Study on Distinct Approaches for Sentiment Analysis

Rupali P. Jondhale
Department of Computer Engineering
VIIT, Kondhwa(Bk.)
Pune-48, India

Manisha P. Mali
Department of Computer Engineering
VIIT, Kondhwa(Bk.)
Pune-48, India

ABSTRACT

Now-a-days many researchers work on mining a content posted in natural language at different forums, blogs or social networking sites. Sentiment analysis is rapidly expanding topic with various applications. Previously a person collect response from any relatives previous to procuring an object, but today look is different, now person get reviews of many people on all sides of world. Blogs, e-commerce sites data consists number of implications, that expressing user opinions about specific object. Such data is pre-processed then classified into classes as positive, negative and irrelevant. Sentiment analysis allows us to determine view of public or general users feeling about any object. Two global techniques are used: Supervised Machine-Learning and Unsupervised machine-learning methods. In unsupervised learning use a lexicon with words scored for polarity values such as neutral, positive or negative. Whereas supervised methods require a training set of texts with manually assigned polarity values. This suggest one direction is make use of Fuzzy logic for sentiment analysis which may improve analysis results.

Keywords

Sentiment Analysis, Natural Language Processing, Fuzzy logic.

1. INTRODUCTION

Sentiment Analysis or Sentiment Mining is the method of learning public sentiments, feelings or judgments about an object. An object can be anything like some topic, product or service. Those can be collected by reviews posted at different sites. Analyzing sentiments and mining opinions are slightly different task. Mining take out and examines public opinion for an object on other hand Analysis of sentiments recognizes the sentiment present in content after that analyzes it. So, the goal of Sentiment Analysis is to find sentiments expressed in text, then classify sentiment according to polarity.

Analysis is consider as a classification process. Mainly three methods or levels of classification in Sentiment Analysis 1)Document level, 2)Sentence level and 3)Feature level. In first level of analysis categorize documents representing as positive or negative sentiment. Consider complete document about single topic as processing unit. Goal of sentence level is organizing sentiment given in every sentence. Initial step is find out given sentence is subjective or factual. If it is subjective then sentence can be classified as negative or positive sentence. While at Feature level classification is performed at fine grain level. Recently number of reviews present on website has develop incredibly and hence analyzing sentiment become tough than before. Data is given in unstructured form. Such data is very uncertain and not definite to find out sentiments and expressions. Extracting sentiments have some challenges and need helpful techniques or methods to neatly take out and convey public views, which can also be important for selling research. The relevant areas to Sentiment Analysis are transfer learning, feeling detection and construction of resources which could catch attention of researchers. The main aim of paper is to collect and present current directions of research in the sentiment analysis area.

Sentiment Analysis:

Sentiment analysis is the technique to detect and extract subjective information from unstructured text as shown in figure 1. Generally, complete and appropriate polarity detection for sentiment of people about some feature can be governed using sentiment analysis. The main challenge is the sentiment classification in area where sentiment may be a judgment, opinion or evaluation of an objects like book, movie, product, etc which can be in the form of document or sentence or feature that can be classify as positive or negative. Classifying the whole document in relation to the opinions about particular object is called as sentiment classification. One form of mining in product reviews is to produce feature-based summary. To produce a summary on the features, product features are first identified, and positive and negative opinions on them are aggregated. Features are product attributes, components and other aspects of the product.

Sentiment Analysis tasks:

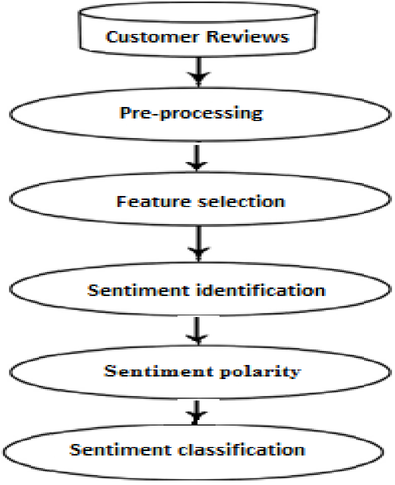

Figure 1: Sentiment analysis process

2. RELATED WORK

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) deals with actual text element processing. Customer reviews are given in unstructured manner so pre-processing is important task to be performed before sentiment classification. In pre-processing unambiguous data is filtered through different processes then it is given for classification. NLP consist of many stages such as tokenization, stop word removal, stemming to filter out ambiguous text. Once tokens are generated different techniques are can be used to relate words representing features and opinions.

