TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

Investigations on Sentiment Analysis and Opinion Mining in Social Media Data

L. Sudha Rani, Asst. Prof, CSE Department, G.Pulla Reddy Engineering College, Kurnool, sudha1021@gmail.com

Dr. S. Zahooer-Ul-Huq, Professor, CSE Department, G.Pulla Reddy Engineering College, Kurnool, szahoor@gmail.com

Article Info
Volume 82
Page Number: 8718 - 8723
Publication Issue:
January-February 2020

Article History
Article Received: *5 April 2019*
Revised: *18 Jun 2019*
Accepted: *24 October 2019*
Publication: *08 February 2020*

Abstract:
In the recent years, the advancement in the social media and big data leads to the development of sentiment analysis and opinion mining. The sentiment analysis is the process of data classification based on the Lexicons to define the polarities of the words. Many organizations depend on the data analytics for their business decision making. This paper presents the extensive survey on the approaches used for sentiment analysis and opinion mining. This survey concentrated on research issues and challenges faced by the sentiment analysis and opinion mining algorithms.

Keywords: *Opinion mining, Lexicon approaches, Domain based algorithms, Machine learning algorithms*

I. Introduction

Big Data analytics is one of the emerging fields in the recent years due to the development of web 2.0 and also with the internet advancement [1-3]. The users can post any data in the internet without the knowledge of programming and also the social media is accessible throughout the world with the help of the internet from anywhere and anytime. Due to the advancements in the field of social media, this created opportunities for researchers and academicians to continue their research. More than 1 billion users are using the social media which are posting the unstructured data in seconds. The huge volumes of data posted in the social media is referred as "big data".

II. Sentiment Analysis in Social Media

The information is gathered from various social media such as twitter, blogs etc., Sentiment analysis is one of the approaches to classify the data and make the opinion of the data. This opinion can be utilized by the users based on the requirements. The applications of the sentiment analysis are business, finance, public auctions and politics. The sentiment analysis in the politics will give the idea about the position of the politicians. The opinion mining is also used for taking the interest of the public whenever government introduced policies. The classification of the sentiment is the major motive of the sentiment analysis. Figure 1 shows the model of sentiment analysis. Aspect level, sentence level and document level are the three categories used for sentiment analysis. Aspect level analysis deals with the opinion mining where it considers the different aspects of the document. Sentence level analysis deals with the analyzing of sentences in the content posted by the users and document level analysis deals with negative and positive class in the data.

The classifiers for sentiment analysis are broadly classified into lexicon based techniques and machine learning based techniques. The correct results prediction can be obtained using the supervised learning compared to the unsupervised and semi supervised learning. The supervised learning approaches require the data labeling which is time consuming and more expensive process. The semi supervised learning uses both labeled data and unlabelled data. This method can be used in the pre-processing mechanism. Lexicon based learning methods are used to analyze the sentiment in the reviews gathered from different domains

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

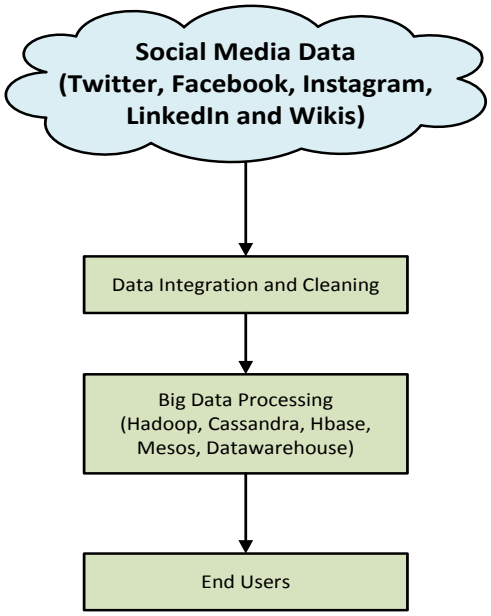

Figure 1: Process for Big Data in social Media

a. Data Pre-processing

The data gathered from different domains contains the noise which needs to be normalized before it is classified. In [4], the authors explore the usage of pre-processing stage in the SVM classifier to find the appropriate features. The selection of appropriate features lead to the improvement in the classification mechanism.

