Sentiment Analysis Based on Data Mining and Natural Language Processing

Deepak Kumar Yadav
Department of Computer Applications, SSTC, SSGI, FET Bhilai, India

Sampada Vishwas Massey
Department of Computer Science & Engineering, SSTC, SSGI, FET Bhilai, India

***Abstract*— Sentiment analysis refers to a text classification that analyzes the text which are oriented from opinions called opinion mining. Sentiments can be determining on different types of levels. For instance, human sentiments can be positive, negative Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. Now day's users use internet to share their suggestions, reviews which help the other user in making decision. In this paper we have used porter stemming algorithm for removal of stop words and a parser named Stanford for the grammatical structure and the KNN algorithm**

***Index Terms*—computational linguistics, Data Mining, Natural language processing, analysis, twitter social media.**

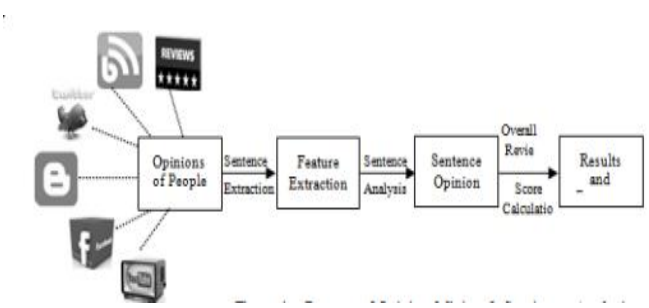
I. INTRODUCTION

Now days on web the number of reviews, suggestions, feedbacks are increasing in enormously manner. Because every person wants to share his views and experience about the product like review on product, review on movie, Person tweets etc. Reviews play vital role in helping and suggesting other person in their decision making. But on the other hand it becomes difficult to read all reviews and make decision as per.

Thus, mining this data, identifying the user opinions this is done by performing detailed sentiment analysis on the data. The fields of opinion mining and sentiment analysis are distinct but deeply related. Opinion mining focuses on polarity detection [positive, negative or neutral] whereas sentiment analysis involves emotion recognition. Because detecting the polarity of text is often a step in sentiment analysis, the two fields are usually combined under the same umbrella.

The fig:1 shows the process of opinion mining and sentiment analysis

Fig:1 Process of opinion mining and sentiment analysis



This methodology provides the summary of total number of positive and negative document which help the users in their decision making. We took one review of customer from amazon.in. In which he is giving his suggestions and sharing his experience about the camera he purchased from the amazon.in. The review is a combination of some critic sentence and positive sentence.

ISSN: 2278 – 1323

Fig: 2 presents an example of document based opinion mining.

INPUT DOCUMENT

Awesome camera!!

Excellent picture quality and both the lenses are awesome. Must buy if you are serious about photography. Invest some more money and buy this camera. The battery life lasts for about 600 shots as promised. I am using this camera from almost a month now. Only downside is that the bag provided is local made and of cheap quality and offers little to no protection to the body and lens and you have to cramp the camera if you want to keep both the lenses and the body. [3]

OUTPUT DOCUMENT

POSITIVE DOCUMENT-

Excellent picture quality and both the lenses are awesome. Must buy if you are serious about photography.
Invest some more money and buy this camera. The battery life lasts for about 600 shots as promised. I am using this camera from almost a month now.

NEGATIVE DOCUMENT

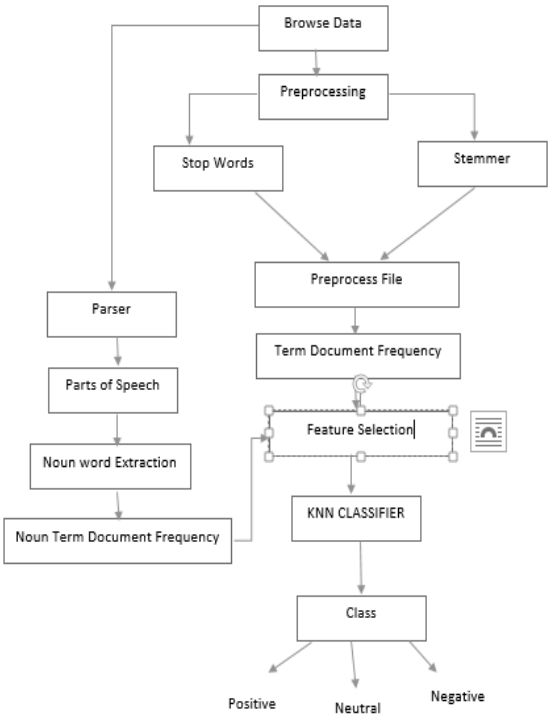
Only downside is that the bag provided is local made and of cheap quality and offers little to no protection to the body and lens and you have to cramp the camera if you want to keep both the lenses and the body.

