Journal of Advanced Engineering Research
ISSN: 2393-8447
Volume 2, Issue 2, 2015, pp.69-73

Analyzing Variations of Opinions on Twitter

R. Nisha Pauline

Department of Computer Science, Auxilium College, Vellore 632 006, India

*Corresponding author email: nishapauline@gmail.com, Phone: +91 94874 87401

ABSTRACT

In today's era of technical enhancement, sharing one's outlook on a social network has become inevitable. Enormous views are being blogged each day on Twitter. Based on these views prospective decisions can be made exceptionally. Sentiment Analysis on Twitter helps to analyze the open views of the public and to take major decisions on various domains. The data repository in Twitter is large enough to support the sentiment classification. The positive and the negative views are collectively considered for the analysis of the products. Interactions on Twitter on timely refresh provide a good platform for this analysis of sentiment variations. Public sentiments are evaluated in every sector to enforce actual judgments. The public view, sentiments and opinions are classified using various machine learning algorithms like Support Vector Machines and naive Bayes through Natural Language Processing (NLP). The objective of this paper is to discover the concept of Sentiment Analysis and Sentiment Variations in the field of Natural Language Processing and to study the various techniques in this field.

Keywords- Public sentiment, Sentiment Analysis, Sentiment Classification, Sentiment Variations, Twitter.

1. INTRODUCTION

Users can post their opinion on Twitter, to tell others what they are doing, what they are thinking, or what is happening around them. Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". According to the latest Twitter entry in Wikipedia, the number of Twitter users has climbed to 190 million and the number of tweets published on Twitter every day is over 65 million [1].

Enormous opinions are found in social media including news, forums, product reviews and blogs. Opinion is defined as "a personal belief or judgment that is not found on proof or certainty". NLP systems such as review summarization systems, dialogue systems and public media analysis systems are used for automated identification of diverse opinions which can be constructive in this field. Sometimes it is directly requested by the user to obtain articles or sentences with a certain opinion. Some social networks like Twitter allow users to add opinions to articles [2]. Some of these opinions are sentiment tags which assign one or more opinion to a tweet. In this paper, we propose a way to utilize such tagged Twitter data for classification of a wide variety of opinions from text.

2. THE PROBLEM OF OPINION MINING

'The opinion mining problem, which enables us to see a structure from the intimidating unstructured text and to provide a unified framework for the current research is defined. The abstraction consists of two parts: opinion definition and opinion summarization [3]. There are two main types of opinions: regular opinions and comparative opinions. Regular opinions are often referred to simply as opinions in the research literature. A comparative opinion expresses a relation of similarities or differences between two or more entities, and/or a preference of the opinion holder based on some of the shared aspects of the entities [4, 5]. A comparative opinion is usually expressed using the comparative or superlative form of an adjective or adverb, although not always.

3. CLASSIFICATION FEATURES

We utilize four basic feature types for opinion classification: single word features, n-gram features, pattern features and punctuation features. For the classification, all feature types are combined into a single feature vector [2].

3.1 Word-based and n-gram based feature

Each word appearing in a sentence serves as a binary feature with weight equal to the inverted count of this word in the Twitter corpus.

3.2 Pattern-based features

Our main feature type is based on surface patterns. For automated extraction of patterns, we followed the pattern definitions given in [6]. We classified words into high-frequency words (HFWs) and content words

(CWs). A word whose corpus frequency is more (less) than FH (FC) is considered to be a HFW (CW). We estimate word frequency from the training set rather than from an external corpus.

3.3 Efficiency of feature selection

Since we avoid selection of textual features which have a training set frequency below 0.5%, we perform feature selection incrementally, on each stage using the frequencies of the features obtained during the previous stages. Thus first we estimate the frequencies of single words in the training set, then we only consider creation of n-grams from single words with sufficient frequency, finally we only consider patterns composed from sufficiently frequent words and n-grams.

3.4 Punctuation-based features

In addition to pattern-based features, we used the following generic features: (i) Sentence length in words, (ii) Number of "!" characters in the sentence, (iii) Number of "?" characters in the sentence, (iv) Number of quotes in the sentence, and (v) Number of capitalized/all capitals words in the sentence.

4. SENTIMENT ANALYSIS (SA)

In recent years, Sentiment Analysis (SA) has become a hot topic in the NLP research community. A lot of papers have been published on this topic.

4.1 Target-independent SA

Specifically, [21] proposes an unsupervised method for classifying product or movie reviews as positive or negative. In this method, sentimental phrases are first selected from the reviews according to predefined part-of-speech patterns. Then the semantic orientation score of each phrase is calculated according to the mutual information values between the phrase and two predefined seed words. Finally, a review is classified based on the average semantic orientation of the sentimental phrases in the review. In contrast, [7] treat the sentiment classification of movie reviews simply as a special case of a topic-based text categorization problem and investigate three classification algorithms: Naive Bayes, Maximum Entropy, and Support Vector Machines. According to the experimental results, machine learning based classifiers outperform the unsupervised approach, where the best performance is achieved by the SVM classifier with unigram presences as features.