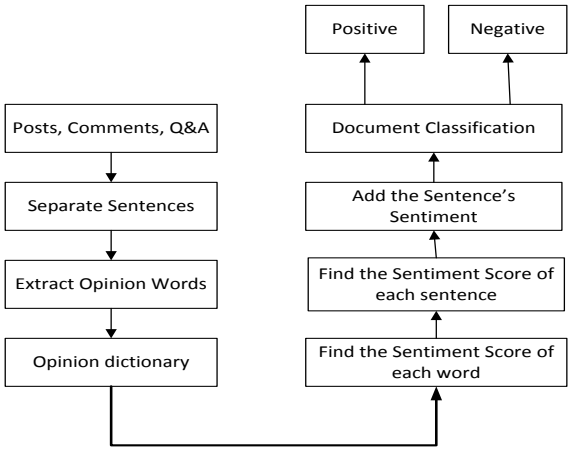

Figure 2: Model for Sentiment Analysis

The different approaches used for feature selection are Feature presence, Feature frequency and Inverse document frequency. In [5], the authors accessed the effect of repeated letters, urls, lemmatization and negation on the classification mechanism.

The data gathered from different domains contains the noise which needs to be normalized before it is classified. In [4], the authors explore the usage of pre-processing stage in the SVM classifier to find the appropriate features.

The selection of appropriate features leads to the improvement in the classification mechanism. There are different approaches used for feature selection are Feature presence, Feature frequency and Inverse document frequency.

The data pre-processing contains three steps such as normalization, tokenization, and part of speech. The lemmatization is a pre-processing technique which is more efficient in text classification [6]. Bigrams and Unigrams are the other feature selection approach used for classification. Unigram is more efficient compared to the Bigrams [7].

b. Sentiment Analysis using Machine learning Approaches

The machine learning approaches are used to extract the features of data which is used for classification of big data. These mechanisms have the ability to classify the large volumes of online data and automatically classify the information based on their domains. Figure 3 shows the classification of machine learning approaches.

Supervised learning: Many algorithms in machine learning approaches use the supervised learning for classification of a data. The data in the supervised learning is divided in to two sets. First one is labelled data and another one is test data. The algorithm is first trained with training data set then it will forward to the test data set. The major algorithms in the machine learning approaches are Decision tree classifiers, Linear Classifiers, Rule based classifiers, Support vector machines and Naïve Bayes.

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

Unsupervised learning: This algorithm does not use the trained set of data or labelled data for classification instead it used unlabelled data where ever it is not possible to label the data. But, the unsupervised learning requires huge volumes of data to train the model; otherwise it leads to inconsistent results.

Semi supervised learning: These algorithms had the benefit of both supervised and unsupervised learning and uses both labeled and unlabelled data [17].

These algorithms overcome the drawback of unsupervised learning by adding the labeled data to the early understanding [18].

