On Classifying Sentiments and Mining Opinions

Iasleen Kaur 1, Dr. Jatinderkumar R. Saini 2
1Assistant Professor, 2Associate Professor & Director I/C
1Shroff SR Rotary Institute of Chemical Technology, Ankleshwar
2Narmada College of Computer Application, Bharuch
1, 2 Gujarat, INDIA

Abstract: Due to presence of large volume of opinionated text on web in form of review sites, social media, blogs, discussion forums, People are intended to develop a system that can identify and classify opinion or sentiment expressed in opinionated text. Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Opinion mining can be useful in several ways in marketing, it helps to judge the success of a launch of new product, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features. This paper tries to cover some techniques and approaches that are used in this area. This paper presents a survey covering the techniques and methods in sentiment analysis.

Keywords: Opinion mining, sentiment classification, supervised, unsupervised

I. Introduction

Emotions and opinions plays an important role in human being’s life. Every person’s life is filled up with a lot of emotions and opinions. We cannot imagine the world without them. Emotions and opinions influence the way humans think, what they do, how they act and how they share. Opinions have a relevant impact on our everyday life. Opinions give us information about other person’s experience.

In today’s world, people can express their views/opinions freely and these views/opinions become very useful source of information for making various policies/decisions. With advent of World Wide Web, social media become good source containing opinionated text and these opinions/emotions can be extracted from the text available online. This field of computer science is known as Sentiment classification, Opinion mining. Sentiment Analysis or Subjectivity Analysis. It is the study of person’s views, emotions or opinions associated with some events, some products and their features, some places etc. This task is very challenging and practically very useful. Opinion Mining and Sentiment Classification is a field of study at the crossroad of Information Retrieval (IR) and Natural Language Processing (NLP) and share some characteristics with other disciplines such as text mining and information Extraction.

II. Basic Terminology

Textual information in the world can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people’s sentiments, appraisals or feelings toward entities, events and their properties. Opinion can be expressed in quintuple [4]

(feature, sentiment, polarity, holder, time)

Opinion feature is the target of the opinion. In general an opinion is either had on a specific object or on a feature of such object. An opinion sentiment is the word or set of words that conjointly describe the opinion given on the feature. That is, these are the descriptive terms or opinionated words used to transmit the opinion. The orientation or polarity of an opinion can be either positive, negative or neutral. The holder of an opinion is the person or organization that expresses the opinion. In the case of product reviews and blogs, opinion holders are usually the authors of the posts. An opinion will be held by an opinion holder over some object feature at some point in time.

III. Data Source

User’s opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services [8].

A. Blogs

As internet usage is increasing day by day, blogging and blog pages are growing rapidly. Expressing ones personal opinions through blogs/ blog pages is becoming popular day by day. Bloggers express their opinions,
feelings, and emotions in a blog [10]. Many of these blogs contain reviews on many products, issues, events topics etc.

B. Review sites

The important factor considered for making a decision by a purchaser before purchasing is to know the experience of previous buyer regarding that product. For that a large number of user-generated reviews are available on the Internet. The reviewer’s data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNET download.com (product reviews) and www.reviewcentre.com [16][22].

C. Dataset’s

Most of the work in the field uses movie reviews data for classification. Movie review data are available as dataset (http://www.cs.cornell.edu/People/pabo/movie-review-data) [7]. Other dataset which is available online is multi-domain sentiment (MDS) dataset. (http://www.cs.jhu.edu/mindredziedatasets/sentiment), (http://www.cs.uic.edu/liub/FBS(CustomerReviewData.zip which consists of reviews of five electronics products downloaded from Amazon and CNET [31] [21].

D. Micro-blogging

Most popular micro blogging service like twitter, facebook where information is represented in short messages like tweets or status. These short messages may contain Opinions about different topics which can be used for task of sentiment analysis task [32].

E. Others

Opinion Mining can also be done on Legal Weblogs or Legal Documents [17]. Person can express their emotion/views on politics, fashion trends, about any product, any movie. Discussion forums are also gaining popularity in Medical world. So Opinion mining can be applicable to any of above specified data. Literary Pieces, lyrics and unsolicited bulk mails [18] can also be mined.

IV. Granularity of Opinion Mining task

A. Mining Opinion at Document level

Document level opinion mining is about classifying the overall opinion presented by the authors in the entire document as positive, negative or neutral about a certain object [3]. The assumption is taken at document level is that each document focus on single object and contains opinion from a single opinion holder. The challenge in the document level classification is that the entire sentence in a document may not be relevant in expressing the opinion about an entity. Therefore subjectivity/objectivity classification is very important in this type of classification.