2.2 Point-wise Mutual Information (PMI)

The Point-wise mutual information method (Jian et. al. 2011) gives a proper path to co-relate information among number of aspects and their classes. Information model is used to develop such PMI model. The common point-wise information Mi for word w can be represented using number of times relation occurs to relate class named as i and w. The mutual information is defined in terms of the ratio between these two values and is given by the following equation:

$$\boldsymbol{M_i(w) = log\left(\frac{F(w)\ .\ p_i(w)}{F(w)\ .\ P_i}\right) = \mathbf{log}\ (\frac{p_i(w)}{P_i})} \qquad \textbf{[3]}$$

If Mi(w) is > 0, then w is optimistically associated word to given group i, otherwise w is pessimistically associated to the group i. One time root words have extended, then root words and extended words are use to arrange sentiment.

2.3 Latent Semantic Indexing (LSI)

LSI is well known aspect based alteration technique. LSI transforms given information to system that is a linear combination of real pattern of aspects. Principal Component Analysis techniques (PCA) are use for this purpose. It calculate axis of a system which gives the information of variations in the underlying attribute values. The main drawback of LSI is an unsupervised method which is unknown to class-division. LSI is used at document level classification. Many other statistical techniques used for Feature Recognition are Hidden Markov Model (HMM) and Latent Dirichlet Allocation (LDA) were proposed by Duric and Song. Split the entities in document from particular patterns represent such entities as orientations. This was their proposed new feature selection schemes. LDA are creative method to allow documents to be describe by latent subjects. HMM-LDA is topic based model that concurrently models topics and their syntactic arrangement in set of documents. The feature selection schemes proposed by Duric and Song attained competitive results for document polarity categorization specially when using only the syntactic classes and minimize overlaps with the semantic words in their result aspect sets used movie reviews and Maximum Entropy (ME) classifier.[9]

2.4 Challenges in feature selection

A challenging task in extracting features is irony detection. The aim is to identify irony comments. This work was proposed by Reyes and Rosso. The goal is to give aspect based model to present some part of subjective information which display reviews and try to explain prominent features of irony. They created system to give verbal irony in the form of six groups of features: n-grams, funny profiling, POS-grams, positive/negative profiling, pleasantness profiling and affective profiling. To construct openly available data groups with sarcastic reviews from news and satiric articles and customer reviews gathered from amazon site. They were posted on basis of an online viral effect contents that cause a chain reaction in public. They make use of Naive bayes, SVM for classification. Their results with the three classifiers are satisfactory, both in terms of accuracy as well as precision, recall, and F-measure.

3. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment Classification techniques can be roughly divided as lexicon based, machine learning approach as shown in figure 2. The Machine Learning Approach use well-known ML techniques and uses linguistic features. The Lexicon-based Approach relies on a lexicon of sentiments, a set of well-known and previously compiled some sentiment words. It is classified into corpus-based and dictionary-based approach make use of general numerical or semantic techniques to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing an important role in the majority of methods.