III. WORK FLOW OF SYSTEM

A. 3.1 Steps use for opinion mining [3] here are as follows:

1. Gathered the relevant comments related to product. This is done by surfing the internet finding reviews on twitter, facebook related to product then crawl it down and picks relevant sentences.
2. Find the opinion word and its semantic orientation. In this step, words will be saved in database and then which words are negative and which ones are positive is to be found.
3. Check link between product characteristics and its semantic orientation. In this step, what is the feature and related to that what is the semantic orientation of the word is to be found.
4. Find strength of word and depending on this overall view on product is made. In this step, the value of negativity and positivity of the words is calculated then these words are analyse and over all perception about the product is made.
5. For this the classifier is used which will help in classifying the reviews. So that the overall view if the product is made.

Fig.3: Workflow of system



B. Work flow in system

1. Browse Data – Data (.txt file) will be browsed from system. The text file will contain reviews. This will be shown in the text area
2. Pre-Processing– This module is used to pre-process the review document by identifying the relevant portion of a text document and by removing the stop word. For this we have used Poter Stemmer algorithm.
3. Parser– This module is used to extract product features from text documents. All subjective sentences are parsed using Stanford Parser, which assigns Parts-Of-Speech (POS) tags on the context in which they appear. The typed dependency diagram given by the parser will be used for the extraction of the features after applying the rules as shown in the next section.
4. Term document Frequency– This will give the total frequency of unique words and unique nouns after the pre-processing.
5. Feature Selection-1– This feature will give total number of words and total number of unique words.
6. Feature Selection-2– This feature will give total number of nouns and total number of unique nouns.
7. Classifier– For classifying the review we have used KNN algorithm. After the evaluation positive, negative or neutral sentiments can be determined.

ISSN: 2278 – 1323

IV. METHODOLOGY

A. BOW representation:

In the first approach, we use the commonly used BOW method as the feature set. In this approach, considering all the documents in the corpus, a vocabulary list is constructed and each document is represented with a vector indicating the existence of a term in the document. There are different methods to weigh each term in the BOW representation such as binary, term occurrence and term frequency-inverse document frequency (tf-idf) [4].

In binary weighting, if the term presents in the document, the weight is 1 and if it doesn't present in the document, its weight is 0. In term occurrence scheme, the weight of each term is equal to the number of times it is appeared in the document. In this paper we have computed the tf-idf as follows:

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{1+\max_{w\in d}\mathrm{f}(w,d)} \quad (1)$$

$$\mathrm{idf}(t,D) = \log\frac{|D|}{|\{d\in D : t\in d\}|} \quad (2)$$

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d)\times\mathrm{idf}(t,D) \quad (3)$$

In equation (1) the frequency of term t in document d (tf (t, d)) is normalized by maximum frequency of terms in that document. |D| is the number of all documents in the corpus in equation (2).

B. Score Representation

Score representation: In score based representation three scores are computed for each term (ti) in our vocabulary list: positive score (s^+i), neutral score (s^0i) and negative score (s^- i). These scores are computed as:

$$s_i^+ = \frac{f_i^+}{f_i^+ + f_i^0 + f_i^-},$$
$$s_i^0 = \frac{f_i^0}{f_i^+ + f_i^0 + f_i^-}, \quad (4)$$
$$s_i^- = \frac{f_i^-}{f_i^+ + f_i^0 + f_i^-}$$

where f^+i , f^0i , f^- i are the frequencies of term ti in positive, neutral and negative documents respectively. Using these scores, we compute the positiveness, neutralness and negativeness of each sentence (x) as:

$$S^+ = \sum_{i\in \mathbf{x}} w_i s_i^+,$$
$$S^0 = \sum_{i\in \mathbf{x}} w_i s_i^0, \quad (5)$$
$$S^- = \sum_{i\in \mathbf{x}} w_i s_i^-$$

where x contains all the terms in a sentence and wi could be either of binary, term occurrence or tf-idf weights in the BOW representation of the sentence. Now each sentence is represented as a 3-dim vector S as follows:

$$\mathbf{S} = [S^+, S^0, S^-]^T \quad (6)$$

In emotion recognition literature the authors of [5] have computed six scores for each term based upon the provided scores of SentiWord Net [6]. It is worth noting that the three scores that we compute for each term in our vocabulary list are not some arbitrary scores that we just assign to each one of them. These scores are actually learned from the existing data (without using any external lexical resource) and reflect the positivity, neutrality and negativity of terms in the related content.

