International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 1 pp. 889-894 Jan 2019-June 2019 www.csjournals.com

Analysis of Sentiments using K-Nearest Neighbor

[1]Pooja Rani, [2]Jaswinder Singh
[1]M.Tech Scholar, [2] Assistant Professor, Department of Computer Science & Engineering,
Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

Abstract: The great impact of social media has led to the discovery of sentiment analysis. The social media provides an ability to share thoughts, opinions, and emotions. The available data on social media has contributed to vast research using sentiment analysis. This paper highlights certain discussions regarding the review-based classification and performance analysis of collected data through the use of K-Nearest Neighbor classifier. Instance-based learning method is the simplest, easy to learn, very popular, highly effective algorithm that is useful in categorizing the object that has similar functionality.
Keywords: Sentiment analysis, Reviews, K-Nearest Neighbor.

I. INTRODUCTION

The term assessment examination (SA) is prevalently known as sentiment mining which is a procedure of feeling characterization generally passed on by a content that might be sure, negative or nonpartisan. The symbols and punctuations are basically used to express opinions such as emoticons, smileys etc. The concept of online networking is expanded each and everywhere throughout the technology of Internet.A large number of individuals offers and expresses their feelings over the media examining about the brands whom they associate with. At the point when individuals post their thoughts and suppositions on the web, one gets muddled, unstructured content. When such types of sentiments are identified over the media, then the information gained from such sentiments represents fruitful results benefiting large companies or organizations. This data is very helpful to monitor performance of different brands and to locate time periods and aspects receiving polar sentiments. The brands can be celebrities, political parties or events, products, movies etc. Sentiment classification is very important for several industries or business. While the areas of sentiment analysis are all about boosting the performance via analysis of shifts in public opinion.This paper includes six sections. The first section describes the introduction about sentiment analysis, second section of paper explains levels of sentiment analysis, third section describes the k-Nearest Neighbor, fourth section of paper explains the related work, fifth section of paper explains the steps of methodology followed, sixth and seven discusses the results and conclusion respectively.

II. SENTIMENT ANALYSIS: LEVELS

There are three major levels discussed below and classification of such levels depends upon the analysis level.

Document Level: It is usually known as the classification representing a sentiment at the document-level as the main objective is to determine the whole opinion on the document basis having a positive or negative sentiment. For example, it was assumed that for a known form of text, the text basically presents the overall opinion whether positive or negative on the basis of single entity or event [2][19][20]. As such method contains only single entity, they are not considered suitable for the text-based entities.

Aspect Level: It is also known as feature level sentiment analysis. Such level tells or discovers the opinion about a specific target. For each opinion, it finds a target instead of caring units of language such as documents, paragraphs or sentences. One of the major goals of aspect level is the identification of sentiment or opinion on entities and their distinct aspects [17].

Sentence Level: It is also known as phrase level sentiment analysis. This level is similar to document level having a major difference that in such a level, each of the sentence gets analysed on individual basis in order to look if it expresses positive, neutral or negative opinion thereby adding a great flexibility than the level of document as it becomes simple to distinguish the subjective and objective nature of sentences acting as a filter. However, it generalized that the sentences that are of objective nature expresses opinions and the subjective sentences do not transmit any kind of sentiment.

III. K-NEAREST NEIGHBOR

K-Nearest Neighbor represents the modest form of classification technique where it describes the distance among different data points and locates the data points that are not labeled. It is based on the some of the important conditions i.e. if ‘k’ denotes the value equal to one, then object gets simply assigned to its neighboring value. But if the value of ‘k’ is large then its prediction is very difficult in such case. K-NN representing an algorithm based on lazy learning so these are also known as the lazy learners. Lazy learner simply stores the training tuple and wait for test tuple. When test tuple arrived then it compares the testing tuple with training

tuple. The algorithm of K-NN indicates a very easy type of algorithm for the process of machine learning [16]. K-Nearest Neighbor also called as case-based reasoning, example-based reasoning, instance-based learning, memory-based reasoning. The performance of method based on K-NN classifier mainly depends over various distinct key factors, such as proper measure of distance, K parameter, and a measure of similarity [18]. Closeness is characterized by distance metric like Minkowski distance, Manhattan distance, Euclidean distance. KNN is widely used in the field of finance, medicine, agriculture, news and banking for problem solving, pattern recognition, functional learning.

