International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey on Discrimination Analysis and Sentimental Analysis in Text Mining by using NLP Method

Bhagyashri Sawana[1], P. K. Bharne[2]

[1]Department of Computer Science, S.S.G.M.C.E Shegaon

[2]Professor, Department of Computer Science, S.S.G.M.C.E Shegaon

Abstract: ***Discrimination Analysis is the prejudicial treatment which involves denying opportunities to members of one group in favor of other groups. It is unfair to discriminate people because of their gender, religion, nationality, age and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. the training data without harming their decision-making utility is therefore highly desirable which forms the primary goal of anti-discrimination techniques in data mining. The evolution of web technology, there is a huge amount of data present in the web for the internet users. These users not only use the available resources in the web, but also give their feedback, thus generating additional useful information. Due to overwhelming amount of user's opinions, views, feedback and suggestions available through the web resources, it's very much essential to explore, analyze and organize their views for better decision making. In this paper we are survey Opinion Mining or Sentiment Analysis is a Natural Language Processing and Information Extraction task that identifies the user's views or opinions explained in the form of positive, negative or neutral.***

Keyword: Discrimination, Sentimental SVM,NLP etc

1. Introduction

Data mining is an increasingly important technology. It is a process of extracting useful knowledge from large collections of data. There are some negative view about data mining, among which potential privacy and potential discrimination. Discrimination means is the unequal or unfairly treating people on the basis of their specific belonging group. In other words, discrimination means treating people differently, negatively or adversely without a good reason. Opinion Miner system designed in this work aims to mine customer reviews of a product and extract high detailed product entities on which reviewers express their opinions. Opinion expressions are identified and opinion orientations for each recognized product entity are classified as positive or negative. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. The purpose of the analysis is to extract, organize, and classify the information contained in the required documents. The main highlights are Stanford parser and POS tagging where different tags are used to identify the different parts of speech in the opinion sentences that are parsed. Here the document is applied to formal and informal classifiers.

2. Literature Survey

J.Domingo-Ferrer et al. (2011) have developed a paper for rule protection for the indirect discrimination prevention in data mining. The datasets are trained and developed to make the classification rules to be extracted. Indirect discrimination rules cannot be extracted from the trained dataset. (i.e.) the trained datasets are free from indirect discrimination. Datasets are modified if any indirect discrimination occurs. Standard data mining algorithms are used to prevent the indirect discrimination from the training dataset.

S. Shivashankar. (2010) have developed a paper for discrimination aware decision tree learning. The decision tree models leads to the lower discrimination than the other models but with a little loss in the accuracy. The decision tree models are effective at removing the discrimination from the original datasets. The problem is the datasets are cleaned away for discrimination before the discovery of the classifier in the dataset[2].

Sara Hajian et al. (2011) have developed a paper for prevention of discrimination in data mining for intrusion and crime detection. Data mining algorithm are used to prevent the direct and indirect discrimination. The data set obtained is free from the discrimination. In addition to detect the discrimination intrusion fraud and crime is also detected in the given dataset[1].

3. Related Work

Proposed Methodology

The implementation part is mainly explained to prevent the discrimination in the dataset and to maintain the data quality for the given dataset.

A. Data Analysis

In this Methodology Data analysis is to gather the data from the external disk. Dataset contains real life dataset and synthetic dataset. First of all we have to check if all the

attributes are placed in a correct manner if any null values are present then those dataset attributes cannot be processed by the metric and other computation process. Data analysis is generally termed as the process of gathering and analysis of dataset individually in a given two dataset.

B. Utility Measures

In this Methodology Utility measures are used to remove the discrimination on the given dataset. Dataset are analyzed with certain measures to remove the discrimination from the specified data's. Indirect discrimination removal and measuring data quality of that process are computed by the mathematical functionality like metric and rule protection and rule generalization. With the use of these techniques and algorithm records are filter in short time.

C. Decision Making

In this Methodology Decisions could be depend on the attributes like gender, race and religion and so on. Each user gets score for their personal attributes in direct discrimination.

D Sentimental Analysis

In this Methodology Sentiment analysis is a natural language processing technique, helps to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of the writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation

The extracted features contribute to a document vector upon which various machine learning techniques can be applied in order to classify the polarity (positive and negative opinions) using the obtained document vector and finally the opinion impact is obtained based on the sentiment of the web users.