Fig: 4. Parsing

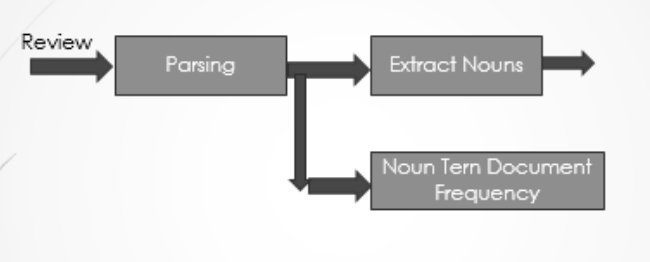
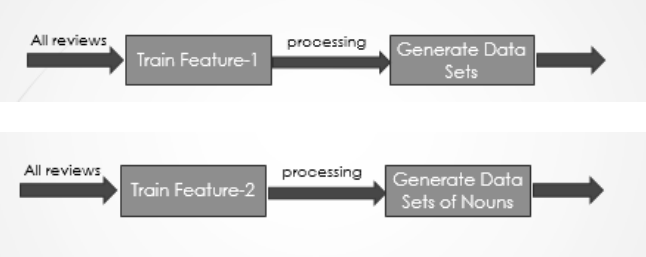


Fig: 5 Processing data sets



V. PROPOSED SYSTEM

Large numbers of user reviews are collected from different-different websites. User reviews contains critic and reviews, there are various websites available on the web which contain user reviews like www.amazon.in[7], www.flipkart.com [8] etc.

An overview of steps and techniques commonly used in sentiment classification approaches, as shown in Figure 3. Part of speech model in which a document is represented as a vector, whose entries correspond to individual terms of a s.

.

A. 5.1. Porter Stemmer algorithm

This is particularly used for the removal of stop words. This algorithm will reduce the English input words or suffixes to its basic stem for e.g. (running to run) so that whatever the variations on a word like (run, ran, running) are considered equivalent during search. The foremost use of this stemming in keyword indexing for search. In the proposed system List of suffix has explicitly defined and with each suffix, the

ISSN: 2278 – 1323

criterion under which it may be removed from a word to leave a valid stem.

B. 5.2 Stanford NLP parser

A natural language parser is a program that accomplish the grammatical structure of sentences for instance, which groups of words become (as "phrases") and which words are the subject or object of a verb.

For this application, we utilized the Stanford Parser [9], which is also a statistical parser with a high accuracy rate, and written in Java itself. The parser provides Stanford Dependencies [10] output as well as phrase structure trees.

5.2.1 Steps in parsing

Break a review into individual sentences 'S'.

Where S = {S_1, S_2, S_3... S_n} for each sentence 'S_i' □S parse and tag the sentence into its linguistic tags corresponding to each token, as well as generating the dependency relations existing in the parse tree.

Let the set of tags generated for each statement S_i□S be T = {T_1, T_2, T_3… T_n}

Once all the dependencies are explored we need to take into account the relevant dependencies only and ignore others. Let the set of dependencies be represented by 'D' and the relevant dependencies by R_D

Where D = {D_1, D_2, D_3… D_n} and Di = {W_1, W_2} where W_1 and W_2 are words dependent on each other.

Fig.6: List of relevant Dependencies

neg: negation modifier

nsubj: nominal subject

amod: adjective modifier

rcmod: relative clause

dobj: direct object

C. 5.3 K-Nearest Neighbor algorithm

This algorithm takes new point{x} and classifies it according to a majority vote of the k-nearest points in the dataset.

The algorithm can be summarized as:

1. A positive integer k is specified, along with a new sample.
2. k entries will be selected from our database which are next to the new sample.
3. Find the most common classification of these entries.
4. This is the classification we give to the new sample.

D. 5.4 Feature based sentiment classification

Sentiment Analysis. Converting a piece of text to a feature vector is the basic step in any data driven approach to Sentiment Analysis. It is important to convert a piece of text into a feature vector, so as to process text in a much efficient manner. In text domain, effective feature selection is a must in order to make the learning task effective and accurate. In text classification, with the bag of words model, each position in the input feature vector corresponds to a given word or phrase. In the bag of words framework, the documents are often converted into vectors based on predefined feature presentation including feature type and features weighting mechanism, which is critical to classification accuracy. The major feature types contain unigrams, bigrams and the mixtures of them, etc. The features weighting mechanism mainly includes presence, frequency, tf*idf and its variants [11]. The commonly used features used in Sentiment Analysis and their critiques [12] are Term Presence, Term frequency, term position, Subsequence Kernels, Parts of Speech, Adjective-Adverb Combination, Adjectives, n-gram features etc.

