A Contemporary Overview on Feature Selection and Classification Techniques in Opinion Mining

Aansi A. Kothari
Department of Computer Science and Engineering
Parul Institute of Technology
Waghodia, Vadodara - 391760, India

Warish D. Patel
Department of Computer Science and Engineering
Parul Institute of Technology
Waghodia, Vadodara - 391760, India

ABSTRACT
Opinion mining is a booming area which has swiftly and definitely captured a lot of attention recently. Right from education to shopping, job or home, be it a political or a social affair, professional or social task, humans need opinions in anything they do. Opinions can be either manual or online. As the era of internet has taken layman along, we have centered our study towards the study of online opinions and reviews. Opinion mining comprises of various ways to track how opinion techniques evolve over time to help identify opinions and patterns and generate recommendations or take decisions. In this paper we summarize the opinion mining process along with various computational techniques, algorithms and models that contribute towards mining of opinion components from various reviews or comments from one or more sources. We further provide future directions for research in this field.

Keywords
Opinion Mining, Review Mining, Sentiment Analysis, Opinion Mining Techniques

1. INTRODUCTION
With the advent of World Wide Web, the way with which we interact with or manage the information, has budged into a different direction. It has become effortless and sort of absolute to obtain our required information within seconds and from multiple resources, on WWW. Even a layman now does not simply surf or simply read data on internet but annotate this information and build a new piece of information on basis of it. They now actively participate in reading and sharing information as well as expressing their own experience and views on it. Today people not only comment, share or bookmark but also present their own ideas and views at large via various social and other online mediums like facebook, twitter, blogs, forums, etc. This way, a rich and wonderful source of information on various aspects ranging from politics to health, travelling, books, day-to-day activities, products we use and at last even on ongoing incidents is represented by them. Social media has also gained a lot of attention here and the comments and reviews placed by people definitely secure a room of importance. These opinions play a vital role in human life today. Opinions can either be expressed online or manually.

Opinion mining is an area where evaluation or study of people's reviews, comments, attitudes, habits, judgment towards various entities, individuals, places, events and attributes, sentiments, etc. takes place based on the knowledge or experiences that they possess. An opinion can be positive, negative or neutral [21].

Opinion can be termed as a sentiment or a view or a judgment made any object, individual, task or a process based on knowledge or experience. Opinion holder is the person or an organization that holds the review or sentiment regarding any object. Object is an entity which may be a person, topic, event, product or an organization about which an opinion or a review can be expressed [21].

Opinion mining usually occurs on three levels given as Document-level, Sentence-level where the whole document is broken into sentences and Feature-level where features, aspects or contexts are concentrated for opinion mining.

2. LITERATURE REVIEW
The task of opinion mining is segregated into five main steps as collection of opinions from World Wide Web, Pre Processing, feature selection, opinion classification and summarization and performance evaluation as shown below:

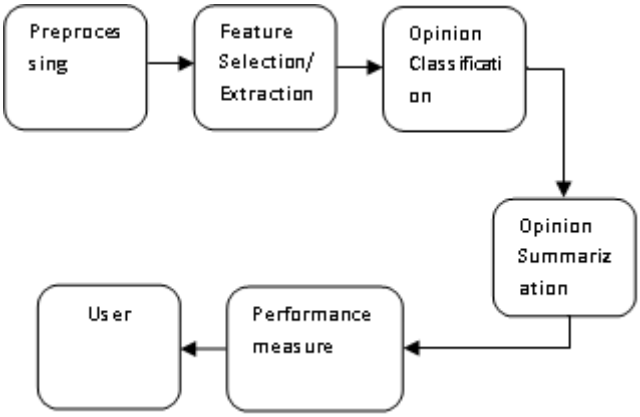

Fig 1: Steps of Opinion Mining Process

Pre Processing
Raw data is first collected from various sources and then it undergoes preprocessing steps that is divided mainly into three common phases: Tokenization, Stop-word (a, an, the, etc.) Removal, Stemming phase and lastly Case Normalization phase where the entire document is either converted in upper case or lower case[20].

