IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

Sentiment Analysis Using Machine Learning: A Survey

Pooja Mahaling[1*], P.V Bhaskar Reddy[2]

[1] School of Computing & Information Technology, REVA University, Bangalore, India
[2] School of Computing & Information Technology, REVA University, Bangalore, India

Corresponding Author: mahalingpooja@gmail.com, Tel.: +91-8105863763

DOI: https://doi.org/10.26438/ijcse/v7si14.6871 | Available online at: www.ijcseonline.org

Abstract— Social media is flooded with data that is generated by bloggers, committee, business, health, marketing, education, etc., in large amount. Extracting the data information from various fields like social media, marketing, reviews, conference publications and advertisement is done to perform sentiment analysis. These text data have some emotions hidden in it, and data analysing is carried out by natural language processing (NLP). NLP is application of artificial intelligence that help machine to read text by simulating the human capability to know language. Sentiment analysis is type of data mining that measures the opinion of the users or the customer or the blogger through the natural language processing, which can be utilized to extricate and dissect emotional data from web for the most part web based life. The main purpose of sentiment analysis is to classify emotions into positive, negative and neutral. The applications of sentiment analysis are in the financial market, area of reviews of consumer services and products to monitor customer sentiment and catch the trending topics. Sentiment analysis has challenges like multilingual sentiment analysis, emotion detection, and data sparsity from the different data by social media, marketing, emails, advertisement, movie review etc.

Keywords— Sentiment analysis, natural language processing, artificial intelligence.

I. INTRODUCTION

From the last few years have so much discussion of artificial intelligence (AI), and role of artificial intelligence in health, productivity, and marketing. Artificial intelligence is the investigation of making machines elegant, and minimizes the human effort. This can be explained by Machine learning [7]. AI is the utilization of man-made brainpower this development that empowers systems to gain directly from models, information, and experience and furthermore enables PCs to perform explicit tasks intelligently, by learning from examples. Machine learning is used in various fields like malware detection, marketing, bioinformatics, research, healthcare analytics etc. AI frameworks are set an undertaking, and given a tremendous measure of dataset to use as instances of how this activity can be accomplished and from which to distinguish pattern [1][2]. This large amount of data is generated from the social media like twitter, email, face book, marketing, conference publications, movie review, healthcare etc. The framework at that point figures out how best to accomplish the ideal yield by applying the reasonable calculation.

Extracting the data information from various fields like social media, marketing, reviews, blogger, email, face book, twitter etc for carrying out sentiment analysis. Among all these generated data from various fields like social media, web, bloggers and twitter. Twitter is best data set as it contains the hash tags, emotions, emojis, images etc, this data analysis is done by sentiment analysis. The elucidation of convolution depends on Twitter information examination to design noteworthy neural structure, to improve the exactness and examination speed. First we learn generally speaking vectors for word portrayal by unsupervised learning on wide Twitter information gathering, which gives the word estimation data as the words embeddings. A short time later, link this word depiction with the earlier extremity score highlight and best in class includes as opinion highlight set. These capabilities is joined and sustained into profound convolution neural systems to prepare and foresee the assumption grouping names of the tweet [3]. The main purpose of sentiment analysis is to classify emotions into positive, negative and neutral [4].

The generated data from the social media, emails, twitter, conference publications, product or movie review, news, bloggers etc, contains noisy and unstructured data which will have an effect on the performance of the sentiment classification. This can be done by the two methods that are lexicon polarity method [2] and machine learning method [5][6][7].

Lexicon-based based strategies make utilization of the rundown of words which were at that point characterized and wherein each word is related with a particular feeling [2]. Dictionary feeling is to distinguish word-conveying

conclusion in the announcement and after that to foresee assessment communicated in the content. AI strategies frequently depend on regulated order approaches where assumption recognition is done as positive and negative [5][6][7].

