IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498, Vol.2, No. 2, April 2012

Sentimental Classification of Text Data

G. Uma Maheswari
School of Information Technology & Engineering
VIT University, Vellore

Abstract—**Sentimental classification is a sub domain Sentiment classification is a sub domain of the natural language processing. In general, sentiment classification means to analyze whether the author holds positive or negative sentiment to a specific subject .Sentiment classification can be used in various applications. It can be used in business to summarize the feedback of their customers. For example, we can use it to classify the favorable or unfavorable reviews towards the products.**

Sentiment classification is a recent sub discipline of text classification which is concerned not with the content a document is about, but with the opinion it expresses. It has a rich set of applications, ranging from tracking users opinions about products or about political candidates as expressed in online forums, to customer relationship management. Functional to the extraction of opinions from text is the determination of the orientation of ``subjective'' terms contained in text, i.e. the determination of whether a term that carries opinionated content has a positive or a negative suggestion. The main idea of the project is to classify the sentiments in the given text data either Positive or Negative. Two types of approaches are used to classify the sentiments. First approach is by using Term counting method and the second approach is by using Machine Learning Algorithm – Support vector machine. To get more accuracy overstatements and understatements will also be weighted.

Keywords-Sentimental Classification, Favourable review, Unfavourable review, Term counting method, Machine learning algorithm, Support Vector machine.

I. INTRODUCTION *(SENTIMENTAL CLASSIFICATION)*

Sentiment classification is the task of labeling a review document according to the polarity of its frequent opinion (favorable or unfavorable). Documents can be categorized in various ways, for example by subject, kind of literacy, or the sentiment expressed in the document. We focus on sentiment classification (into positive or negative opinions). One useful application of sentiment classification is in question answering. Two approaches to classify sentiments are compared in this paper.

The first approach is to count positive and negative terms in the text, where the text is considered positive if it contains more positive than negative terms, and negative if there are more negative terms. A text is neutral if it contains an equal number of positive and negative terms. Instead of having a strict equality for neutral reviews, we can allow a margin of several terms. Positive and negative terms are initially taken from the General Inquirer GI is a dictionary that contains information about English word senses, including tags that label them as positive, negative, negation, overstatement understatement.

An enhanced term-counting method also takes contextual valence shifters into account. Valence shifters are terms that can change the semantic orientation of another term, for example they make a positive term become negative. Examples of negation terms are not, never, none, nobody. There are many other factors that affect whether a particular term is positive or negative, depending on how it is used in a sentence. However we do not address all of them. Terms that change the intensity of a positive or negative term are also examined. These terms increase or decrease the weight of a positive or negative term. We also add positive and negative terms from several other sources and test their contribution to the accuracy of the classification.

The second approach uses Machine Learning (ML) in order to determine the sentiment of the text data. We trained Support Vector Machine classifiers that use unigrams (single words) as features. An enhanced version of this method uses as features, in addition to unigrams, some specific bigrams. We selected only bigrams that contain a combination of a negation, intensifier, or diminishers with another feature word. Rather than having bigrams such as very good where very is an intensifier, we identify the bigrams as int good where int indicates any intensifier. There are similar features for diminishers and negations. Methods based on Machine Learning are much more effective in terms of the accuracy of classification. By combining the two methods we are able to improve the results over either of the method alone.

II. RELATED WORK

One of the first approaches in the sentiment classification is term counting method which concentrates on manual calculation of positive and negative terms. Many papers have also concentrated on the same method but only on specific areas like product review, movie review etc. But for classifying the general text; it has been done only in the manual way right now. The following papers explain in detail about various types of reviews for classifying the phrases.

A. Sentimental Classification using phrase patterns

This paper presents a phrase pattern-based method in classifying sentiment orientation of text. That is to analyze whether the text expresses a favorable or Unfavorable sentiment for a specific subject. In this method, they construct some phrase patterns and calculate their sentiment orientation by unsupervised learning algorithm. When they classify a document, they first add special tags to some words in the text, and then match the tags within a sentence with some phrase patterns to get the sentiment orientation of the sentence. At last,

they add up the sentiment orientation of each sentence. We classify the text according to this summation. The method achieves an accuracy rate of 86% when used to evaluate sports reviews from some websites.