4.2 Target-dependent SA

Besides the above mentioned work for target independent sentiment classification, there are also several approaches proposed for target-dependent classification, such as [3, 8, 9]. [8] adopts a rule based approach, where rules are created by humans for adjectives, verbs, nouns, and so on. Given a sentiment target and its context, part-of-speech tagging and dependency parsing are first performed on the context. Then predefined rules are matched in the context to determine the sentiment about the target.

In [10], opinions are extracted from product reviews, where the features of the product are considered opinion targets. The sentiment about each target in each sentence of the review is determined based on the dominant orientation of the opinion words appearing in the sentence. As mentioned, target-dependent sentiment classification of review sentences is quite different from that of tweets. In reviews, if any sentiment is expressed in a sentence containing a feature, it is very likely that the sentiment is about the feature. However, the assumption does not hold in tweets.

4.3 SA of Tweets

As Twitter becomes more popular, opinion analysis on Twitter data becomes more attractive. [2, 11, 12] all follow the machine learning based approach for sentiment classification of tweets. Specifically, Dmitry Davidov *et al.* [2] proposes to classify tweets into multiple sentiment types using hashtags and smileys as labels. In their approach, a supervised KNN-like classifier is used. In contrast, [13] proposes a two-step approach to classify the sentiments of tweets using SVM classifiers with abstract features. The training data is collected from the outputs of three existing Twitter sentiment classification web sites. As mentioned above, these approaches work in a target-independent way, and so need to be adapted for target-dependent sentiment classification.

5. CLASSIFICATION BASED ON SUPERVISED LEARNING

Any existing supervised learning methods can be applied to sentiment classification, e.g., naive Bayesian classification, and support vector machines (SVM). [7] took this approach to classify movie reviews into two classes, positive and negative. It was shown that using unigrams (a bag of individual words) as features in classification performed well with either naive Bayesian or SVM. Subsequent research used many more features and techniques in learning. As most machine learning applications, the main task of sentiment classification is to engineer an effective set of features.

Terms and their frequency: These features are individual words or word n-grams and their frequency counts. In some cases, word positions may also be considered. The TF-IDF weighting scheme from information retrieval may be applied too. These features have been shown quite effective in sentiment classification.

Part of speech: It was found in many researches that adjectives are important indicators of opinions. Thus, adjectives have been treated as special features.

Opinion words and phrases: Opinion words are words that are commonly used to express positive or negative sentiments. For example, beautiful, wonderful, good, and amazing are positive opinion words, and bad, poor, and terrible are negative opinion words. Although many opinion words are adjectives and adverbs, nouns (e.g., rubbish, junk, and crap) and verbs (e.g., hate and like) can also indicate opinions. Apart from individual words, there are also opinion phrases and idioms, e.g., cost someone an arm and a leg. Opinion words and phrases are instrumental to sentiment analysis for obvious reasons.

Negations: Clearly, negation words are important because their appearances often change the opinion orientation. For example, the sentence “I don’t like this camera” is negative. However, negation words must be handled with care because not all occurrences

For example, [7] compares Naive Bayes, Support Vector Machines, and maximum-entropy-based classification on the sentiment-polarity classification problem for movie reviews. More extensive comparisons of the performance of standard machine learning techniques with other types of features or feature selection schemes have been engaged in in later work [14-16]. We note that there has been some research that explicitly considers regression or ordinal-regression formulations of opinion mining problems [17-19]: example questions include, “how positive is this text?” and “how strongly held is this opinion?”.

6. CLASSIFICATION BASED ON UNSUPERVISED LEARNING

It is not hard to imagine that opinion words and phrases are the dominating indicators for sentiment classification. Thus, using unsupervised learning based on such words and phrases would be quite natural. For example, the method in [20] uses known opinion words for classification, while [21] defines some phrases which are likely to be opinionated. Be low, we give a description of the algorithm in [21], which consists of three steps:

Step 1: It extracts phrases containing adjectives or adverbs as adjectives and adverbs are good indicators of opinions. However, although an isolated adjective may indicate opinion, there may be insufficient context to determine its opinion orientation (called semantic orientation in [21]). For example, the adjective “unpredictable” may have a negative orientation in an automotive review, in such a phrase as “unpredictable steering”, but it could have a positive orientation in a movie review, in a phrase such as “unpredictable plot”. Therefore, the algorithm extracts two consecutive words, where one member of the pair is an adjective or adverb, and the other is a context word.

Two consecutive words are extracted if their POS tags conform to any of the patterns in Table 1.1. For example, the pattern in line 2 means that two consecutive words are extracted if the first word is an adverb and the second word is an adjective, but the third word cannot be a noun. NNP and NNPS are avoided so that the names of entities in the review cannot influence the classification.

Example: In the sentence “This camera produces beautiful pictures”, “beautiful pictures” will be extracted as it satisfies the first pattern.