Feature Selection And Extraction
Then follows the Feature Selection and Extraction Step in which the feature is first identified, followed by the selection procedure and then extraction and reduction process if required. Feature Identification includes understanding of various feature types such as Term frequency, Term Co-occurrence, Part-Of-Speech and Opinion Words, for identification purpose. We have brooded in the following section on feature selection techniques.

Feature Selection Techniques
There are various techniques for feature selection such as Stemmed Terms, Based on minimum, Dependence Y-relation, Graph Distance, TF-IDF, Opinion Words, Document Frequency, MI, IG, CHI, N-gram out of which some famous methods are discussed below:

Document Frequency
The number of documents that comprises of a given term is stated as a Document Frequency. In this method firstly the document frequency is compared with the predefined lower Document Frequency Threshold and upper threshold. Those words or terms that hold frequency lower than lower threshold and the ones that hold higher frequency than upper threshold are eliminated. It is assumed that the most uncommon words and even the one that are too common are either not contributing to influence in global performance or are non-informative for predicting categories. This method is the simplest yet effective selection method for categorization of text [9].

CHI
The CHI statistic method helps in computing the association between term and category. It is formulated as follows [9]:

$$CHI\ (t, c_i) = \frac{\mathrm{N} \times (\mathrm{AD} - \mathrm{BE})^2}{(\mathrm{A} + \mathrm{E}) \times (\mathrm{B} + \mathrm{D}) \times (\mathrm{A} + \mathrm{B}) \times (\mathrm{E} + \mathrm{D})}$$

and,

$$CHI_{max}\ (t) = \max_{\mathrm{i}}(CHI\ (t, c_i))$$

Where, A gives the number of times t along with c_i; B is the number of times t occurs c_i does not; E is the number of times c_i occurs and t does not; D is the number of times neither c_i nor t occurs; N is the total number of documents.

Mutual Information
Mutual information is a criterion that is used generally in statistical language modelling of association of terms and their correlated applications. It is expressed as following [9]:

$$MI\ (t, c_i) = \log\left(\frac{\mathrm{A} \times \mathrm{N}}{(\mathrm{A} + \mathrm{E}) \times (\mathrm{A} + \mathrm{B})}\right)$$

and,

$$MI_{max}\ (t) = \max_{\mathrm{i}}(MI\ (t, c_i))$$

Where, A gives the count of occurrence of t along with c_i; B is the number of occurrences of t when c_i does not; E is the number of times c_i occurs and t does not; N shows the total number of documents.

Information Gain
This method by knowledge of the presence or absence of a term in the document, calculates the number of bits that are used for prediction of category.

This method is often employed as a goodness standard for term. It is expressed as follows [10]:

$$IG(t) = \sum_{i=1}^{|C|} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{|C|} P(c_i|t) \log P(c_i|t) + P(\bar{t}) \sum_{i=1}^{|C|} P(c_i|\bar{t}) \log P(c_i|\bar{t})$$

Where $P(c_i)$ represents the probability of occurrence of c_i class; $P(\mathrm{t})$ represents the probability of occurrence of word t; $P(\bar{t})$ represents the probability of non-occurrence of word t.

TF-IDF
It presents a weighting method that is frequent in opinion mining. In the document in a corpus, this method is used to compute the importance of a term. The number of times a given term appears in that given document is defined as Term Frequency (TF). The measure of general importance of a term is known as Inverse Document Frequency (IDF). TF-IDF can be articulated as follows [3]:

$$W_i = tf_i * \log(D/df_i)$$

Where, tf_i demonstrates the term frequency of term i in a given document, D is the number of documents in the corpus, and df_i demonstrates the document frequency or number of documents containing term i. Hence, $\log(D/df_i)$ is represented as the inverse document frequency.