IV. RELATED WORK

Kanika Sharma, *et.al* [1] compared the technique of wde-lstm with wde-knn classification. Both the existing and proposed techniques were implemented in python. It was analyzed that accuracy of wde-lstm technique was approx. 87.12 percent and when the technique wde-knn was applied it increased up to approx.94 percent. The performance of wde-lstm and wde-knn technique was also compared in terms of execution time. Execution time of the proposed method was less as compared to existing method. Naresh Sharma, *et.al* [2] presented an approach for sentiment classification using a classifier named K-NN classifier with the use of bag of words method taken as feature selector. The results obtained presents that K-NN classifier outperforms well for analyzing sentiments and provides a good high form of accuracy. The results have also shown that the classifier along with bag of words-based feature selection outperforms well dependent over polarity-based sentiment classification. Radhi Desai [4] performed Twitter based sentiment analysis. Twitter was considered to be an impressive point to the researchers in several areas such as democratic event prediction, movie box-office, popularity linked with celebrities. SA represents the opinion or feelings of one person to another domain type. The sentiment analysis classification and opinions based on the mechanism of polarity performs a major challenging target. The other challenges perform overwhelming information amount and all of these are expressed in distinct ways. Lots of work on sentiment analysis has to be done for sentiment analysis. Ahmad Ali, *et.al* [6]proposed techniques to classify the sentiment label accurately that were based on real tweets. Two methods were used for the research. First method was known as sentiment analysis method based on k-nearest neighbor and another was established on support vector machine. Paper focused on partition of reviews into positive and negative. The proposed approach concluded that sentiment analysis approach based on KNN performs better than SVM. B. Kaur, *et.al* [8] presented an approach which was used to identify the reviews. The combination of SVM and KNN was used, in which k-nearest neighbor classifier was working best for small length reviews. The approach was experimented on SuperFetch reviews. The approach concluded that combination of SVM and KNN produced better results on the basis of Recognition rate, Precision, TP rate, and F-Score. Onam Bharti, *et.al* [9] proposed an approach using KNN**,** Naïve Bayes, and the modified version of k-means clustering, and it found that the modified version is more accurate than the KNN techniques and Naïve Bayes individually. The researchers obtained classification accuracy of 91% on overall basis over the 500 mobile reviews of test-set. The algorithm running time is O (n + V log V) for the process of training where n represents the word number in a document and the V represents the vocabulary reduced size. It runs faster than the algorithms of Support Vector Machines and Naïve Bayes classification that takes more time in converging optimally in regard to set of weights. The level of accuracy was comparable to the existing algorithms used for the classification of sentiments based on reviewing mobile. Medhat, *et.al* [12] conferred about various applications of SA, recent modernized advancements in algorithm which were presented and investigated briefly in paper. Recently, the articles were reviewed gathering the reader's interest in technology offered by sentiment analyses. The algorithm based on Emotion detection was used for analyzing and enhancing emotions, it could either be implicit or explicit. Several types of algorithm were used for presenting the emotions and sentiments. Some of them are Point-wise Mutual information, Latent Semantic Indexing, Chi-square. The opinion-based techniques of classification were disjointed into hybrid, lexicon-based, and machine learning approaches.Disha Kohli, *et.al* [18] focused to analyze the expressed sentiments on Twitter demonetization such that the opinions of public and certain views were extracted, and analysed and further used to understand the positive and negative impact of such an impact on the Indian people. After analysing the sentiments of results, it was observed that many of the sentiments were of neutral type. The remaining tweets have shown that the positive type of sentiment remains over higher side.

