Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 241-245
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Sentiment Analysis of Online Movies' Reviews Using Improved k-Nearest Neighbor Classifier

Tanya Arora[1], Sanjeev Dhawan[2] and Kulvinder Singh[3]

*[1]M.Tech.(Computer Engineering) Student, University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra, Haryana, India*
*[2]Department of Computer Science & Engineering, University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra, Haryana, India*
*[3]Department of Computer Science & Engineering, University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra, Haryana, India*
E-mail: [1]tanya.arora121@gmail.com, [2]rsdhawan@rediffmail.com, [3]kshanda@rediffmail.com

Abstract—*Others' opinion can be decisive while choosing among various options; especially when those choices involve worthy resources like spending time and money buying products or services. A significant reason behind information gathering of opinions has always been there to figure out what other people think. Until a decade ago, the only sources of information were family, friends or acquaintances. But now, electronic word-of-mouth has become a flourishing frontier. Customers relying on their peers' past reviews on e-commerce websites or social media have drawn a considerable interest to sentiment analysis due to the realization of its commercial and business benefits. In this paper, sentiment analysis of movie reviews has been carried out to ascertain the sentiment behind the review- whether positive or negative and an improved k- Nearest Neighbor (ImpkNN) classifier has been designed which uses the concept of attribute weighted-kNN and the weights associated are trained using cross validation. At last, the results of both Basic kNN and ImpkNN are evaluated using graphs.*

1. INTRODUCTION

In this internet era, online data is getting manifold each and every second. However, the gigantic information available on web has sometimes become a vigorous issue for the users. Due to the increasing diversity of information, it has become quite cumbersome for the users to buy a relevant product or service. To overcome this, many websites including e-commerce and social media have facilitated users with online discussion forums or online reviews section. Reviews expressed electronically comprising opinion or sentiment of the people across areas such as buying products or rating movies or events. Hence, e-opinions given by numerous users on shopping websites and social media have lately known as the hotspots which can be analyzed to judge sentiments. Sentiment analysis is a cognitive process of ascertaining peoples' opinion or attitude towards object, events and their attributes. A lot of research has been done to detect the polarity of the sentiment using text mining techniques-classification and clustering. Text classification requires supervised learning while in clustering, unsupervised learning method is followed. In this paper, text classification approach is adopted to characterize the polarity of an opinion in a movie dataset which consists of movie reviews given by the users. The Java-ML (Java Machine Learning Library) tool is used for the purpose of data classification. In the present research, an improved k-Nearest neighbor (ImpkNN) classifier is presented and then comparisons have been made between Basic kNN and ImpkNN by analyzing various features like accuracy, precision, recall, F-measure and execution time. Besides introduction, this paper consists of five sections. The section II surveys the previous researches done in this area, section III describes the proposed work, section IV depicts the experimental setup and at last section V concludes the discussion.

2. RELATED WORK

This section exemplifies the researches that have been done till now in the context of sentiment analysis to detect emotional polarity of texts using different text mining techniques including natural language processing, text classification and text clustering. Using movie reviews as data; three machine learning techniques-Naïve Bayes, maximum entropy classification and Support vector Machine (SVM) were employed by Pang and Lee [1] and discovered the factors that make sentiment classification more challenging than topic based categorization. An unsupervised learning algorithm was designed by Turney [2] for classifying reviews of four different domains (reviews of automobiles, books, movies and travel destinations). A data-driven method was introduced by Barzilay and Lapata [3] for learning the content-selection component for a concept-to-text generation system which determined which part of the information should be the output of a natural language generation system. The method presented by Chunling Ma [4] for providing emotional estimations for natural-language texts which was based on a keyword spotting technique, i.e. the system divides a text into words and performs an emotional estimation for each of these words, as well as a sentence-level processing technique. A new approach for sentiment analysis of phrases was described by Theresa, Janyce and Paul [5] that worked in two steps: firstly, it determines whether an expression is factual or carries any sentiments and then clarifies their polarity. For movie review mining, a multi-knowledge path [6] was adopted which integrated WordNet, statistical analysis and movie knowledge; to automatically generate a feature class based summary for arbitrary online movie reviews. While text classification, kNN[7] is the first classifier that was invented. However, because of its characteristic of postponing decision-making without pre-modeling, kNN [7, 8] is not cost efficient to categorize new documents when training set is large. Rocchio algorithm [9] is also a well-established and extensively applied classifier but has the limitation of restricting the hypothesis space to the set of linear separable hyper plane regions [9]. A hybrid algorithm [10] was developed based on inconsistent precision rough set to amalgamate the efficiency of both kNN and Rocchio classifiers to overcome their weaknesses. In an another research, a reassessment of the kNN model was approached by Cucala, Marin, Robert and Titterington [11] to conceive a framework deduced from a proper probabilistic model for directing Bayesian inference on the values of the conforming model. To analyze online sports forum for their hotspot detection, a combined approach of k-means clustering and SVM classification was followed by Nan and Desheng [12]. A decision support model (DSM) was made by Yue Da [13] that leveraged various text mining technologies and SWOT (Strength, Weakness, Opportunity, Threat) analysis to search and analyze the unstructured textual data from various online reviews. Na, Thet and Khoo [14] conducted sentiment analysis of consumers' reviews by collecting reviews from four different types of movies to detect sentiments judging their lingual facets such as vocabulary, length of sentences and categories of words in grammatical context. Some research was also carried out to introduce a technique [15] to collect a corpus of documents automatically which could be used in training of a classifier to investigate sentiments associated. The classifier formed by Pak and Paroubek [15] is derived from the multinomial Naive Bayes classifier that uses N-gram and POS-tags (Part-Of-Speech tags) as fundamental features. A group of three classifiers comprising two statistical classifiers (Naïve Bayes and Maximum Entropy learner) and a knowledge based tool has also been introduced in the recent research [16]. Subsequently, inferences made from the discussed researches have encouraged us to develop a new classifier for sentiment analysis of movie reviews that is an improvement of previous kNN classifier.