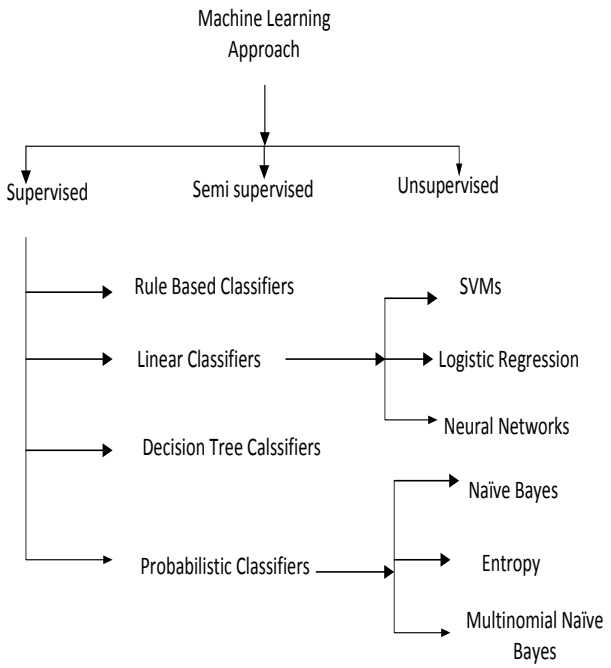

Figure 3 Classifications of Machine Learning Approaches

c. Sentiment Analysis using the Lexicon approach

Lexicon approach is used to find the sentiment lexicons of words which contain score for each word; it may either positive, negative or neutral. Figure 4 shows the classification of Lexicon approaches.

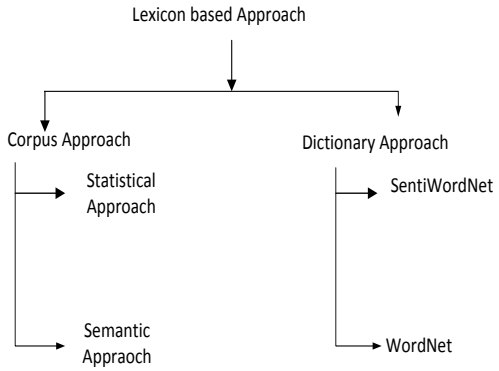

Figure 4 Classifications of Lexicon Approaches

III. Deep Learning Approach for Sentiment Analysis

From the past decade, neural networks have become more popular in the natural language processing (NLP). The sentiment analysis is one of the areas where the NLP is applied. The usage of neural networks is extended for solving many machine learning problems. The only requirement for the neural networks is to define the architecture based on the number of hidden layers used, weights at the hidden layer, activation function, interconnections, threshold value for the data, etc.

The deep learning approaches are used in the neural networks when there is good training data with enough time for better classification. The deep learning is subset of machine learning algorithms where it attempts the high level abstraction from the submitted data with complex models. Deep learning algorithms utilize the deep neural networks to train the good represents of the input data to perform the particular tasks.

a. Traditional Neural Networks

Neural Networks plays vital role in the cognitive science and machine learning algorithms. The neural networks are extensively used in the field of pattern recognition and image processing and also they are popular in solving the problems of natural language processing (NLP). The neural network is applied in word embedding and sentiment classifications. Figure 5 shows the structure of the fully connected neural network.

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

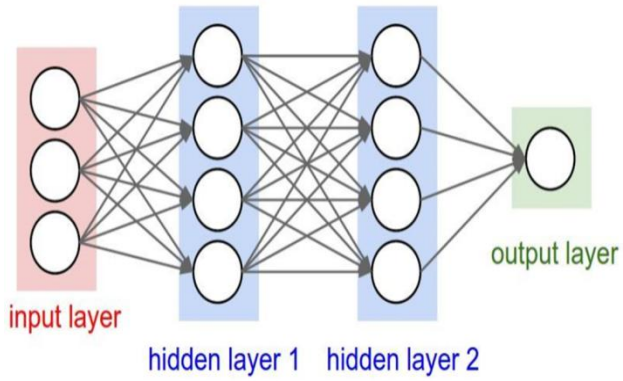

Figure 5: Fully connected Neural Network

b. Word Embedding using deep learning

The deep learning network in the NLP cannot take the raw words as input in the classification. The network only understands the functions and numbers. Therefore, the words need to be converted in to vectors or else in to the word embeddings. These word embeddings captures the semantics and word characteristics it they are explained properly. The training of word vectors can be done by loading huge volume of raw corpus into the deep network and train them by giving the sufficient time. In [13], the author proposed the first deep learning model for large scale networks to train the network for "distributed representation of words". Figure 7 explains about the training of network using the raw corpus, where it is represented as word sequence. The major idea behind this is to relate each word in the vocabulary, to represent the probability function for word sequence with respect to words feature vectors, and to learn the probability function parameters along with feature vectors of words.