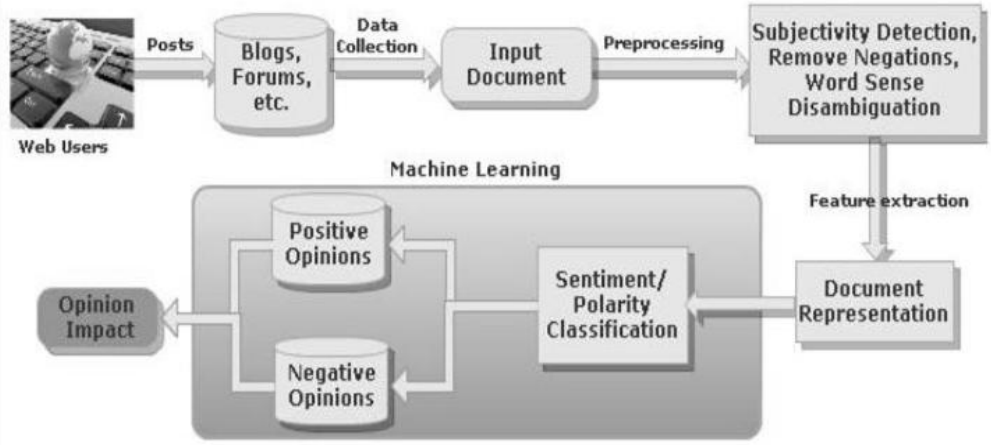

Figure 1: System Flow

Opinion words encode an emotional state, which can be desirable or undesirable. Opinion words that encode desirable states (beautiful, nice, happy, awesome) have a positive orientation, while the ones that encode undesirable state (bad, terrible, disappointing) have a negative orientation. Sentences may contain one or more words and usually analyzing the relation of these words is sufficient to understand if an opinion is negative, positive or neutral.

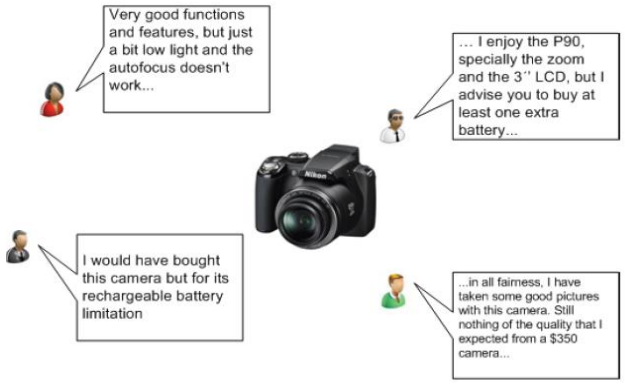

Figure 2: Users opinions for a digital camera - The nature of customers opinions show that the same sentence can share both positive and negative aspects about products features.

A new technique to automatically propagate information of each expert tagged entity to its synonyms, antonyms, similar words and related words.

4. Sentiment Classification

Sentiment Classification broadly refers to binary categorization, multi-class categorization, regression and ranking. Sentiment Classification mainly consists of two important tasks, including sentiment polarity assignment and sentiment intensity assignment [4]. Sentiment polarity assignment deals with analyzing, whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment deals with analyzing, whether the positive or negative sentiments are mild or strong. There are several tasks in order to achieve the goals of Sentiment Analysis. These tasks include sentiment or opinion detection.

A new technique to automatically propagate information of each expert tagged entity to its synonyms, antonyms, similar words and related words. Figure 3 illustrates an example. As mentioned below, an entity can be a single word or a phrase. By expanding each single word to a list of its related words, different word combinations can be formed. In Figure 3, the sentence *"Good picture quality"* is an expert tagged opinion sentence. During the training course, the system looks up synonyms and antonyms for opinion entities. The tag of the original opinion entity ~~*good*~~", <OPINION_POS_EXP>

(*positive opinion*), gets propagated to each synonym of ~~good~~" (red box on the left in Figure 3). The negative tag <OPINION_NEG_EXP> gets propagated to ~~good~~ "'s antonyms (dark red box on the bottom left). Similarly, for each single word in other entity types, similar words and related words are looked up. The tag of the original word gets propagated to each newly discovered related word (blue boxes). Using this expansion, a number of bi-gram combinations (green arrows) can be obtained. In this example, there are several possible instances derived from *"Good picture quality"*, such as ~~D~~ecent picture quality", ~~P~~oor image quality", and etc. Obviously, only *"Good picture quality"* is the expert tagged truth data.

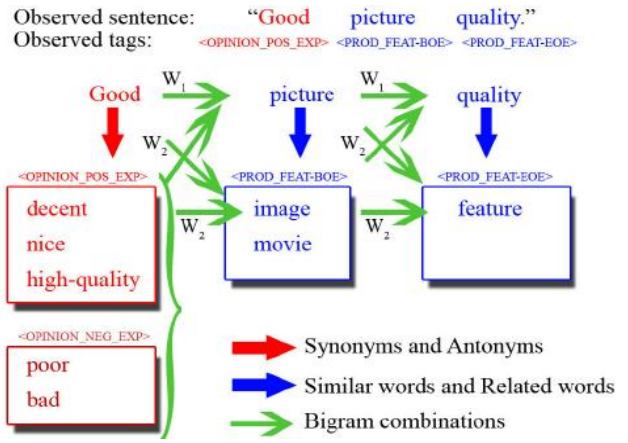

Figure 3: Information propagation using entity's synonyms, antonyms, similar words and related words

5. Feature based Sentimental Classification

Feature engineering is an extremely basic and essential task for 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.