Both supervised and unsupervised learning methods can be used for the document level classification. Turney [21] present a work based on distance measure of adjectives found in whole document with known polarity i.e. excellent or poor. The author presents a three step algorithm i.e. in the first step; the adjectives are extracted along with a word that provides appropriate information. Second step, the semantic orientation is captured by measuring the distance from words of known polarity. Third step, the algorithm counts the average semantic orientation for all word pairs and classifies a review as recommended or not. In contrast, Pang et al. [7] present a work based on classic topic classification techniques. The proposed approach aims to test whether a selected group of machine learning algorithms can produce good result when opinion mining is perceived as document level, associated with two topics: positive and negative. He present the results using naive bayes, maximum entropy and support vector machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets. Jatinder Saini [18] tried to find the polarity of unsolicited bulk mails through sentiment analysis. In this experiment, Sentiment-depicting words in the whole document are analyzed, scaled and extremes of positive and negative opinions are identified. It has been found that for almost 50% of cases, the opinions expressed through such UBE have positive polarity; almost 30% cases are negatively opined whereas almost 20% cases contain neutral opinion.

B. Mining Opinion at Sentence level

In the sentence level sentiment analysis, the polarity of each sentence is calculated. The same document level classification methods can be applied to the sentence level classification problem. Objective and subjective sentences must be found out. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes. In case of simple sentences, single sentence bears a single opinion about an entity. But there will be complex sentences also in the opinionated text.

The sentence level opinion mining is associated with two tasks [4] [6]. First one is to identify whether the given sentence is subjective (opinionated) or objective. The second one is to find opinion of an opinionated sentence as positive, negative or neutral. The assumption is taken at sentence level is that a sentence contain only one opinion. Riloff and Wiebe [12] use a method called bootstrap approach to identify the subjective sentences and achieve the result around 90% accuracy during their tests. In contrast, Yu and Hatzivassiloglou [14] talk...
about sentence classification and orientation. For the sentence classification, authors present three different algorithms: (1) sentence similarity detection, (2) naive Bayes classification and (3) multiple naive Bayes classification. For opinion orientation authors use a technique similar to the one used by Turney [21] for document level. Wilson et al. [28] pointed out that not only a single sentence may contain multiple opinions, but they also have both subjective and factual clauses. Like the document-level opinion mining, the sentence-level opinion mining does not consider about object features that have been commented in a sentence. For this the feature level opinion mining is discuss in the next sub-section.

C. Mining Opinion at Feature level

The task of opinion mining at feature level is to extracting the features of object that have been commented and after that determine the opinion of the object i.e. positive or negative and then group the feature synonyms and produce the summary report. Liu [5] used supervised pattern learning method to extract the object features for identification of opinion orientation. To identify the orientation of opinion he used lexicon based approach. This approach basically uses opinion words and phrase in a sentence to determine the opinion. The working of lexicon based approach [31] is explained in 3 steps: Identification of opinion words, Role of Negation words and But-clauses. In contrast, Hu and Liu do customer review analysis [20] through opinion mining based on feature frequency, in which the most frequent features is accepted by processing many reviews that are taken during summary generation. In opposite to Hu and Liu, Popescu and Eitzioni [1], improved the frequency based approach by introducing the part-of relationship and remove the frequent occurring of noun phrases that may not be features.

V. Approaches for Classifying Sentiments and mining Opinions

In relation to sentiment analysis, the literature survey done indicates two types of techniques –Supervised and Unsupervised learning.

A. Supervised learning

In a machine learning based classification, two sets of data are required: training and a test set. A training set is used is an automatic classifier to learn the differentiating characteristics of data, 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 well-known machine learning methods in the natural language processing area are K-Nearest neighborhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model.

Naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem. The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The Naive Bayes algorithm is widely used algorithm for document classification [22][24][26][33]. The support vector machine is a statistical classification method proposed by Vapnik [30]. SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Support vector machines (SVM), a discriminative classifier is considered the best text classification method [17][24][26][33]. The idea behind the centroid classification algorithm is extremely simple and straightforward [26]. Initially the prototype vector or centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is computed, finally based on these similarities, document is assigned to the class corresponding to the most similar centroid. The K-nearest neighbor (KNN) is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document. Given a test document \( d \), the system finds the \( k \) nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document [26]. Winnow is a well-known online mistaken-driven method. It works by updating its weights in a sequence of trials. On each trial, it first makes a prediction for one document and then receives feedback; if a mistake is made, it updates its weight vector using the document. During the training phase, with a collection of training data, this process is repeated several times by iterating on the data [26]. Besides these classifiers other classifiers like ID3 and C5 are also investigated [23].