3. PROPOSED WORK

An outline of the steps and techniques followed for sentiment analysis are depicted in Figure1. At first movie reviews dataset files are read thoroughly, afterwards the further process begins as shown below:

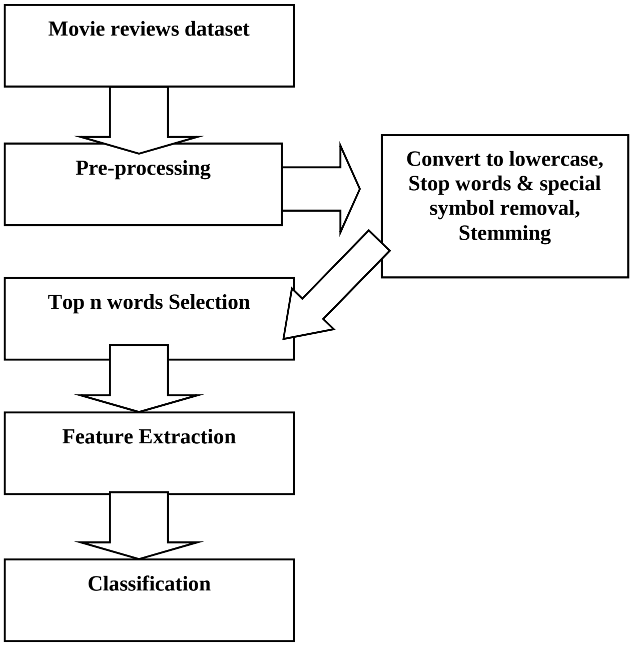


Fig. 1: Proposed Framework for Sentiment analysis

3.1 Pre-processing

Text pre-processing is defined as the task of cleaning the unwanted data from the whole and preparing the relevant data for classification. Assorted pre-processing methods are-

3.1.1 Convert to Lowercase:

This is done to avoid make out between words simply on case.

3.1.2 Stop words Removal:

Stop words are common words found in English language like for, of, are, is, and, or etc. These words are of no use while sentiment analysis, thus should be removed.

3.1.3 Symbols, Numbers and Punctuation Removal:

Special symbols and numbers are not pertinent while observing sentiments. Although Punctuation can provide grammatical context which supports understanding, yet it is irrelevant in determining sentiments.

3.1.4 Apply Stemming:

Stemming is applied to transform any word to its root word or dictionary based form by removing verbs, 'ing', 's' or 'es', 'ed' and many other common endings. Porter algorithm is used for stemming in this paper.

3.2 Top n words selection

After pre-processing, words are selected from data according to their recurrence or count. A threshold cut of (n>i), where "i" is a natural number; is applied to select top n-words that are recurring frequently. These frequent words amount to the interesting features that need to be analysed and further leads to more accurate results while polarizing sentiments.

