International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 4261-4264 www.ijaem.net ISSN: 2395-5252

A Study on Different Approaches of Sentiment Analysis for Big Data

Seema Singh, Praveen Mishra

[1,2]*Assistant Professor, Institute of Technology & Management, Gida Gorakhpur, Uttar Pradesh*
[3]*Student, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh*

Submitted: 25-07-2021 Revised: 04-08-2021 Accepted: 06-08-2021

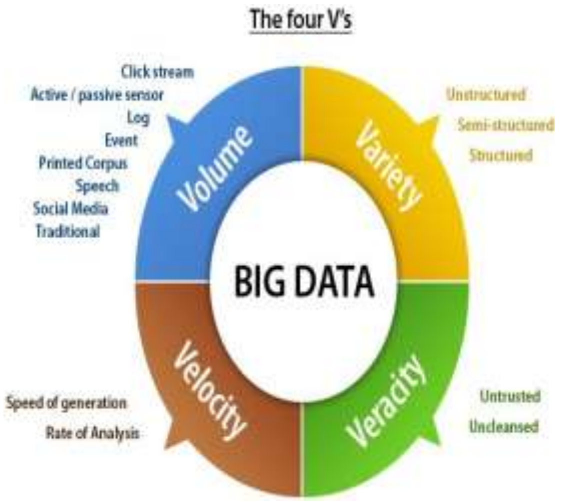
ABSTRACT

Sentiment computing (SA) plays a very vital role in opinion mining and promotes new application opportunities in product reviews and decision making. From a given piece of text, sentiment Analysis does the job of labelling the people's opinions in three different categories as neutral, positive and negative. The Sentiment Analysis techniques are used in a various departments like ecommerce sites, government organization and entertainment industries. There are various algorithms prevalent for sentiment analysis. The present study compares and contrasts the different techniques of sentiment analysis.

Key Words: Sentiment analysis, Big data, Machine Learning, Lexicon-based approach, Hybrid Approach

I. INTRODUCTION

Big data work as fuel and analytics as engine in digital transformation. With the increase in the volume and variety of information on social networking websites (Facebook, Twitter, LinkedIn etc.) an increasing interest among researchers can be seen regarding sentimental analysis (SA). Many industries and organizations have started applying analytics to such web generated data to make better business decisions. Apart from the data gathered from social networking sites, companies also create their own web sites to collect reviews about their products. By applying approaches of data analytics enterprises can bring a great change in their way of functioning and decision making. Big data analytics and computational knowledge based solution that may decrease the complexity and analytical burden over accessing and rectifying huge amount of data. Big data is acronym of four Vs- volume, velocity, variety and 4th can be represented as value, variability, or veracity. Volume refers to data set and is an important characteristic of big data. Velocity indicates data growth rate at which is generated makes large amount of dataset. For example- in each second thousands of tweets is tweeted in twitter account. Variety refers to the data composition (e.g., unstructured, structured and semi structured data



Source: Blue Copper Technologies

Fig. 1: Four Vs of Big Data

In SA domain, the texts classes are associated to either of positive, negative or neutral/irrelevant. It is broadly classified into three level categories namely, document level, sentence level and aspect based. Its online data retrieval system is based techniques that analyse the web text. These methods started by retrieving the pertinent web texts, splitting the texts, checking the spelling and counting the recurrence of specific words. However, it is regarded that its abilities are very restricted in interpreting the sentences and separating significant information.

Recently a significant number of methods and techniques have been suggested by researchers to improve the efficiency of SA. Therefore, this study aims to classify and compare the different SA techniques.

II. REVIEW OF LITERATURE

SA emerged as a favourite research topic that extracts people's opinions, thoughts, feelings and behaviours [1]. As data is quite large, the extraction of sentiment information from web text data now becomes a challenging task [2]. The advent of social networks after early 2000s, people started to share their opinions, feelings and emotions through various social media platforms.

Thus, with increased coverage of social media, people's opinions shared through networking sites can be significantly meaningful for the decision making.

The prominent approaches for SA are: 1. Lexicon Based approach, 2. Machine Learning (ML) based approach and, 3. Mixed approach [3]. Lexicon-based approach is an unsupervised method, wherein text data are categorised into a defined sentiment clusters. Sentiment scores are calculated from the sentiment lexicon (dictionary of words) [4]. Whereas, ML approach can either be supervised (e.g. classification) or unsupervised (e.g. clustering) [1].