The related work concluded that, K-nearest neighbor is a classification technique of supervised learning approach. A straight forward classification technique is KNN, where elements are arranged in the class of their nearest neighbor. Next section of paper describes the details related to methodology followed.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 1 pp. 889-894 Jan 2019-June 2019 www.csjournals.com

V. METHODOLOGY

The information is gathered for experiment and put away in database for pre-handling. This section of paper describes the description of steps of methodology followed to achieve results. The steps of methodology are shown in fig. 1.

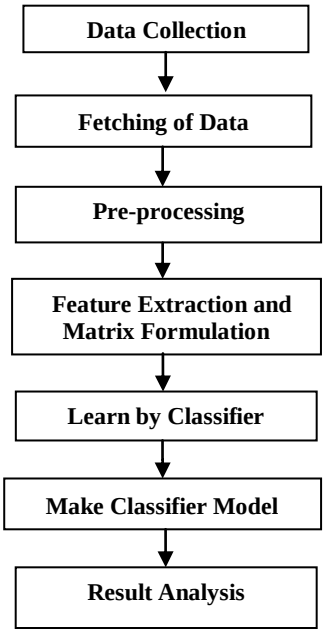

Fig. 1 Methodology followed

Step1: Data Collection
To conduct the experiment, the data which is given as input was taken from Movie Reviews. All the movie reviews have been scanned from www.imdb.com.IMDb stands for internet movie data base. It is the most authentic source for movie reviews and ratings. The data gathered from web is not complete and create uncertainty. Un-wanted data will be removed in the step of pre-handling.

Step 2: Fetching of data
The recovered data are put away as .csv arrangement records, and after that these documents are brought in the PyCharm instrument of Python. A large number of reviews are stored to prepare and test the datasets. Information mining calculation is utilized for preparation and experiment of the brought reviews.

Step 3: Pre-Handling/Processing of data
Pre-processing is an important phase of information mining. This process was done before the review-based usage of feature extractor in order to design or build the feature vector. Some pre-handling steps are required to separate the valuable data from the dataset. Such steps convert the plain tweet text into the elements of processing nature with additional information utilized by the feature extractor. The features are extracted from reviews in the next step.

Step 4: Feature Extraction and Classifier Modeling
Extracted feature are very large in number. It increased the storage time and computation time so the features are reduced and presented in the form of matrix of finite dimensions. Reduced features are then fed into classifier i.e. k-Nearest Neighbor. KNN has the capability to classify the reviews. Next process is the classifier modeling that performs the classification of sentiments.KNN learns to assign a label to the text and give a polarity score like +1 or -1, it should be classified a text as positive and negative.

Step 5: Optimization of result
In the last step, result will be analyzed on the basis of occurrences of classified reviews and the words with maximum rating are considered as most important words. To construct the model in Python, KNN is used for

the training and testing of the data. Cross validation is used for splitting the dataset. K-fold cross validation is applied because it is easy to use. Cross validation method first partitioned the dataset into k mutually exclusive subset, almost equal size. During ith iteration each subset is used as a test set and remaining subset is used as a training set. This technique makes efficient use of data and avoids the overlapping of test set.

VI. RESULTS

The proposed approach is implemented in python and results are analyzed in terms of accuracy. This section deals with the analysis of the outcomes acquired after experiment. The recommended method represents the comparative analysis of classifier K-Nearest Neighbor. Experimental analysis shows the overall Accuracy, Precision, Recall and F-measure.

Table 1 Accuracy of KNN

No. of Validation	KNN (Accuracy)
5-Fold	49.23
10-Fold	51.23

Table 1 shows the accuracy of KNN in different forms of validations. Accuracy shown by 5-Fold cross validation process is 49.23 and 10-Fold shows 51.23. Accuracy increases with the increase of number of validations. The overall accuracy is obtained by the averaging all accuracy obtained from each fold.