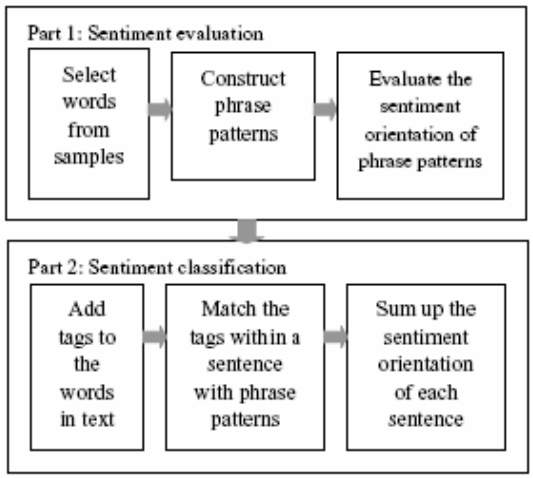

Figure 1. Sentimental Classifier

For examples,

The team is slack these days. …….. (1)

Millard is playing starring roles. …… (2)

In the sentence (1), team is the subject to review and slack is an adjective to describe the subject team. They are put together to show that the state of the team is poor. The sentiment of this sentence is negative. In the sentence (2), Millard is the subject to review, and starring is the word to describe the verb playing. These three words are put together to express favorable sentiment to the team Millard.

Tags are added to the words manually to generate the phrase patterns. Adding the tags to the words are done only in the manual way and so there is no limitation for the usage of words in the phrase patterns.

With these tags, we can arrange them in pairs or groups to form phrase patterns that are often appeared in sentences.

For example,

Roma was defeated by AC Milan

In the sentence, the word Roma is the subject to review. Its tag is n. The word defeated describes the action of Roma, and its tag is dv. The phrase pattern n+dv appears in the sentence.

In our experiment, we mainly use phrases patterns that are composed of the subjects to review, adjectives, verbs, adverbs and nouns. We also use some prepositions and conjunctions. The phrase patterns we choose are all often used.

TABLE I. TAGS ADDED TO WORDS

Type	Tag	Example
Subject to review	n	team
Positive adjective	aj	good
Negative adjective	dj	bad
Positive adverb	ad	perfectly
Negative adverb	dd	poorly
Positive noun	an	star
Negative noun	dn	garbage
Positive verb	av	achieve
Negative verb	dv	frustrate
Special Prep.	np	without
Special Conj.	nc	however

Sentiment orientation will be obtained from these phrase patterns.

TABLE II. PHRASE PATTERNS

n+aj/dj/an/dn
n+av/dv+aj/bj/an/dn
n+ad/dd+av/dv+aj/bj/an/dn
av/bv+n
ad/dd+av/bv+n
n+an/dn
n+av/dv+np+an/dn

From the samples, we also find that in some phrases, conjunctions such as however, yet and but is also very important in expressing sentiment. These words often change the sentiment into the opposition orientation.

For example, Rainier's side is hot favorites but have no hope. There is the word but in this sentence. If we only consider the words favorites and Rainier, we will mistake the sentiment for positive. However, the word but in the sentence changes its sentiment orientation, actually it is negative. These words may change the sentiment orientation too. At present, the words with sentiment orientation are selected manually, and the phrase patterns are also created manually. It takes a lot of time and the result is not precise enough. So we will also try to find a way to get the words and phrase patterns with machine learning method to improve the objectivity and precision.