[5] stated that SA measures the polarity of data either using machine learning approach (supervised/unsupervised) or using lexicon-based approach (unsupervised) or hybrid/combined approach. [6] Posited that the usefulness of linguistic features and existing lexical resources used in micro-blogging to detect the sentiments of twitter messages.semantic analysis are the methods of latent topic extraction. One of the major limitations of unsupervised learning technique is that “for the trained accuracy, it generally need a large amount of data. It is observed that unsupervised models often leads to incoherent topics as the objective functions of topic models cannot be align well with human judgments always.

Supervised learning

Supervised learning is regarded as a mature and successful solution for sentiment analysis. In supervised learning a machine is trained using labelled data. Once machine is trained, it is provided new set data to analyse and produce the correct outcome based on learned logarithm. The widely known supervised classification algorithms are: K-Nearest Neighbors (KNN), Naïve Bayes, and Support Vector Machine (SVM). The notable limitation can be seen with supervised learning is that it is prone t quantity and the quality of the training data and may most likely fail when training data are biased or unclear.

Semi-Supervised learning

As the name suggest, Semi-Supervised learning (SSL) models comprise the characteristics of supervised and unsupervised learning approaches. The strength this technique is that it learns from labeled as well as unlabeled data. SSL is a rather new ML approach to opinion mining and is appreciated as it efficient when data is large heterogeneous and unlabeled.

A. Lexicon-based approach

The lexicon technique involves with the extraction of sentiment orientation from the word or phrases in the document. There are two main approaches for the sentiment analysis using Lexicon technique- Dictionary based and corpus-based approach. Dictionary-based Lexicon can be created manually, or automatically, by indentifying the opinion seed words. By utilizing Tokenizer the input data is transformed into tokens. For the every new token created is matched for the lexicon in the dictionary. A score is added or subtracted to the total score based on the sentiment polarity match i.e. positive/negative.

In past few years Corpus-based approaches have been extensively used to explore written texts. This approach has many advantages such as calculating the frequency of words, analysing the actual pattern in natural, etc.

B. Hybrid approach

Hybrid approach exhibits the combined properties of lexicon approach (speed) and machine learning approach (accuracy)**.** This approach keeps the robustness of the machine learning of the

1. Sentiment Analysis Approaches

There are three broad approaches, as stated above, used to study the opinions - machine learning approaches, lexicon based approaches, and hybrid approaches. Support Vector Machine (SVM) and Naïve Bayesian classification are approaches based on Machine learning. SVM, a supervised learning model, is utilized mainly to analyse and identify the data patterns that can be used analytics. Naïve Bayesian Classification follows the Naïve-Bayes theorem and Bayesian probability and uses the concepts of maximum likelihood. These approaches based on NLP and lexical resources mainly use WordNet and parts of speech information [7].