Based on the user requests, the words may embed in any dimension. The higher dimensionality represents the capturing of more information; meanwhile it incurs more computational cost. Therefore, a trade-off between the computation cost and high dimensionality is required. To help the researchers in this regard *Google* [14]has released the trained data set with vectors on the *Google News.* The *Google* data set contains three million words and phrases which are represented in three hundred dimensional vectors. The data set can be directly used in the deep networks to process any NLP application.

In general NLP uses the windows approach for processing the sentences. The windows approach follows the tagging mechanism to the words based on their neighboring words in the sentences. There a fixed size of window is selected and the fixed amount of words is loaded in the deep network to perform the tagging to the middle words. The padding is done for the starting and ending words in the sentences. Figure 9 shows the window based approach for deep networks

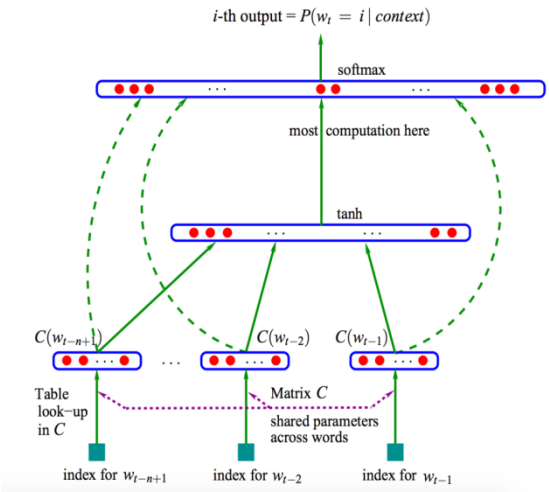

Figure 6: Deep learning model for large scale data sets

c. Applications of deep learning

There are many machine learning algorithms which show good performance to process the NLP applications, but still there are some drawbacks which deep learning should look after to overcome those drawbacks. The major advantages of deep learning approaches are:

Features Optimization: The deep learning approaches take the input as word embeddings instead of features. This word embeddings contains the information about the text. The hidden layers in the deep learning can learn the features when they are at the training phase. Therefore, the

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

classification techniques of the machine learning algorithms are not required for the deep learning algorithms.

Good Representation: In the deep leaning algorithms, the features are learned by the networks at the time of training stage for a specific task, the word representation contains the context information which is learned from the raw corpus. Therefore, there is no need of manual representation of word features in the deep learning algorithms.

Adaptability: The sentiment analysis contains task variations. The deep learning have the ability to adapt the changes in the architectures.

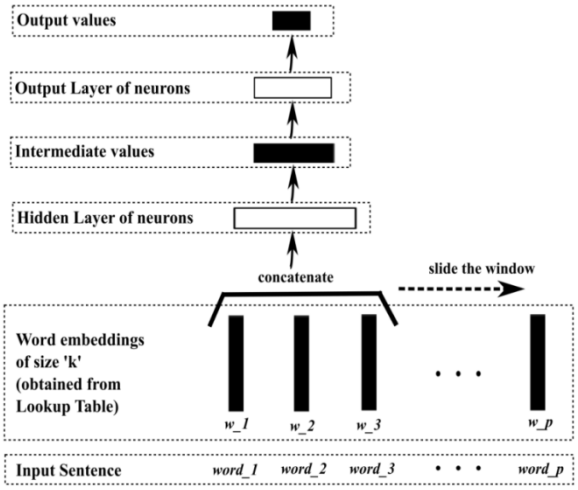

Figure 7: Window approach for deep network architecture

IV. Evaluation of Sentiment Analysis Algorithms

The performance of the existing algorithms is compared using the precision, recall and F-score. Precision is used to retrieve the data that are more likely to be relevant and applicable. Recall is the process of retrieving the data that are relevant. F-measure is defined using the Precision and Recall. Table 3 shows the comparison of algorithms in Sentiment Analysis.