5.4.1 Feature Extraction-Let us consider the n-gram features for feature extraction. An n-gram is a contiguous sequence of n items from a given sequence of text or speech. An n-gram could be any combination of letters [49] (syllables, letters, word, part-of speech (POS), character, syntactic, and semantic n-grams). The n-grams typically are collected from a text or speech corpus and n-gram features captures sentiment cues in text. Fixed n-grams are exact sequences. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. n-gram features can be classified into two categories: 1) Fixed n-grams are sequences occurring at either the character or token level. 2) Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining [13]. Documents are often converted into vectors according to predefined features together with weighting mechanisms [14]. Correlation is a commonly used method for feature selection [15], [16]. The process of obtaining n–gram can be given as in the steps below, 1) Filtering – removing URL Links 2) Tokenization – Segmenting text by splitting it by spaces and punctuation marks, and forming bag of words 3) Removing Stop Words – Removing articles("a", "an", "the") 4) Constructing n-grams – from consecutive words

VII. CONCLUSION

This paper illustrates the research area of Sentiment Analysis and its latest advances. It affirms the terminology, the major tasks, the granularity levels, and applications of sentiment analysis. Most work has been done on product reviews – documents that have a definite topic. In this work views from amazon.in, flipkart.com online shopping sites has been collected. These views are both structured and unstructured. So for finding the strength of opinions expressed by reviewers on object, data is pre-processed and the relevant words are taken from the comments. Words are illustrate as positive and negative by giving negative sign in front of value of negative word.

ISSN: 2278 – 1323

REFERENCES

[1] Erik Cambria, National University of Singapore Bjo_rnSchuller, Technical University of Munich Yunqing Xia,Tsinghua University Catherine Havasi, Massachusetts Institute of Technology, "New Avenues in Opinion Mining and Sentiment Analysis "Published in Intelligent Systems, IEEE (Volume: 28, Issue: 2) [ISSN: 1541-1672] pp 15-21March/April 2013.
[2] Amazon.in
[3] Ritesh Srivastava and M.P.S Bhatia, "Quantifying Modified Opinion Strength: A Fuzzy Inference System for Sentiment Analysis",International Conference on Advance Computing, Communications and Informatics (ICACCI), 2013.
[4] C. D. Manning, P. Raghavan, and H. Sch¨utze, Introductionto information retrieval. Cambridge University Press Cambridge,2008, vol. 1.
[5] D. Das and S. Bandyopadhyay, "Sentence-level emotion andvalence tagging," Cognitive Computation, vol. 4, no. 4, pp.420–435, 2012.
[6] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in In Proceedings of the 5th Conference on Language Resources and Evaluation,2006, pp. 417–422.
[7] Amazon.com
[8] Flipkart.com
[9] Richard Socher, John Bauer, Christopher D. Manning and Andrew Y.Ng. 2013. Parsing With Compositional Vector Grammars, in theProceedings of ACL 2013.
[10] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning "Generating Typed Dependency Parses from Phrase Structure Parses". In LREC 2006.
[11] Lisa Hankin, "The effects of user reviews on online purchasing behavior across multiple product categories", Master's final project report, UC Berkeley School of Information, 2007.
[12] George Forman, "An Extensive Empirical study of feature selection Metrics for Text Classification", Journal of Machine Learning Research, Vol. 3, pp. 1289-1305, 2003.
[13] Michael Wiegand and Alexandra Balahur, "A Survey on the Role of Negation in Sentiment Analysis", Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, 2010.
[14] Yuming Lin, Jingwei Zhang, Xiaoling Wang and Aoying Zhou, "Sentiment Classification via Integrating Multiple Feature Presentations", WWW 2012 – Poster Presentation, pp. 569-570, 2012.
[15] M. Hall and L.A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach", Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855- 858, 1997.
[16] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research, Vol. 3, pp. 1289-1305, 2004.

First Author I. Deepak Kumar Yadav is pursuing M.E. (Master of Engineering) in Computer Technology & Applications from SSTC, SSGI, FET Bhiai Chhattisgarh Swami Vivekanada University ,Bhilai,, India and has done his MCA (with Honors) from BIT, Durg, India. Also working as an Assistant Professor in Department of Computer Application SSTC, SSGI, FET Bhilai. His interest area is Data Mining and Knowledge Management Process, Information and data security.

Second Author II. Sampada Vishwas Massey has done MTech. In Computer Science from Chhattisgarh Swami Vivekanada University ,Bhilai,, India and has done her BE in Computer Science & Engineering from CSVTU, Bhilai, India. Also working as an Assistant Professor in Department of Computer Science & Engineering SSTC, SSGI, FET Bhilai. Her interest area is Image Processing and Neural Network.

ISSN: 2278 – 1323