II. LITERATURE SURVEY

AI is a utilization of man-made brainpower that makes PC frameworks to gain straightforwardly from precedents, information, and experience [7]. These frameworks are set an undertaking, and given a lot of information to use as instances of how this assignment can be accomplished or from which to recognize designs. This large amount of data is generated from social media, twitter, face book, Google applications, email, marketing, healthcare, biomedical etc [4]. The obtained data can be classified by using machine learning algorithms. These AI calculations can be ordered into two they are administered (structured) learning, unsupervised learning.

In supervised machine learning is the system that is labelled. This marked classifications every datum point into at least one gathering, for example, 'mangoes' or 'grapes'. The structure makes sense of how this data known as preparing information is sorted out, and utilizes this to anticipate the classifications. Unsupervised learning will be learning without names. This intends and recognizes attributes this makes information guides pretty much comparative toward one another, for instance by making groups and allotting information to these bunches. AI is firmly identified with the fields of insights which give a scope of devices and educate how AI frameworks manage probabilities or vulnerability in basic leadership.

Social media is flooded with data that is generated by bloggers, committee, business, healthcare, marketing, education, conference, biomedical etc., in large amount. These text data have some emotions hidden in it, and this data analysis is carried out by Natural language processing [8]. The main purpose of sentiment analysis is to classify human emotions into positive, negative and neutral [5][6][7]. The main reason of studying the sentiment analysis is that there was little amount of text is available on the some organization website of the manufacture review. The organization wanted to find the opinion or the sentiment of the public towards the product and services therefore it have conducted the opinion polls, surveys and different groups. And the survey of sentiment is been carried out by the two main types they are truth and opinions [4]. Truth is objective expressions about entities, actions and their property. Sentiment analysis gives the association the office to overview the open feelings towards the occasion, produces, and surveys or identified with them. The vast majority of the investigations is concentrating on acquiring estimation examination by dissecting lexical and syntactic component that are communicated expressly through supposition words, feelings, shouts marks, emoticons and so forth. Generally the generated data is unstructured. This can be done by machine learning algorithms.

Among all these generated data from social media, web, bloggers and twitter. Twitter is best data set as it contains the hash tags, emotions, emojis, images etc. All those tweets for the most part convey individual perspectives or feelings of product. Estimation examination is a method that separates the client conclusions and feeling from tweets. This is a less demanding approach to recover client perspectives and suppositions, contrasted with poll or reviews. Twitter is a famous miniaturized scale blogging administration, enables clients to post tweets, status message with length up to 140 characters [2]. These tweets normally convey individual perspectives or feelings towards the subject referenced in the tweets. Slant examination is a method that separates the client assessments and notion from tweets. It is a less demanding approach to recover client perspectives and feelings, contrasted with poll or overviews.

Sentiment analysis first finds the objectives on which suppositions have been communicated in a sentence, and afterward decides if the sentiments are certain, negative or impartial [9]. The objectives are objects, and their parts, properties and highlights. An item can be an item, administration, singular, association, occasion, social insurance theme, and so forth. In an item survey information sentence, it recognizes item includes that have been remarked on by the analyst and decides if the remarks are sure or negative. Sentiment analysis for twitter data can be carried out by Lexicon polarity and is extended by Senti bag [2][13] to get the tweet feeling extremity score. Lexicon method uses dictionary based approach and finding opinion words from the data and searches for synonyms and antonyms [2]. And wherein sentiment polarity score is vocabulary based estimation include, and a couple methodologies ordinarily use it as an assumption include for tweet estimation split achievement. This score can be calculated by estimating the PMI (Point-wise shared information) between the word and the negative or positive end gathering of the tweet [2].

Tweet commonly made out of partial expression a verity of noise and unstructured sentence, poor syntax structure and non lexicon words. These commotion and unstructured twitter data will affect the execution of tweet opinion characterization. The clamour can be expelled as: [10]

- Removal of all non grammar characters in the tweet.
- Eliminate web links.

- Removal of numbers as these numbers do not contain sentiment information
- Replace negative references like no, not, can`t
- Expand acronyms to their full forms like "asap" which is "As soon as possible"
- Removal of stop words. These can be an, the, than etc.
- Replace feelings and emoticon. The feelings and emoticon are author state of mind articulation as symbols in the tweet.
- Tokenization using natural language procession in data.
- Removal of punctuations.