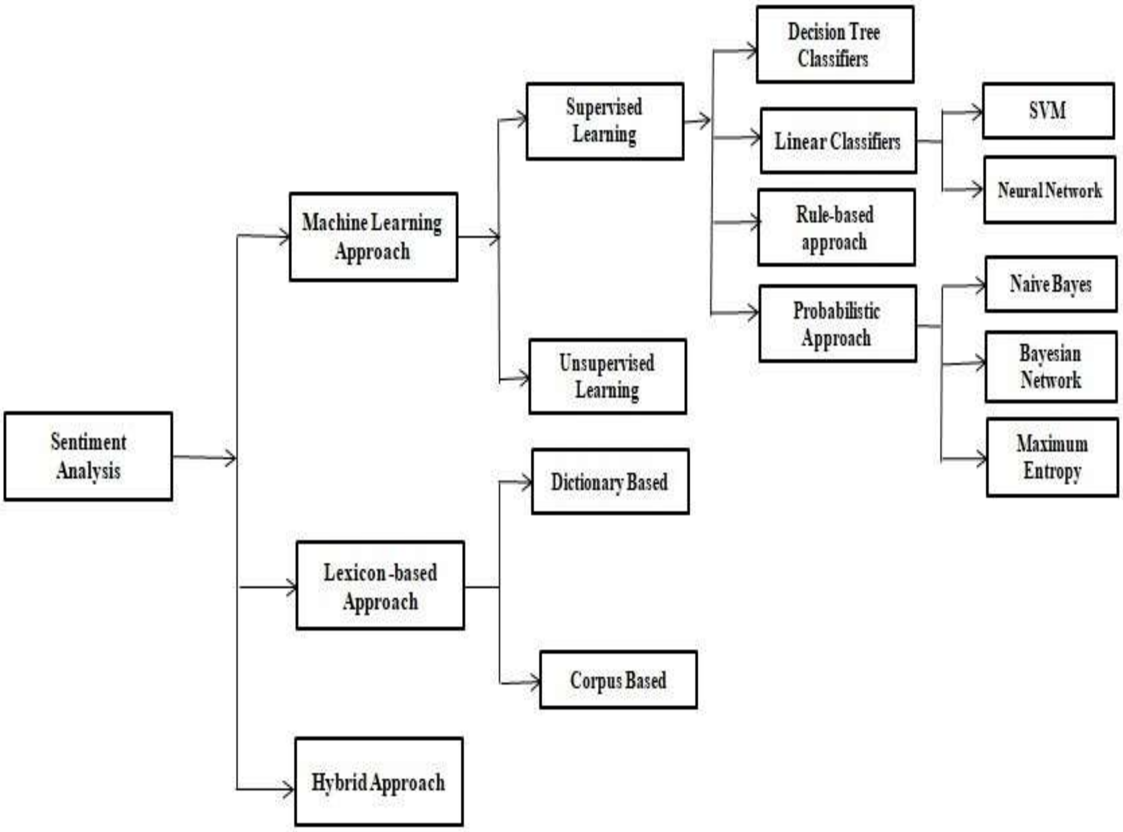
C. Machine Learning-Based approach

ML approaches are more practical approaches among other types of approaches due to its ability to fully automatic implementation. Its algorithms are such that it addresses a combination of methods to automatically detect the available pattern in the given data set. ML based sentiment classification methods can be unsupervised, supervised or semi-supervised.

[i]Unsupervised Learning

Collection of unlabeled documents is easy, but gets sometimes difficult to create labelled training documents in the text classification. There it is unsupervised learning methods which eliminate the issues of creating training data. Linear Discriminant Analysis and Probabilistic latentstatistical method and orients the training parallelly on a manual configuration of the symbolic method. The advantage of their hybrid approach is to obtain the best of both mechanism-stability and readability from a well designed lexicon, and the high accuracy from a robust supervised learning algorithm.

Fig. 2: Sentiment Analysis Approaches and Techniques



III. CONCLUSION

In this study we have tried to cover different approaches of sentiment analysis. We examined that all of the above algorithms and techniques/approaches, which is used in sentiment analysis, are not 100 percent accurate. These approaches have their own merits and demerits. The supervised machine learning approaches have better performance than the unsupervised ones. There are various popular algorithm are in use for supervised machine learning approach such as naïve bayes, maximum entropy and support vector machine. However, the merits of unsupervised methods cannot be ignored as supervised methods require large amounts of labeled training data that are very costly. A significant research is being carried out to enhance the performance sentiment analysis using hybrid approach which combines machine learning approach as well as lexicon approach.

REFERENCE

[1] Danneman N, Heimann, R (2014) Social Media Mining with R: Deploy Cutting-Edge Sentiment Analysis Techniques to Real-World Social Media Data Using R, www.it-ebooks.info.

[2] Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González- Castaño, F J (2016) Unsupervised method for sentiment analysis in online texts. Expert System Application, 58: 57–75.

[3] Chauhan A P, Patel K M (2015) Sentiment Analysis Using Hybrid Approach: A Survey. International Journal of Engineering Research and Applications. 5(1): 73-77.

[4] Sun S, Luo C, Chen, J (2017) A review of natural language processing techniques for opinion mining systems. Information Fusion, 36 (July): 10–25.

[5] Prabowo R, Thelwall M (2009) Sentiment Analysis: A Combined Approach. Journal of Informetrics, 3(2): 143-157.

[6] Kouloumpis E, Wilson, T, Moore J D (2011) Twitter sentiment analysis: The good the bad and the omg!. Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21.

[7] Hu X, Tang L, Tang J, Liu H (2013) Exploiting social relations for sentiment analysis in microblogging.in Proceedings of the sixth ACM international conference on Web search and data mining, 2013: 537-54