3.3 Feature Extraction

Feature extraction is used to decrease the dimensional reduction of the feature space [14]. The basic approach of statistical feature extraction is to use recurring words in the corpus as feature values. A vector space model (VSM) represents the words and their frequency score in a form of matrix (rows contain words and columns contain their corresponding terms (weights); any greater than zero entry indicates the presence of the word. TF-IDF (Term Frequency and Inverse Document Frequency) weighting is simply used for calculating weight for each word. TF and IDF are defined as below:

TF (w) = (Count of word "w" occur in document/Number of words in the document)

IDF (w) =log (total count of documents/ Count of documents with word "w")

3.5 Classification

Text classification is defined as the process of assigning documents to suitable pre-defined classes/categories. Classifiers like kNN, SVM, Naïve Bayes and many more can be used for this purpose. The oldest and the simplest classifier used for text classifying is kNN [7], which is an illustration-based method that delays the decision to extrapolate outside the training instances until another query comes across [8]. Whenever there is a new instance to classify, its k-nearest neighbours are explored by selecting a reasonable distance measure such as Euclidean distance. Then, the class of these k instances is inspected to choose an apt class C that represents the most of instances. However, there is a curse of dimensionality in basic kNN as the similarity metrics do not take into account the relation of attributes which results in inaccurate distance and affects the classification precision. Thus, to subdue this problem, ImpkNN is purported that is much more efficient than the basic kNN.

3.5.1 ImpkNN:

Classification using ImpkNN uses the concept of attribute (word) selection and weighting. As it has been usually seen that some attributes are more relevant than others while classification. Filtering approach is deployed for choosing relevant attributes by fixing a threshold value. A method in which those relevant attributes are given more importance than the rest is known as attribute weighting. Normalized weighting [17], a type of attribute weighting, normalizes weights using standard deviation. Using this measure, random weights are assigned to each attribute which are then trained by cross validation.

3.5.2 Cross validation:

It is done as the predictive evaluation of the classifier in which each record is used for the same number of times for training but only once for testing.

4. EXPERIMENTAL SETUP

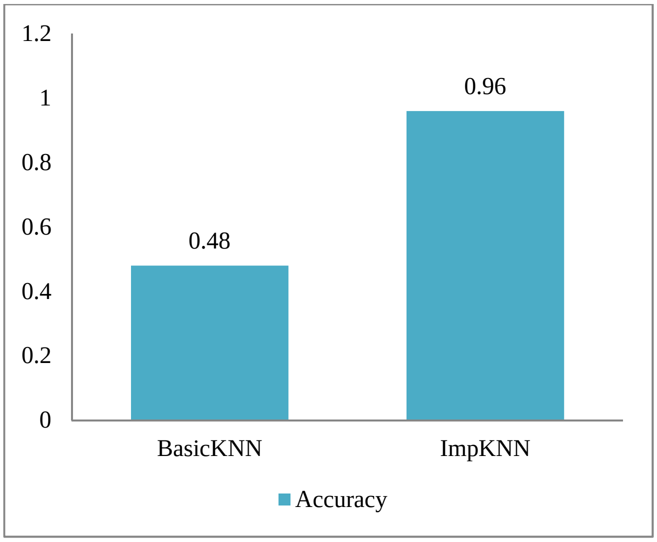
4.1 Dataset

The dataset used for analysing sentiments is Cornell movie-review corpora, containing users' reviews about movies which are further polarized using proposed sentiment analysis method. There are 100000 entries of users' reviews about movies. This dataset is obtained from http://www.cs.cornell.edu/people/pabo/movie-review-data/ introduced in [2].

4.2 Evaluation (Basic kNN versus ImpkNN)

4.2.1 Accuracy:

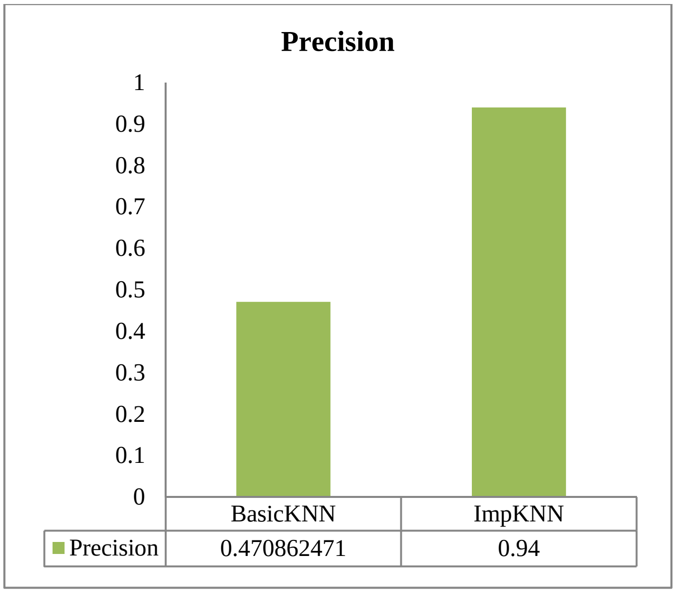
In sentiment analysis, it is the ratio of rightly classified words to the total count of words. Comparison between accuracy in results of Basic kNN and ImpkNN is shown in Graph 1.