Opinion Classification
In the fourth step, we try to classify selected features by various classification techniques discussed in our paper as follows:

Naive Bayes
This is one of the simplest and very frequently used classification techniques. With use of probabilistic model that takes independent assumptions by distributions of different terms, it performs distribution of documents in each class. Naive Bayes classification uses two classes of models that calculate posterior probability for each class depending on the word distribution in the class. Disregarding actual position of the words in the document it assumes bag of words [10]. To formulate posterior probability,

$$P(H|X) = \frac{\mathrm{P(X|H)P(H)}}{\mathrm{P(X)}}$$

One of its disadvantages is that it does not take into consideration of the fact that attributes are dependent on each other. Despite of the fact, it still performs feasibly. The main drawback of Naïve Bayes is that it gives very poor results when features are correlated to each other. It can be further divided into:

- Multivariate Bernoulli Model [9]: As a feature this model takes presence or absence of the word in a text document and it assumed with two values, either of presence or absence. Hence, it can be assumed to model binary values thus making the model in each class document to be a multivariate Bernoulli model.
- Multinomial Model [10]: This model takes into account, frequency of terms present in the document. The document may contain lots of words and thus can be expressed as "bag of words". As a result, the conditional probability of a document given a class is simply a product of the probability of each observed word in the corresponding class.

Opinion Word
A review may contain opinion sentences that can be either positive or negative and we need to identify it. They may contain opinions that have not only one but more than one product features. We define three sub-tasks: 1) Identify a Group or a set of opinion words. Another simple hint to this is, if any adjective appears close to product feature, it can be considered as opinion word. 2) Now, determine the semantic orientation for each opinion word as either positive or negative. 3) With the use of dominant orientation, we choose the orientation of the opinion [4].

Lexical approach
A glossary or a word list can be termed as lexicon. It can be of a positive or a negative term. We make use of this knowledge to compute the frequency of the occurrence of these terms in each document. The probability of test document being positive and negative is given as [4]:

$$P(+|D) = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}}$$

Here, a and b denote the occurrence of positive and negative terms in the given documents respectively. If P(+|D) > t , document is classified positive, else negative. t is termed as classification threshold. When there is uncertainty of information about positivity or negativity of terms we take t=0.5.

Centroid classification
This is one of the simplest and basic algorithms used for classification purpose. Firstly, for each training class prototype vector (centroid vector) is computed, after which is calculated the similarity between the testing documents and all the centroids. Now on the basis of these we allocate d to the class that has the most similar centroid. Following is the formula that helps in computing K centroids {C_1,C2 ,. . ., C_K} for the K classes [9]:

$$\mathrm{C_i} = \frac{1}{|C_i|}\sum_{d \epsilon\, c_i} d$$

Where |z| denotes the cardinality of set z and d indicates the document in class C_i. For each test document d, we calculate its similarity to each centroid $C_i$using cosine measure as follows:

$$sim(d, C_i) = \frac{\mathrm{d}\,.\,\mathrm{C_i}}{||\mathrm{d}||_2||\mathrm{C_i}||_2}$$

Winnow classifier
This method updates its weights in a sequence of trials. On every single trial, the prediction is made for one document first and then the feedback is received. After the feedback, if a mistake is found to have occurred, the weight vector is updated using d document [9]. In training phase, by iterating on data, same process is repeated for several times, with a collection of training data. It is divided into various variants like positive winnow, a balanced winnow and a large margin winnow, out of which the balanced winnow consistently gives outstanding performance. This algorithm keeps, w_{kt}^{+} and w_{kt}^{-} for each feature. Hence, for any given instance (d_{k1}, d_{k2}, . . . ,d_{kW}), the document is considered relevant iff,

$$\sum_{t=1}^{W} (w_{kt}^{+} - \; w_{kt}^{-})\, d_{kt} \;\geq\; \tau$$

Where, τ is expressed as a given threshold and k defines the class label.