The text categorization schemes using ML method can be split into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents. The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

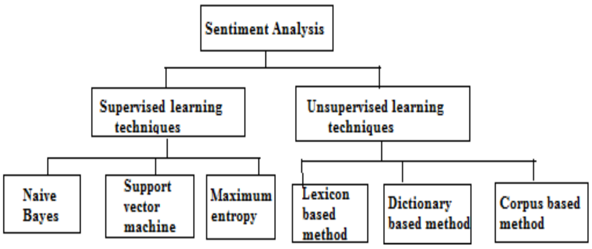

Figure 2: Sentiment classification techniques

3.1 Unsupervised learning methods

3.1.1 Lexicon-based approach

Sentiment words are used in sentiment categorization tasks. Positive words of opinion are used to express some definite states and negative words of opinion are used to express some indefinite states. Also opinion phrases and idioms which together are called sentiment lexicon. There are three main approaches in order to compile or collect the opinion word list. Manual approach is very time consuming and is normally combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The following are two automated approaches.

3.1.2 Dictionary-based approach

The key policy of the dictionary-based approach is small set of opinion words is grouped manually with well-known polarities. Such set is expanded by finding the famous thesaurus or corpora like WordNet for searching words synonyms and antonyms. The newly disclosed terms are inserted to the main list of seed words after that further processing starts. The iterations ends when novel words are not found to add in the seed list. Once procedure is completed, manually checking can be carried out to delete or correct errors.

3.1.3 Corpus-based approach

The Corpus-based method is use to resolve difficulty of searching sentiment words with domain specific polarities. It is based on syntactic patterns or patterns that take place combinely along with root list of sentiment words to search other sentiment words into huge corpus. Using the corpus-based method alone is not much efficient as the dictionary approach the reason is, it is tough to construct large corpus to collect all English words, but it has great benefit that it can help to search class and domain specific sentiment terms and their orientations with the help of particular domain specific corpus.

3.2 Machine learning methods

The machine learning methods applicable to analysis mainly fit into supervised classification and text classification techniques hence it is called Supervised learning. In a supervise learning categorization, require two sets are: training set and testing set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and Support Vector Machines (SVM) have achieved great success in text categorization. The other most popular machine learning methods in the natural language processing area are K-Nearest neighborhood, ID3, C5, centroid classifier, and the N-gram model.

3.2.1 Naive-Bayes Classifier

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is a probabilistic method and it allows us to imprison ambiguity about model in a standard way by determining probabilities of the outcomes. It can solve analytical and prognostic problems. Such categorization is named after Thomas Bayes, he presented Bayes Theorem. Bayesian classification offers practical ways of learning algorithms and previous knowledge and studied information can be merged. This categorization gives helpful point of view for accepting and estimating learning algorithms. It examines explicit possibility for suggestion and is tough to noise in the input information. Bayes classifiers are the majorly doing well than other known algorithms for knowledge to classify the text documents.

3.2.2 Maximum Entropy Classifier

A maximum entropy classifier can be used to extract sentences from documents. Experiments using technical documents show that such a classifier tends to treat features in a categorical manner. Importantly, unlike Naive Bayes, Maximum Entropy makes no assumptions about the relationships between features and so might potentially perform better when conditional independence assumptions are not met.

3.2.3 Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning techniques can be used for regression or categorization type of problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to separate data based on the labels or outputs defined. Support Vector Machine (SVM) carry out classification by developing an *N*-dimensional hyper plane which best take apart information into two classes. Support Vector Machine techniques are adjacent to neural networks. Support Vector Machine method using a sigmoid kernel function is equivalent to perceptron of two-layer, perceptron which is neural network type. LIBSVM is a popular collection for Support Vector Machine that supports multiple class classification.

3.3 Other approaches

3.3.1 Fuzzy Concept

Fuzzy logic is recently used in many fields of Artificial Intelligence to deal with ambiguous information. In sentiment analysis sentiments present in reviews are ambiguous in nature. So predicting sentiments through these words gives imprecise results. Fuzzy logic is flexible. It mainly depends on human beings nature or behavior. Experts knowledge is used for reasoning purpose.

Fuzzy logic can be helpful to predict sentiment present in text posted at different websites, blogs. Comments are given in natural language consists uncertain words, so to remove such fuzziness or ambiguity fuzzy logic can be used. But some disadvantages of fuzzy such as it is not able to handle negation present along with opinion words.