Graph 1: Basic kNN versus ImpkNN

4.2.2 Precision:

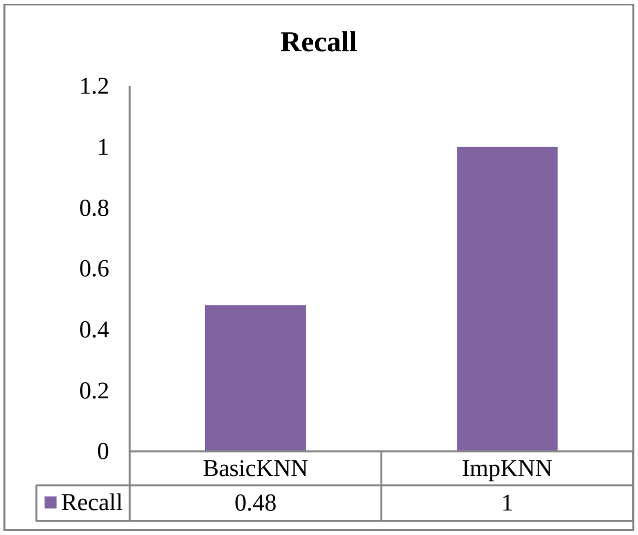
It is the count of words rightly labelled as positive or negative multiplied by the inverse of total number of times that words are classified as positive or negative. Precision in results given by Basic kNN and ImpkNN is depicted in Graph 2.



Graph 2: Basic kNN versus ImpkNN

4.2.3 Recall:

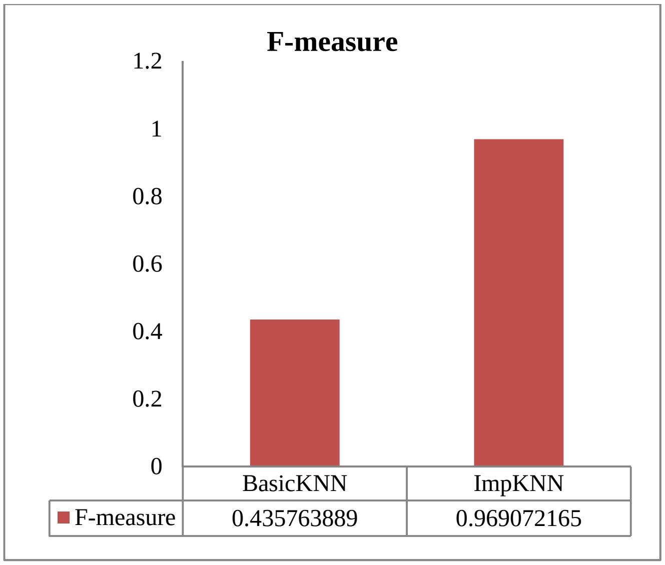
It is described as the count of words aptly labelled as positive divided by the total count of words that are truly positive. Recall calculated using Basic kNN and ImpkNN is shown in Graph 3.



Graph 3: Basic kNN versus ImpkNN

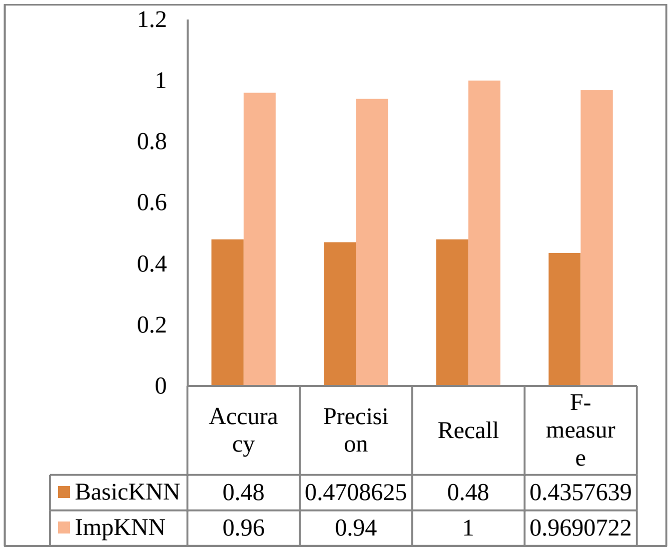
4.2.4 F-measure:

It is given by the harmonic mean of precision and recall. Graph 4 compares F-measure of Basic kNN and ImpkNN.