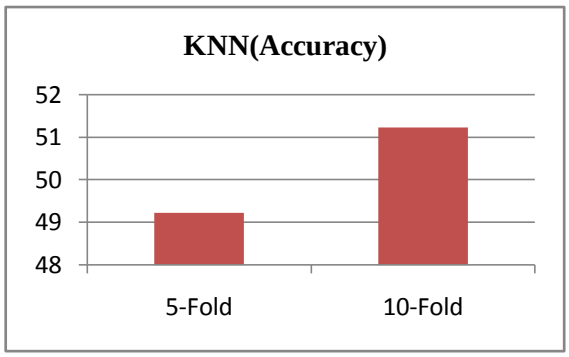

Fig. 2 Accuracy of KNN

Fig. 2 depicts the accuracy of the classifier that is KNN. The X-axis on graph shows the validation fold and Y-axis presents the values of accuracy. The algorithm represents the accuracy of used classifier in 5-Fold and 10-Fold. 49.23 is the minimum accuracy represented by KNN in 5-fold cross validation process.

Table 2 Precision after using KNN

No. of Validation	KNN (Precision)
5-Fold	50.23
10-Fold	51.23

Table 2 represents the precision after using KNN in distinct forms of cross validation folds i.e. 5-Fold and 10-Fold. Precision shown in 5-Fold validation is 50.23 and in 10-Fold is 51.23.

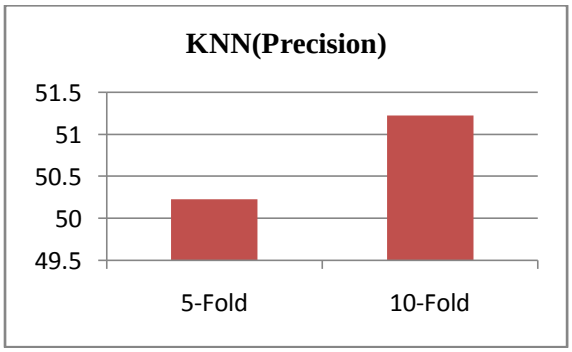

Fig. 3 Precision on different Folds

Fig. 3 depicts the Precision of the classifier that is KNN. The X-axis on graph stands for the validation fold and Y-axis stands for the values of Precision. The algorithm represents the precision in folds i.e. 5-Fold and 10-Fold of cross validation. 50.23 is the minimum precision represented by KNN in 5-fold cross validation process. There are various parameters which can be used for accuracy measure are given in Table 3.

Table 3 Parametric analysis of KNN

Classifier	Overall Accuracy	Precision	Recall	F-Measure
KNN	52.13	60.13	62.23	61.13

After experiment analysis Table 3 represents the parametric values of k-nearest neighbor. Value of overall accuracy is 52.13, precision is 60.13, recall is 62.23, and F-measure is 61.13.

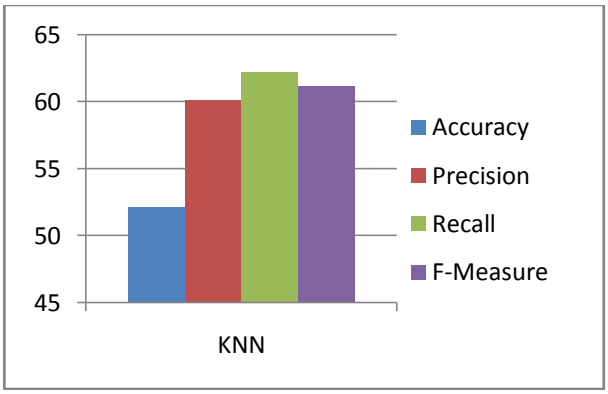

Fig. 4 Result on all Parameters

Fig. 4 shows the results on all parameters. The X-axis on graph correlates to the classifier and Y-axis corresponds to the values of different parameters like Accuracy, Precision, Recall, and F-measure.

VII. CONCLUSION

Classification of sentiments plays major part in the domain of research. Sentiment classification is not only the concept of text mining, it also includes the concept of information extraction. This paper discusses about widely used machine learning classification technique called k-nearest neighbor and experimental analysis was done to measure the accuracy, precision, recall and F-measure. More investigation is needed to increase the accuracy of KNN with good values of K. Performance of KNN can be further improved by combining it with other classifier and the hybrid approach can be used to make the sentiment analysis more effective.