III. METHODOLOGY

From the study results come across different types of algorithm, literature survey on sentiment analysis. Machine learning application of artificial intelligence uses the natural language processing for understanding of the human language. This can be done by machine learning algorithms like SVM (Support Vector Machine) and Nive Bayes algorithms. The aim is to classify the collected dataset using SVM [4].

The building blocks of sentiment analysis are processing, segmentation, tokenization, emotion detection, token score assignation, POS (Parts of speech) tagging and syntactical analysis, polarity calculation and finally the score [4][11][12].

A. Processing: Collecting of data from the social media like twitter, face book, bloggers, marketing etc. which is unsupervised data.

B. Segmentation: In this stage, we find boundaries of sentences in text and by using regular expressions or supervised machine learning models; it decides the finish of a sentence and the start of another sentence.

C. Tokenization: This process breaks a sentence into a group of tokens (words) for the rest of the pipeline to process. In this process the transformation from unstructured format to a structured format and breaking a sentence string into words (smaller strings).

D. Emotion Detection: Finding the emotion in the previous step where we have divided the sentence and finding the emotions in that and else apply the tokenization again. These tokenized sentences are fed for the next block.

E. Interjection Detection: Among the received tokenised sentence find the connection for the comment towards the product, the emotion or the opining.

F. Token Score Assignation: From the last step sentence ids distinguishes emotions as positive and negative and decide the score as positive opinion or the negative opinion.

G. POS Tagging and Syntactical Analysis: POS (Parts Of Speech) tagging is labeling each token by what it corresponds to in a certain dataset is English grammar. Therefore, according to parts of speech that are defined label each token by its proper grammar value, that is, the corresponding part of speech.

H. Polarity Calculation: after the POS tagging and comparing with dataset with the collected bank of words, score is calculated whether it is positive statement or the negative or the neutral.

I. Score: once it is been completed the final score is given in between 0 to 1.

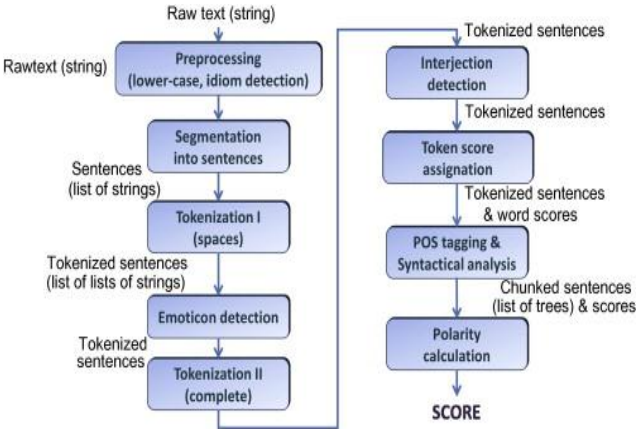

Fig1. Building blocks of sentiment analysis

IV. ALGORITHM

Machine learning algorithms can be classified into three they are: Supervised Learning, Unsupervised Learning and Reinforcement Learning. These further classified into algorithms like decision tree, Logistic Regression, Linear Regression, Support vector machine in supervised leaning. Unsupervised learning algorithms like KNN, principal component analysis (PCA). From a last survey of algorithms Reinforcement Learning is not used in sentiment analysis.