TABLE III. STATISTIC OF PHRASE PATTERNS

Phrase pattern	Appear in positive reviews	Appeared in negative reviews	Sentiment orientation
n+aj	108	24	1.504
n+dj	37	119	-1.168
n+an	67	31	0.7707
n+dn	19	33	-0.5521
n+av+aj	9	0	3.802
n+dv+dj	10	72	-1.974
n+av+np+an	1	1	0
n+av+np+dn	1	7	-1.9459
n+dv+np+an	2	1	0.6931
n+dv+np+dn	0	9	-4.3026
av+n	52	15	1.2432
dv+n	0	19	-4.9957
ad+av+n	2	0	3.0986
dd+dv+n	0	4	-3.6094
n+ad+av+aj	6	0	3.9459
n+dd+av+aj	1	11	-2.3979
n+ad+dv+aj	12	6	0.6931
n+ad+dv+aj	8	0	4.1972
n+ad+dv+an	2	20	-2.3026
n+ad+dv+dn	6	12	-0.6931
n+dd+dv+aj	0	16	-4.8332

B. Sentimental Classification of Movie Reviews Using Contextual Valence Shifters

The author has presented two methods for determining the sentiment expressed by a movie review. The semantic orientation of a review can be positive, negative, or neutral. They examine the effect of valence shifters on classifying the reviews. They examine three types of valence shifters: negations, intensifiers and Negations are used to reverse the semantic polarity of a particular term, while intensifiers and diminishers are used to increase and decrease, respectively, the degree to which a term is positive or negative. The first method classifies reviews based on the number of positive and negative terms they contain. They use the General Inquirer in order to identify positive and negative terms, as well as negation terms, intensifiers, and diminishers. We also use positive and negative terms from other sources, including a dictionary of synonym differences and a very large Web corpus. To compute corpus-based semantic orientation values of terms, we use their association scores with a small group of positive and negative terms. They show that extending the term-counting method with contextual valence shifters improves the accuracy of the classification. The second method uses a Machine Learning algorithm, Support Vector Machines. They started with unigram features and then added bigrams that consist of a valence shifter and another word. The accuracy of classification is very high, and the valence shifter bigrams slightly improve it. The features that contribute to the high accuracy are the words in the lists of positive and negative terms. Previous work focused on either the term-counting method or the Machine Learning method. They show that combining the two methods achieves better results than either method alone. The positive and negative terms may not all be equally positive or negative. Positive and negative terms can 20 be given weights to show just how positive or negative they are Overstatements and understatements could also be weighted.

C. Proposed System

The main idea of the project is to classify the sentiments in the given text data either positive or Negative. Two types of approaches are used to classify the sentiments. First approach is by using Term counting method. Second approach is by using Machine Learning Algorithms – Support vector machine or Bayesian classification. In the previous work they have not concentrated on more positive and more negative statements. So to overcome and to get more accuracy overstatements and understatements will be weighted in the current work.

D. Conclusion

This paper present, the words with sentiment orientation is selected manually, and the phrase patterns are also created manually. If the words we selected are excessive, it causes some phrase patterns to appear in positive samples negative samples nearly with the same probability. In such condition, Tai and Tdi are nearly the same. The value of W_i is very low. It is hard to distinguish whether a phrase pattern Expresses positive or negative expression.Therefore some phrase patterns are valid only in part of the test samples. It takes a lot of time and the result is not precise enough. So we will also try to find a way to get the words and phrase patterns with machine learning method to improve the objectivity and precision.

REFERENCES

[1] Zhongchao Fei, Jian Liu, Gengfeng Wu, Sentiment classification using phrase patterns, Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), 2004 IEEE.

[2] Alistair Kennedy and Diana Inkpen, University of Ottawa, Sentiment Classification of Movie Reviews Using Contextual Valence Shifters, 2005.

[3] Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens,Automatic Sentiment Analysis in On-line Text, Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007

[4] Artificial Intelligence by Patrick Henry Winston, 3rd edition Pearson Education Asia.

[5] Artificial Intelligence a modern approach by Stuart Russell and Peter Norving, 2nd edition, Pearson Education Asia.

[6] Introduction to AI and Expert Systems by Dan W Patterson, Prentice Hall of India, 1990.

[7] Artificial Intelligence by Knight, Tata McGraw-Hill, 2nd edition.

[8] www.ualberta.ca

[9] www.sports.yahoo.com