Author	Feature Selection Approach	Concept	Precision	Recall	F-measure	Accuracy
Zhen et al. [8]	Intrinsic domain relevance And Extrinsic Domain relevance	Opinion feature	0.65	0.61	0.63	-
Li Cheng et al. [9]	Semi Supervised learning	SVM Classifier	-	-	-	0.63
Jianping et al. [10]	Sentiment Score	Rule based Classifier	0.84	0.30	0.82	-
Wen et al. [11]	Heterogeneous augmentation	SVM Classifier	-	-	-	0.54
Rui et al .[12]	Data Expansion	Naïve Bayes	-	-	-	0.81

Table 3:Comparison of algorithms in Sentiment Analysis

V. Conclusion

This paper concentrated on extensive survey of sentiment analysis and the opinion mining algorithms. This paper also discussed about sentiment classification algorithm with respective to their performance. The domain adaption techniques, opinion extraction methods, machine learning approaches and Lexicon approaches are discussed. Further, the research challenges and issues from the conventional algorithms related to sentiment analyses are figured out.

VI. Reference

[1] Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. *arXiv preprint arXiv:1312.4617*.

[2] Aisopos, F., Tzannetos, D., Violos, J., &Varvarigou, T. (2016). Using n-gram graphs for sentiment analysis: an extended study on Twitter. Paper presented at the Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on.

[3] Cheng, Q., Li, T. M., Kwok, C.-L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk

January-February 2020
ISSN: 0193-4120 Page No. 8718 - 8723

and emotional distress in Chinese social media: A text mining and machine learning study. Journal of medical Internet research, 19(7).

[4] K. R. Venugopal, K. G. Srinivasa, and L. M. Patnaik, Soft Computing for Data Mining Applications. Springer, 2009.

[5] D. Sejal, K. G. Shailesh, V. Tejaswi, D. Anvekar, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "Query Click and Text Similarity Graph for Query Suggestions," in Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, pp. 328{341, 2015.

[6] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment Classification using Machine Learning Techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Languag Processing, ACL, vol. 10, pp. 79-86, 2002.

[7] X. Glorot, A. Bordes, and Y. Bengio, Domain Adaptation for Large-scale Sentiment Classification: A Deep Learning Approach," in Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 513-520, 2011.

[8] ZhenHai, K. Chang, J.-J. Kim, and C. C. Yang, Identifying Features in Opinion Mining *via* Intrinsic and Extrinsic Domain Relevance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 623-634, 2014.

[9] Li Cheng and S. J. Pan, Semi-Supervised Domain Adaptation on Manifolds," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 12, pp. 2240 - 2249, 2014.

[10] Jianping Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, D. Wen, and Y. Gao, Web-Based Traffic Sentiment Analysis: Methods and Applications," IEEE transactions on Intelligent Transportation systems, vol. 15, no. 2, pp. 844-853, 2014.

[11] D.Wen,J. Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, , and Y. Gao, Web-Based Traffic SentimentAnalysis: Methods and Applications," IEEE transactionson Intelligent Transportation systems, vol. 15,no. 2, pp. 844-853, 2014.

[12] Rui Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, Dual Sentiment Analysis: Considering Two Sides of One Review," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2120-2133, 2015.

[13] YoshuaBengio, RejeanDucharme, Pascal Vincent, and ´ Christian Janvin. 2003. A neural probabilistic language model. The Journal of Machine Learning Research, 3:1137–1155.

[14] Google data set available at https://code.google.com/ archive/p/word2vec/.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In ICLR.

[16] Ronan Collobert, Jason Weston, Leon Bottou, Michael ´ Karlen, KorayKavukcuoglu, and PavelKuksa. 2011. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537.

[17] Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," in Proceedings of ACL, 2014, pp. 347-358.

[18] K. Hanhoon, Y. S. Joon, and H. Dongil., "Senti-lexicon and improved naı¨vebayes algorithms for sentiment analysis of restaurant reviews," Expert SystAppl ,39:6000–10, 2012.

[19] L. Cheng and S. J. Pan, Semi-Supervised Domain Adaptation on Manifolds," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 12, pp. 2240 - 2249, 2014.