Machine Learning Approaches

The aim of Machine Learning is to develop an algorithm so as to optimize the performance of the system using example data or past experience. The Machine Learning provides a solution to the classification problem that involves two steps:
1) Learning the model from a corpus of training data
2) Classifying the unseen data based on the trained model. In general, classification tasks are often divided into several sub-tasks:

1) Data preprocessing
2) Feature selection and/or feature reduction
3) Representation
4) Classification
5) Post processing

Feature selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features) for the remaining steps of the task. The classification phase of the process finds the actual mapping between patterns and labels (or targets). Active learning, a kind of machine learning is a promising way for sentiment classification to reduce the annotation cost[5]. The following are some of the Machine Learning approaches commonly used for Sentiment Classification.

6. Parts of Speech Tagging

One special application of natural language processing is determining the part of speech of each word in a sentence, known as part-of-speech (POS) tagging. The part-of-speech is a category used in linguistics that is defined by a syntactic or morphological behavior of a word. The traditional English language grammar classifies parts-of-speech in the following categories: verb, noun, adjective, adverb, pronoun, preposition, conjunction and interjection

In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context —i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times.

Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb, or other lexical class marker to each word in a sentence. POS tagging is a necessary pre-module to other natural language processing tasks like natural language parsing, semantic analyzer, information extraction and information retrieval. A word can occur with different lexical class tags in different contexts. The main challenge in POS tagging involves resolving this ambiguity in possible POS tags for a word. We developed a POS tagger which will assign part of speech to the word in a sentence provided as input to the system. Here we have assigned five tags only viz. noun, adverb, adjective, verb and pronoun. Several approaches have been proposed and successfully implemented for POS tagging for different languages[6]. There are various approaches of POS tagging, which can be divided into three categories rule based tagging, statistical tagging and hybrid tagging.

A. **Rule based approach:**
The rule based POS tagging model requires a set of hand written rules and uses contextual information to assign POS tags to words. The main drawback of rule based system is that it fails when the text is unknown, because the unknown word would not be present in the WordNet. Therefore the rule based system cannot predict the appropriate tags.

B. **Statistical approach:**
A statistical approach includes frequency and probability. The simplest statistical approach finds out the most frequently used tag for a specific word from the annotated

training data and uses this information to tag that word in the un annotated text. These systems are having more efficiency than the rule based approach. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language.

C. **Hybrid approach*:***

Hybrid approach may perform better than statistical or rule based approaches. The POS tagger which is implemented using hybrid approach is having higher accuracy than the individual rule based or statistical approach.

Part-of-speech taken from opinions on the Internet:
I. I have a good camera.
II. My camera is bad.
III. The autofocus function never works.
IV. I always have problems with the autofocus.
V. This camera comes with a recharger that works everywhere.

In I and II, both good and bad qualifies the noun camera. In III, IV and V the words never, always and everywhere, also qualify the noun but different from adjectives they modify the verb. After classifying parts of a sentence, many algorithms of data mining can benefit from it, using statistical methods to determine the likelihood of a word to belong to a specific group.

Example
Part-of-Speech from Penn Treebank
I. I **PRP** have **VBP** a **DT** good **JJ** camera **NN** . .
II. My **PRP$** camera **NN** is **VBZ** bad **JJ** . .
III. The **DT** autofocus **NN** function **NN** never **RB** works **VBZ .** .
IV. I **PRP** always **RB** have **VBP** problems **NNS** with **IN** the **DT** autofocus **NN** . .
V. This **DT** camera **NN** comes **VBZ** with **IN** a **DT** recharger **NN** that **WDT** works **VBZ** everywhere **RB** . .
Each tag represent a part-of-speech (JJ, NN, VBZ, etc). In the above annotation JJ stands for adjective while NN means nouns.