Out of these above algorithms SVM is utilized to order the writings as positives or negatives and neutral. SVM functions excellently for content classification because of its points of interest, for example, its capability to deal with expansive highlights. Another favorable position is SVM is powerful when there is an inadequate arrangement of precedents and furthermore on the grounds that the vast majority of the issue are directly distinguishable. Bolster Vector Machine has indicated expected outcomes from the past study in assessment investigation [4]. This has utilized to arrange the

writings as positives or negatives. SVM functions admirably with content classification because of points of interest, for example, its capability to deal with huge highlights. Another preferred standpoint is SVM is strong when there is a meager arrangement of models and furthermore on the grounds that the majority of the issue are directly divisible [4]. Bolster Vector machine has demonstrated expected outcomes in past research in estimation examination. Four viable estimates that have been utilized in this investigation depend on disarray lattice yield, which are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

• Precision (P) = TP/(TP+FP)

• Recall (R) = TP/(TP+FN)

• Accuracy (A) = (TP+TN)/(TP + TN + FP + FN) AUC (Area under the (ROC) Curve) = 1/2.((TP/(TP+FN))+(TN/(TN+FP))

• F-Measure (Micro-averaging) = 2.(P.R)/(P+R)

This content order adequacy had been normally estimated utilizing F1, exactness, and AUC. F1 measures joined viability measure dictated by accuracy and review. The zone under the ROC bend (AUC) has turned into extent estimation of execution of managed classification rules. In any case, the basic type of AUC is just appropriate to instance of two classes [4].
The building blocks of sentiment analysis like processing, segmentation, tokenization, emotion detection, token score assignation, POS (Parts of speech) tagging and syntactical analysis, polarity calculation and finally the score is been carried out by SVM algoritham [4][11][12].

V. CONCLUSION

From the study results is use of SVM algorithm is globally accepted and gives the accurate results. The outcome additionally demonstrates that by utilizing chi-square element choice will essentially improves the classification of the data collected.

REFERENCES

[1] Zhao Jianqiang1, Gui Xiaolin1- Deep Convolution Neural Networks for Twitter Sentiment Analysis IEEE 2017.

[2] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In EMNLP, vol. 14, pp. 1532-1543. 2014.

[3] Nurulhuda Zainuddin, Ali Selamat- Sentiment Analysis Using Support Vector Machine Conference Paper · September 2014

[4] Jianqiang Z. Combing Semantic and Prior Polarity Features for Boosting Twitter Sentiment Analysis Using Ensemble Learning. In Proc. Data Science in Cyberspace (DSC), IEEE International Conference on. IEEE, pp.709-714,2016

[5] Hagen, M., Potthast, M., Büchner, M., & Stein, B.. Twitter Sentiment Detection via Ensemble Classification Using Averaged Confidence Scores. In European Conference on Information Retrieval, Springer, Cham, 2015, pp. 741-754,2015

[6] Bhumika M. Jadav M.E. Scholar, L. D. College of Engineering Ahmedabad, India- Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis , International Journal of Computer Applications Volume 146 – No.13, July 2016

[7] Abdalraouf Hassan1, Ausif Mahmood, Convolutional Recurrent Deep Learning Model for Sentence Classification, 2017 IEEE.

[8] S. Behdenn, F. Barigo, G. Belalem- Document Level Sentiment Analysis: A survey, EAI 2018

[9] Bhumika M. Jadav Ahmedabad, Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis International Journal of Computer ApplicationsVolume 146 – No.13, July 2016

[10] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka- SentiFul: A Lexicon for Sentiment Analysis, IEEE Transactions on affective computing, vol. 2, no. 1, January- 2011

[11] S.M.Shamimul Hasan, Donald A. Adjeroh- Proximity-Based Sentiment Analysis, 2011 IEEE

[12] Montejo-Ráez, A., Martníez-Cámara, E., Martní-Valdivia, M. T., &Ureña-López, L. A..A knowledge-based approach for polarity classification in Twitter.Journal of the Association for Information Science and Technology, 65(2), pp.414-425, 2014

Authors Profile

Pooja Mahaling, Currently pursuing M.Tech in Data Engineering and Cloud Compiuting from REVA University, Bangalore, Karnataka. Year of pass out: 2019, intrested in AI and its applications; and also carried out internship on Machine Learning from CRL, BEL, Banglore.

Dr. Bhaskar Reddy P.V working as a professor in school of C & IT REVA University, Bangalore. He has 10 years of teaching experience. He published & presented many papers in peer-reviewed International journals & Conferences. His research interests include image processing and Data Mining.