K-means clustering
It takes the idea based on k-means clustering algorithm. There are two groups in which documents are classified, a positive and a negative. Yet there is an issue of poor result when we talk about aspects of accuracy and stability. To fix these, we designed three methods namely and out of which TF-IDF (Term Frequency- Inverse Document Frequency) weighting method is applied to the raw data first, after which a voting mechanism is used to extract result of clustering. Lastly, the method term score is used to enhance the results[11].

Maximum entropy
This model was described by Jaynes. This model prefers least biased distribution that exploits the uncertainty present in the distribution subject with respect to given limitations. This model helps in prediction of product feature relation in accordance of the opinions. It helps in classification problem that observes textual context xϵX and thereby, predicts the correct linguistic class yϵY. These classes are bifurcated into opinion-relevant product feature and opinion irrelevant product feature. Then we can apply a classifier X ‘ Y with using conditional probability selecting class y with the highest conditional probability p in the context of x given as [5]:

$$cl(x) = \arg\max p(y|x)$$

This conditional probability p(y|x) can be given as follows:

$$\mathrm{p(y|x)} = \frac{1}{Z(x)}\prod_{i=1}^{k} \alpha^{\mathrm{f_i(x,y)}}$$

where,

$$\mathrm{Z(x)} = \sum_{y}\prod_{i} \alpha^{\mathrm{f_i(x,y)}}$$

K-nearest neighbor classifier
This classifier depends on the categorical labels affixed with training documents alike test documents. It is also termed as lazy learners as it puts back the decision about how to generalize beyond the training data until all new query instances are encountered. The system finds the k nearest neighbor for a given test document amongst the training documents. For the weight of the classes of neighbor documents, we take the similarity score of the nearest neighbor document present. Hence, we can give the weighted sum as follows [9]:

$$\mathrm{score(d, c_i)} = \sum_{d_j \epsilon KNN\,(d)} sim(d, d_j)\, \delta(d_j\, c_i)$$

Where, KNN (d) defines the set of k nearest neighbors of document d. If $d_j \epsilon\, c_i$, $\delta(d_j\, c_i)$ = 1, or otherwise 0. The test document d should belong to the class with the highest resulting weighted sum.

Support Vector Machine
Support Vector Machine (SVM) for text classification was proposed by Vladimir Vapnik in 1995.This method classifies linear as well as non-linear data [3, 1]. For transforming training data into higher dimension, it uses non-linear mapping. After transformation of training data, it looks for linear optimal separating hyper plane.SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. Hence in this type of classifiers the optimal boundaries are found between various classifiers and then use it for classification purpose. The main goal of SVM is to improve the speed of training as well as testing. SVM is extended into various different approaches. The optimization of SVM (dual form) is to minimize $\vec{\alpha}^{*}$ as follows:

$$\vec{\alpha}^* = \arg min \left\{ -\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \alpha_i, \alpha_i \rangle \right\}$$

Where, $\sum_{i=1}^{n} \alpha_i y_i = 0;\ 0 \leq \alpha_i \leq C$

Opinion Summarization
Feature classification helps make feature summarization process and recommendation generation process (if needed) easier and quicker.

Performance Evaluation
To evaluate performance of classification, we need to calculate precision, recall and F-measure [1, 12]. Precision is the fraction of documents retrieved that are actually relevant to the query. It is formulated as follows:

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall is the fraction of documents which are query relevant and that were retrieved actually. Recall is formulated as follows:

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

F measure is the both precision and recall and is expressed as follows:

$$\text{F measure} = \frac{\text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})/2}$$

3. CONCLUSION

This paper mainly endeavors at recognizing, selecting and classifying term orientation of the opinionated text. We enlightened a generic model of opinion mining process and then focused on various techniques used at various levels in opinion mining tasks which determine if the document or a task carried a positive or a negative opinion. We then discussed about feature selection techniques. Focused was also put on multiple classification techniques. We observed that there are still some challenges that are significant in feature selection, and classification techniques. The main challenge lies in dealing with negative expressions or the expressions or terms that are positive yet expressed negatively or vice versa, to construct a summary of opinions based on their product features, accuracy and generating recommendations. In future if these challenges are met, it can help generate better recommendations to the user. We can also work in enhancing performance measure and can also explore various useful domains that are yet untouched or very slightly focused.