7. Sentiment Analysis System Stages [SAS]

1) **Data Collection**: In this step, it is important to get the piece of concern, that is, what we actually want to know
2) the opinion about. It is also important to remove all facts that don't express opinions like news and objective phrases. The focus is on the user's opinions.
3) **Pre-processing:** The preprocessing is also important in order to remove unnecessary words or irrelevant words from the user's opinions. It deals with strings tokenization and punctuations removal. This processing system deals only the description part of each review, here processing means splitting review into sentences to create a plain text file of reviews.
4) **Part of Speech Tagging (POS) and Classification:** The polarity of the content that must be identified. Generally,the polarities used are +ve, -ve or neutral. POS is a process whereby tokens are sequentially labeled with syntactic labels, such as "Verb" or "Noun" or "Adjective". POS tagger was used to define boundaries (split opinions into sentences) and to produce for each word a given part-of-speech.
5) **Stop Words Removal:** We remove words like digits, prepositions, articles, conjunctions and proper nouns like name of Movie etc from the POS tagged review file, as their existence are pointless in this system. It helps better extraction of opinion phrases/words from the POS tagged file.
6) **Negation tagging:** Negation tagging aims at identifying such words and reflecting their effects when determining the sentiment orientation of reviews. For example, "good" and "not good" obviously represent opposite sentiments. We used fuzzy string matching using regular expressions when identifying negation words to handle word variants.
7) **Summarization of Results:** In this step, the categorizations of numerous opinions are summarized in order to be presented to the end user. The goal is to make possible the considerate and give a general comprehension idea to business pundits about what people are talking about an item or product. The summarization is expressed in graphical and textual format.

Sentiment Calculation

We are able to analyse the total text in its positive and negative polarity for the particular feature.
S(W+ve) = Set of Positive Sentiment words
S (W-ve) = Set of Negative Sentiment words
For Nth Feature
S (W+) = (W1 + W2 + W3 +……Wn) Set of Positive Sentiment Words
S (W-) = (W1 + W2 + W3 +……Wn) Set of Negative Sentiment Words
First of all we calculate the negative and positive polarity every sentiment by calculating its probability.
Then we shall calculate the mutual information of its positive or negative sentiment. After calculating the overall polarity of the individual feature we shall calculate the Mutual information for every feature for its final product review Analysis.

8. Conclusion

In this paper we survey on Discrimination Analysis , Sentimental analysis and Part of Speech Tagging is playing a vital role in most of the natural language processing applications. The rule based POS tagger described here is resolving ambiguity and assigning the tags to the ambiguous words using English grammar rules.

9. Future Challenges

There are several challenges in analyzing the sentiment of the web user reviews. In recent advances, there are still several promising new directions for developing and advancing new opinion mining research. For example, much past and current opinion mining research has focused on English, Chinese, Arabic, and several European languages. Advanced and mature techniques have been developed

especially for English. However, in light of the large amount of public opinions expressed by citizens in different parts of the world, new, scalable opinion mining and sentiment analysis resources and techniques need to be developed for various languages.

References

[1] S.Hajain, J.Domingo Ferrer, and A.Martinez Balleste,"Rule protection for Indirect Discrimination Prevention in Data Mining", Proc.Eighth Int'l Conf.Modeling Decisions for Artificial Intelligence (MDAI'11).pp.211-222, and 2011.

[2] S. Shivashankar and B. Ravindran, ―Multi Grain Sentiment Analysis using Collective Classification", *Proceedings of the European Conference on Artificial Intelligence*, pp. 823-828, 2010.

[3] George Stylios, Dimitris Christodoulakis, Jeries Besharat, Maria-Alexandra Vonitsanou, Ioanis Kotrotsos, Athanasia Koumpouri and Sofia Stamou, ―Public Opinion Mining for Governmental Decisions", *Electronic Journal of e-Government,* Vol. 8, No. 2, pp. 203-214, 2010.

[4] AnindyaGhose and Panagiotis G. Ipeirotis, ―Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 23, No. 10, pp. 1498-1512, 2011.

[5] JanyceWiebe and Ellen Riloff, ―Finding Mutual Benefit between Subjectivity Analysis and Information Extraction", *IEEE Transactions on Affective Computing*, Vol. 2, No. 4, pp. 175-191, 2011.

[6] Huifeng Tang, Songbo Tan and Xueqi Cheng, ―A survey on sentiment detection of reviews", *Expert Systems with Applications,* Vol. 36, pp. 10760–10773, 2009.

[7] Ahmed Abbasi, Stephen France, Zhu Zhang and Hsinchun Chen, ―Selecting Attributes for Sentiment Classification Using Feature Relation Networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 3, pp. 447-462, 2011.

[8] 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.

[9] A. Abbasi, H. Chen and A. Salem, ―Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", *ACM Transactions on Information Systems*, Vol. 26, No. 3, Article No. 12, 2008.

[10] SugeWang, Deyu Li , Xiaolei Song , Yingjie Wei and Hongxia Li, ―A feature selection method based on improved fisher's discriminant ratio for text Sentiment Classification", *Expert Systems with Applications,* Vol. 38, No. 7, pp. 8696–8702, 2011.

[11] Q. Ye, Z. Zhang and R. Law, ―Sentiment classification of online reviews to travel destinations by supervised machine learning