Graph 4: Basic kNN versus ImpkNN

Evaluation of Basic kNN and ImpkNN is shown in Graph 5 as below:



Graph 5: Basic kNN versus ImpkNN

5. CONCLUSION

In this research, ImpkNN classifier is used for sentiment analysis of movie reviews dataset which outperforms the Basic kNN as observed from Graph5. Sentiment analysis of movie reviews examines the polarization of opinions expressed by the users which can be used for many business and commercial purposes. In future, classifiers like SVM, Naïve Bayes, Genetic algorithm and fuzzy classification can also be applied on movie reviews dataset to find out the most efficient classifier as compared to ImpkNN. The classification can be done not only on movie reviews but also on many other products' or services' reviews available on social media and e-commerce websites.

REFERENCES

[1] B.Pang, L.Lillian and S.Vaithyanathan,'Thumbs up? Sentiment classification using machine learning techniques',in Proceeding of the conference on empirical methods in natural languge processing, pp.79-86, 2002.

[2] P.Turney, 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews', in Proceedings of the 40th annual meeting of the association for computational linguistics(ACL'02), Philadelphia, Pennysylvania, USA, pp. 417-424, July 2002.

[3] R.Barzilay and M.Lapata, 'Collective content selection for concept-to-text generation',in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Processing (HLT/EMNLP), Vancouver, pp. 331-338, October 2005.

[4] C.Ma, P.Helmut and L.Mitsuru, 'Emotion estimation and reasoning based on affective textual interaction', 3rd edition, Springer, 2005.

[5] T.Wilson, J.Weibe and P.Hoffmann, 'Recognizing contexual polarity in phrase level sentiment analysis', in Proceedings of human language technology conference and conference on empirical methods in natural language processing(HLT/EMNLP), Vancouver, pp. 347-354, October 2005.

[6] L.Zhuang, F.Jing and X.Jhu, 'Movie review mining and summarization', ACM, CIKM'06, Airlington, Virginia, USA,November 2006.

[7] T.Cover and P.Hart, 'Nearest neighbour pattern classification', IEEE Transaction on Information Theory, 13(1), pp. 21–27, 1967.

[8] F.Sebastiani, 'Machine learning in automated text categorization', ACM Computing Surveys, 34(1), pp. 1–47, 2002.

[9] T.Joachims, 'A probabilistic analysis of the Rochhio algorithm with TFIDF for text categorization', In Proceedings of the fourteenth international conference on machine learning, 1997.

[10] D.Miao, Q.Duan, H.Zhang and N.Jiao, 'Rough set based hybrid algorithm for text classification', Journal of Expert Systems with Applications, 36(5), pp. 9168-9174, 2009.

[11] L.Cucala, J.Marin, C.Robert and D.Titterington, 'A Bayesian reassessment of nearest neighbour classification', arXiv:0802.1357v1 [stat.CO], 10 Feb 2008.

[12] N.Li and D.Wu, 'Using text mining and sentiment analysis for online forums hotspot detection', in Decision Support System, Elsevier, September 2009.

[13] Y.Dai, T.Kakkonen and E.Sutinen, 'A decision-support model that combines text-mining technologies with two competitive intelligence analysis method', International Journal of Computer Information System and Industrial Management Applications, vol.3, pp. 165-173, 2011.

[14] O.Kummer and J.Savoy, 'Feature selection in sentiment analysis', Coria 2012, Bordeaux, pp. 273-284, March 2012.

[15] A.Pak and P.Paroubek, 'Twitter as a corpus for Sentiment Analysis and Opinion mining', Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 Orsay Cedex, France, 2012.

[16] I.Perikos and I.Hatzilygeroudis, 'Recognizing emotion in text using ensemble of classifiers', Engineering Applications of Artificial Intelligence, Elsevier, 2016.

[17] M.Syed, K.Iltanen and M.Juhola, 'Attribute weighting in k-nearest neighbor classification', University of Tampere, November 2014.