4. REFERENCES

[1] A. Rashid, N. Anwer, M. Iqbal, M. Sher, "A Survey Paper: Areas, Techniques and Challenges of Opinion Mining", International Journal of Computer Science Issues, Vol.10, Iss.6, 2013, pp. 18-31.

[2] Selvam, S. Abirami, "A Survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering, Vol.2, Iss.9, 2009, pp. 3544-3549.

[3] S. Tan, J. Zhang, "An Empirical Study of Sentiment Analysis For Chinese Documents", Expert Systems with Applications-Elsevier, Vol. 34, Iss.4, 2007, pp. 1-8.

[4] A. Gelbukh, "Computational Linguistics and Intelligent Text Processing", Springer, 2013.

[5] M. Hu, B. Liu, "Opinion Extraction and Summarization on the Web", American Association for Artificial Intelligence, 2006, pp. 1621-1624.

[6] G. Li, F. Liu, "A Clustering-based Approach on Sentiment Analysis", Intelligent Systems and Knowledge Engineering-IEEE, 2010, pp. 331-337.

[7] G. Somprasertsri, P. Laitrojwong "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", Journal of Universal Computer Science, vol. 16, Iss. 6, 2010, pp.938-955.

[8] C.C. Aggarwal, C.X. Zhai, "Mining Text Data", New York-Springer, 2012.

[9] A. Abbasi, H. Chen, A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", Transactions on Information Systems(TOIS)-ACM, Vol. 26, Iss. 3, 2008, pp. 1-6.

[10] Y Yang, X Liu, "A Re-Examination of Text Categorization Methods", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49.

[11] N. Mishra, C.K. Jha, "Classification of Opinion Mining Techniques", International Journal of Computer Applications, Vol.56, Iss.13, 2012, pp. 1-6.

[12] M. Tsytsarau, T. Palpanas, "Survey On Mining Subjective Data On The Web", Data Mining and Knowledge Discovery-Springer, Vol. 24, Iss. 3, 2012, pp. 478-514.

[13] K. Khan, B. Baharudin, A. Khan, "Mining Opinion Components from Unstructured Reviews: A Review", Journal of King Saud University – Computer and Information Sciences, Elsevier, 2014.

[14] A. Patra, D. Singh, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms", Vol.75, No.7, 2013, pp. 14-18.

[15] P. Melville, W. Gryc, R.D. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", Proceedings of the 15th ACM SIGKDD International Conference On Knowledge Discovery And data mining(KDD), pp. 1275-1284.

[16] A.Dasgupta, P. Drineas, B. Harb, V. Josifovski, "Feature Selection Methods for Text Classification", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD)-ACM,2007, pp. 230-239.

[17] T. Liu, S. Liu, Z. Chen, W.Y. Ma, "An Evaluation on Feature Selection for Text Clustering", Proceedings of the Twentieth International Conference on Machine Learning(ICML), 2003.

[18] M. Hu, B. Liu, "Mining and Summarizing Customer Reviews", Proceedings of the tenth ACM SIGKDD

international conference on Knowledge discovery and data mining, pp. 168-177.

[19] N.M. Shelke, S. Deshpande, V. Thakre, "Survey of Techniques for Opinion Mining", International Journal of Computer Applications, Vol.57, Iss.13, 2012, pp. 30-35.

[20] G. Chen and L. Chen," Recommendation Based on Contextual Opinions", User Modeling, Adaptation, and Personalization-Springer, Vol. 8538, 2014, pp. 61-73.

[21] A..Kao, S.R. Poteet, "Natural Language Processing and Text Mining", London-Springer,2007.