Guohong and Xin (2010) presented a fuzzy set theory based on framework for Chinese sentence-level sentiment classification. Animesh and Debael (2011) presented an opinion mining systems called Fuzzy Opinion Miner (FOM). FOM is a supervised opinion polarity detection system that mines reviews using Fuzzy logic.

Fuzzy gives intelligent and excellent control over different techniques. It is also represented in simple way to find out implicit sentiments. Fuzzy set theory provides clear boundaries for multiple terms. Also using fuzzy logic multiple level of classification can be done instead of only binary classification.

4. CONCLUSION

In the area of sentiment analysis text processing given in natural language at different websites is really hard task to perform. Main tasks are: Pre-processing text is very helpful for filtering some unwanted information. For extracting features out of given reviews and then classify them into different classes is the main aim to perform sentiment analysis. Supervised learning techniques surely perform well in particular domain based on training data. But unsupervised methods are domain independent. In future, incorporating Fuzzy concept with unsupervised learning techniques may give better results as it is simple to understand and helpful to remove ambiguity present in unstructured text.

5. REFERENCES

[1] Yan Dang, Yulei Zhang, and Hsinchun Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews", *IEEE Computer Society* Intelligent Systems, 1541-1672, 2010.

[2] Guohong and Xin, "Sentence based analysis using fuzzy set theory framework", *IEEE International conference on* Fuzzy Systems (FUZZ); May 2010.

[3] Janyce Wiebe and Ellen Riloff, "Finding Mutual Benefit between Subjectivity Analysis and Information Extraction", *IEEE Transactions on* Affective Computing, October-December 2011.

[4] Jian, Bai X., Berberidis ," Predicting consumer sentiments from online text", *International Journal of* Decision Support System;50:732–42,June 2011.

[5] Animesh and Debael , "Noise control in document classification based on fuzzy formal concept analysis", Presented at the *IEEE International conference on* Fuzzy Systems (FUZZ); May 2011.

[6] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang and Aijun An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", *IEEE Transactions on* Knowledge and Data engineering, April 2012.

[7] Jaganadh G, "Opinion Mining and Sentiment Analysis", CSI Communications, May 2012.

[8] Chenghua Lin, Yulan He, Richard Everson, Member and Stefan Ruger, "Weakly Supervised Joint Sentiment-Topic Detection from Text", *IEEE Transactions on* Knowledge and Data engineering, June 2012.

[9] Duric Adnan, Song Fei.," Feature selection for sentiment analysis based on content and syntax models", *International Journal of* Decision Support System;53:704–11, 2012.

[10] Jinan Fiaidhi, Osama Mohammed, Sabah Mohammed, "Mining Twitter space for Information: Classifying Sentiments Programmatically using Java", *International Journal of* Computer Trends and Technology (IJCTT), October 2013.

[11] Jayashri Khairnar, Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", International Journal of Scientific and Research Publications, June 2013.

[12] XU Xueke, CHENG Xueqi, TAN Songbo, LIU Yue, SHEN Huawei, "Aspect-Level Opinion Mining of Online Customer Reviews", *China Communications on* Management and visualization of user and network data, March 2013.

[13] Danushka Bollegala, Member, David Weir, and John Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", *IEEE Transactions on* Knowledge and Data engineering, August 2013.

[14] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades, "Ontology-based sentiment analysis of twitter posts", *IEEE Transactions on* Expert Systems with Applications 40, 4065–4074, 2013.

[15] Poongodi S, Radha N, "Classification of user Opinions from tweets using Machine Learning Techniques", *International Journal of* Advanced Research in Computer Science and Software Engineering, September 2013.

[16] Reyes and Rosso, "Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text," *Lecture Notes* in Computer Science, vol. 3932, pp. 167–187, January 2014.

[17] Isidro Peñalver-Martinez a, Francisco Garcia-Sanchez, *"*Feature-based opinion mining through ontologies", *IEEE Transactions on* Expert Systems with Applications, May 2014.