REFERENCES

[1] Kanika Sharma and Akanksha Sambyal, "Sentiment Analysis Using Amazon Data For WDE-KNN Algorithm," *International Journal of Information and Computing Science*, vol. 6, no. 2, pp. 1-9, 2019.
[2] Tyagi Abhilasha, and Naresh Sharma, "Sentiments Analysis of Twitter Data using K-Nearest Neighbour Classifier," *International Journal of Engineering Science and Computing*, vol. 8, no. 4, pp. 17258-17260, 2018.
[3] Anjume Shakir, and Jyoti Arora, "Sentiment Analysis of Twitter Data using KNN Classification Technique," *International Journal for Scientific Research & Development*, vol. 6, no. 3, pp. 2040-2042, 2018.
[4] Desai Radhi, "Sentiment Analysis of Twitter Data: A Survey," International *Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 464-470, 2018.
[5] Ayman Mohamed Mostafa, "An Evaluation of Sentiment Analysis and Classification Algorithms for Arabic Textual Data," *International Journal of Computer Applications*, vol. 158, no. 3, pp. 29–36, 2017.
[6] M. Rezwanul Huq, Ahmad Ali, and Anika Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," *International Journal of Advanced Computer Science and Applications* , vol. 8, no. 6, pp.19-25, 2017.
[7] G. Sneka and CT. Vidhya, "Algorithms for Opinion Mining and Sentiment Analysis: An Overview," *International Journal of Advanced Research in Computer Science and Software Engineering*,vol. 6, no. 2, pp. 455–459, 2016.
[8] B. Kaur and N. Kumari, "SVM and KNN based Hybrid Approach to Sentiment Analysis," *International Journal of Technical Research & Science* , vol. 1, no. 5, pp. 67-74, 2016.
[9] Onam Bharti, Monika Malhotra, "Sentiment Analysis on Twitter Data," *International Journal of Computer Science and Mobile Computing*, vol. 5, no.6, pp. 601-609, 2016.
[10] L. Zahrotun, "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Computer Engineering and Applications Journal*, vol. 5, no. 1, pp. 11–18, 2016.
[11] Gurpreet Kaur, and Manoj Kumar, "A Review on Sentiment Analysis of Social Media Data Using Text Mining and Machine Learning," *International Journal of Advanced Research*, vol. 4, no. 5, pp. 772–775, 2016.
[12] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications : A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
[13] P. Kalaivani and D. K. L. Shunmuganathan, "Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches," *Journal of Computer Science and Engineering*, vol. 4, no. 4, pp. 285-292, 2013.
[14] S Padmaja and S. F. S, "Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey," *International Journal of Ad hoc, Sensor & Ubiquitous Computing*, vol. 4, no. 1, pp. 21–33, 2013.
[15] Kouloumpis Efthymios, Theresa Wilson, and Johanna Moore, "Twitter Sentiment Analysis: The good the bad and the omg!" in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, (*ICWSM*-2011), pp. 538-541, 2011.
[16] Agarwal Apoorv, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Language in Social Media*, pp. 30-38, Association for Computational Linguistics, 2011.
[17] Pabreja Kavita, "GST sentiment analysis using twitter data," *International Journal of Applied Research*, vol. 3, no. 7, pp. 660-662, 2017.
[18] Roy Kaustav, Disha Kohli, RakeshkumarKathirvel Senthil Kumar, Rupaksh Sahgal, and Wen-Bin Yu, "Sentiment Analysis of Twitter Data for Demonetization in India–A Text Mining Approach," *Issues in Information Systems*, vol. 18, no. 4, pp. 9-15, 2017.
[19] Suman, Jaswinder Singh, "Sentiment Analysis: A Survey," *International Journal for Research in Applied Science & Engineering Technology*, vol. 5, issue 8, pp. 1957-1963, 2017.
[20] Suman Rani, Jaswinder Singh, "Sentiment Analysis of Tweets using Support Vector Machine", *International Journal of Computer Science and Mobile Applications*, vol.5, issue 10, pp. 83-91, 2017.