In all supervised approaches, reasonably high accuracy can be obtained subject only to the requirement that test data should be similar to training data. To move a supervised sentiment classifier to another domain would require collecting annotated data in the new domain and retraining the classifier. This dependency on annotated training data is one major shortcoming of all supervised methods.

B. Unsupervised Learning:

All approaches previously described build upon a set of fully annotated data, which is used to train a classifier with one technique or another. This classifier is then used to classify novel incoming text. Unsupervised approaches to sentiment classification can solve the problem of domain dependency and reduce the need for annotated training data. This approach is also known as semantic orientation approach to Sentiment analysis b’coz it measures how far a word is inclined towards positive and negative, Instead of prior training in order to mine data. Much of the research in unsupervised sentiment classification makes use of lexical resources available.
An unsupervised learning algorithm was proposed by Turney [21], which uses semantic orientation of phrases. Zagibalov and Carroll [27] describe a method of automatic seed word selection for unsupervised sentiment classification of product reviews in Chinese. The results obtained are close to those of supervised classifiers and sometimes better, up to an F1 score of 92%. Kamps et al [19] focused on the use of lexical relations in sentiment classification. Andrea Esuli and Fabrizio Sebastiani [2] proposed semi-supervised learning method started from expanding an initial seed set using WordNet. They determined the expanded seed term’s semantic orientation through gloss classification by statistical technique. Takamura et al. [15] determined term orientation according to a “spin model”. An approach was proposed by Chunxu Wu [11] to handle the reviews for which contextual information to determine the orientation of opinion is not enough.

An unsupervised learning algorithm by extracting the sentiment phrases of each review by rules of part-of-speech (POS) patterns was investigated by Ting-Chun Peng and Chia-Chun Shih [29]. They consider only opinionated sentences containing at least one detected sentiment phrase for opinion extraction. Gang Li and Fei Liu [13] developed an approach based on the k-means clustering algorithm. The technique of TF-IDF (term frequency – inverse document frequency) weighting is applied on the raw data. The result is obtained based on multiple implementations of the clustering process. Finally, the term score is used to further enhance the clustering result. Chaovalit and Zhou in [9] compared the Semantic Orientation approach with the N-gram model machine learning approach by applying to movie reviews. They confirmed from the results that the machine learning approach is more accurate but requires a significant amount of time to train the model. In comparison, the semantic orientation approach is slightly less accurate but is more efficient to use in real-time applications. The performance of semantic orientation also relies on the performance of the underlying POS tagger.

VI. Evaluation Metrics

The performance of sentiment classification system can be evaluated by using four metrics: Accuracy, Precision, Recall and F1 measure. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives.

\[
\text{Precision} = \frac{\text{No. of correct extracted opinions}}{\text{Total no. of extracted opinions}}
\]

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives.

\[
\text{Recall} = \frac{\text{No. of correct extracted opinions}}{\text{Total number of annotated opinions}}
\]

Precision and recall can be combined to produce a single metric known as F-measure, which is the weighted harmonic mean of precision and recall. The main advantage of using F-measure is it is able to rate a system with one unique rating.

\[
F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

Accuracy measures the overall degree to which instances have been correctly classified, using the formula as defined below.

\[
\text{Accuracy} = \frac{\text{No. of correctly classified instances}}{\text{Total no. of instances}}
\]

VII. Tools used in Opinion Mining

The tools which are used to track the opinion or polarity from text are [25]: WEKA (Weka is a collection of machine learning algorithms for data mining tasks), Rapidminer (provides software, solutions, and services in the fields of predictive analytics, data mining, and text mining), GATE (General Architecture for Text Engineering is NLP and language engineering tool), NLTK (Natural Language Toolkit), Red Opal, Opinion observer, Review Seer tool, Ling Pipe, OpenNLP and Stanford Parser and Part-of-Speech (POS) Tagger.

VIII. Conclusion

Sentiment classification has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. This paper focuses on the frame work on opinion mining and survey on some of the tasks which have been done in each phases. It is observed that performances of sentiment classifiers are severely dependent on domains or topics. This study shows that supervised approach works well for sentiment classification analysis of movie domain, product, and hotel etc., where as lexicon based approach is appropriate for web content mining such as short text in micro-blogs, tweets, and Facebook status.

IX. References