Step 2: It estimates the semantic orientation of the extracted phrases using the Point wise Mutual Information (PMI) measure given in Equation 1:

$$\text{PMI}(\text{term1}, \text{term2}) = \log_2 \frac{\Pr(\text{term1} \wedge \text{term2})}{\Pr(\text{term1}) \cdot \Pr(\text{term2})} \quad \ldots\ldots(1)$$

Here, $\Pr(\text{term1} \wedge \text{term2})$ is the co-occurrence probability of term1 and term2, and $\Pr(\text{term1}) \bullet \Pr(\text{term2})$ gives the probability that the two terms co-occur if they are statistically independent. The ratio between $\Pr(\text{term1} \wedge \text{term2})$ and $\Pr(\text{term1}) \bullet \Pr(\text{term2})$ is thus a measure of the degree of statistical dependence between them. The log of this ratio is the amount of information that we acquire about the presence of one of the words when we observe the other. The semantic/opinion orientation (SO) of a phrase is computed based on its association with the positive reference word “excellent” and its association with the negative reference word “poor”:

SO (phrase) = PMI (phrase, “excellent”) - PMI (phrase, “poor”)

……(2)

Table 1 Patterns of tags for extracting two-word phrases

S. No.	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR or RBS	VB, VBD, VBN OR VBG	anything

The probabilities are calculated by issuing queries to a search engine and collecting the number of hits. For each search query, a search engine usually gives the number of relevant documents to the query, which is the number of hits. Thus, by searching the two terms together and separately, we can estimate the probabilities in Equation 1. The author of [22] used the AltaVista search engine because it has a NEAR operator, which constrains the search to documents that contain the words within ten words of one another in either order. Let hits (query) be the number of hits returned. Equation 2 can be rewritten as follows:

SO (phrase)

$$= \log_2 \left[\frac{\text{hits(phrase NEAR “excellent”)hits(“poor”)}}{\text{hits(phrase NEAR “poor”)hits(“excellent”)}} \right] \quad \ldots(3)$$

To avoid division by 0, 0.01 is added to the hits.

Step 3: Given a review, the algorithm computes the average SO of all phrases in the review, and classifies the review as recommended if the average SO is positive, not recommended otherwise.

Final classification accuracies on reviews from various domains range from 84% for automobile reviews to 66% for movie reviews.

To summarize, we can see that the main advantage of document level sentiment classification is that it provides a prevailing opinion on an entity, topic or event. The main shortcomings are that it does not give details on what people liked and/or disliked and it is not easily applicable to non-reviews, e.g., forum and blog postings, because many such postings evaluate multiple entities and compare them.

7. CONCLUSION

Twitter sentiment analysis has attracted much attention recently. In this paper, we address target dependent sentiment classification of tweets. Different from previous work using target independent classification, we propose to incorporate syntactic features to distinguish texts used for expressing sentiments towards different targets in a tweet. According to the experimental results, the classifiers incorporating target-dependent features significantly outperform the previous target independent classifiers.

Our goal in this survey has been to cover techniques and approaches that promise to directly enable opinion-oriented information-seeking systems, and to convey to the reader a sense of our excitement about the intellectual richness and breadth of the area. We very much encourage the reader to take up the many open challenges that remain, and hope we have provided some resources that will prove helpful in this regard.

REFERENCE

[1] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, *Target-dependent Twitter Sentiment Classification,* Harbin Institute of Technology, Harbin, China.

[2] Dmitry Davidov, Oren Tsur, Ari Rappoport, *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*. 2010.

[3] M. Hu, B. Liu, Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[4] N. Jindal, B. Liu, Identifying comparative sentences in text documents. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2006),* 2006.

[5] N. Jindal, B. Liu, Mining comparative sentences and relations. In *Proceedings of National Conference on Artificial Intelligence (AAAI - 2006),* 2006.

[6] D. Davidov, A. Rappoport, *Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words*. 2006

[7] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), 2002*.

[8] T. Nasukawa, J. Yi, Sentiment analysis: Capturing favorability using natural language processing. *In: Proceedings of the Conference on Knowledge Capture, K-CA*, 2003.

[9] X. Ding, B. Liu, P. Yu, A holistic lexicon-based approach to opinion mining. *In: Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 231– 240, 2008.

[10] M. Hu, B. Liu, Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[11] Alec Go, Richa Bhayani, Lei Huang, *Twitter sentiment classification using distant supervision.* Technical report, Stanford. 2009.

[12] R. Parikh, M. Movassate, *Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques.* 2009.

[13] Luciano Barbosa, Junlan Feng, *Robust sentiment detection on twitter from biased and noisy data.* In COLING (Posters). 2010.

[14] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, pages 519–528, 2003.

[15] M. Gamon, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

[16] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura, Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2005.

[17] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 115 – 124, 2005.

[18] Theresa Wilson, Janyce Wiebe, Rebecca Hwa, Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI*, pages 761–769, 2004. Extended version in *Computational Intelligence* 22(2, Special Issue on Sentiment Analysis):73–99, 2006.

[19] K. Sundharivel, D. Udaya Suriya Rajkumar, Real Time Nature Disasters of Data Analysis Using Tweet, Journal of Advanced Engineering Research, 1 (1), 2014, 41-47.

[20] M. Taboada, J, Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis. *Computational Intelligence*, 2010.

[21] P. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, 2002.

[22] N. Jindal, B. Liu. Identifying comparative sentences in text documents. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2006)*, 2006.