Com ps Yelena Mejova. pdf [2010-02-03].

[2] **X Fang, J Zhan"** Sentiment analysis using product review data" , Journal of Big Data , Issue 1/2015,2015.

[3] M. Levene , A. Mudinas, D. Zhang, "Combining lexicon and learning based approaches for concept leve lsentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York ,NY, USA, Article 5, pp. 1-8,2 012.

[4] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classificationofreviews",40th annual meeting,

© 2019, IJCSE All Rights Reserved

association for computational linguistics " , Association for Computational Linguistics, pp. 417– 424,2002.
[5] A Kamps, "sentiment analysis in twitter using machine learning techniques", Fourth International Conference on Computing, Communications and , Tiruchengode, pp. 1-5, 2013.
[6] D. Pucci, M. Baroni, F. Cutugno, and A. Lenci, "Unsupervised lexical substitution with a word space model," in presenting of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence ,Citeseer , 2009.
[7] A. Balahur, J. M. Hermida, and A. Montoyo, "Building and exploiting emoti- net, a knowledge base for emotion detection based on the appraisal theory model," Affective Computing, IEEE Transactions on, vol. 3, no. 1,pp. 88– 101, 2012.
[8] P. Domingos , M. Pazzani presnsented , "On the optimality of the simple bayesian classifier under zero-one loss," Machine Learning, 29, 103–130 (1997), Kluwer Academic Publishers.,Netherlands.
[9] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for micro blog by machine learning," in the proceeding of the Computational and Information Sciences (ICCIS), Fourth International Conference on, pp. 286–289, IEEE, 2012.
[10] J. Feng ,L. Barbosa discovered "Robust sentiment detection on twitter from biased and noisy data," in 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010.
[11] J. Feng ,D. Hakkani-Tur, , A. Celikyilmaz, "Probabilistic model-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010. 17).
[12] F. Ren, and Y. "Learning sentimental influence in twitter," in the proceedings of the 2016 International Conference on Future Computer Sciences and Application (ICFCSA), pp. 119– 122, IEEE, 2011.
[13] A. Pak and P. ParoubekK, "Twitter as a corpus for sentiment analysis and opinion mining," in LREC, vol. 2010, 2010.
[14] Rui Xia, ChengqingZong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011)1138–1152.
[15] Melville, WojciechGryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD"09, June 28–July 1, 2009, Paris, France.
[16] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li,"Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications xxx (2011) xxx–xxx.
[17] Qing Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.
[18] Songbo Tan, Jin Zhang, "An empirical study of sentiment analysis for chinese documents", Expert Systems with Applications 34 (2008) 2622–2629.
[19] Ziqiong Zhang, Rob Law ,Qiang Ye, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.
[20] Ion SMEUREANU, Cristian BUCUR, "Applying Supervised Opinion Mining Techniques on Online User Reviews", Informatical Economic vol. 16, no. 2/2012.
[21] Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach." Journal of Informetrics 3 (2009)143–157.
[22] Hui-Ju Chiu ,Long-Sheng Chen, Cheng-Hsiang Liu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Info 5 (2011).
[23] S. Vaithyanathan, and B. Pang, L. Lee, and "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002.
[24]B.Liu, Zhang, R. Ghosh, M. Dekhil "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories,2011
[25] 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 Issue3,pp.1 34,2008.
[26] S. Li , R. Xia, C. Zong, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.
[27] A Palve, R. D.Sonawane, A.D. Potgantwar," Sentiment Analysis of Twitter Streaming Data for Recommendation using, Apache Spark", International journal of scientific research in network security and communication (IJSRNSC), Vol.5 , Issue.3 , pp.99-103, Jun-2017.
[28] K .Sarvakar, U. K Kuchara **"**Sentiment Analysis of movie reviews: A new feature-based sentiment classification**"**, Isroset-Journal (IJSRCSE) Vol.6 , Issue.3 , pp.8-12, Jun-2018.

Authors Profile

Ms sakshi Koli pursed Bachelor of Science from dehradun Insititute of technology, Dehradun in 2014 and Master of computer Science from DIT University in year 2016. She is currently working as Assistant Professor in Department of Computational Science and engineering at Tulas Insitute, Dehardun ,India. Her main research work focuses on Machine Learning , Big Data Analytics and Data Mining.She has 2/5 years of teaching experience .

Mr. Ram Narayan pursed Bachelor from UP Technical university and Master from Graphic Era University Dehradun Uttrakhand, India. He is presently working as Assistant Professor in Computer Science and Engineering Department at Tula's Institute Dehradun Uttrakhand, India. He has more than 7 years of teaching experience. His area of interest is Pattern Recognition, digital signal processing.

© 2019, IJCSE All Rights Reserved