Mining Sentiments and Sequential Rules for Event Prediction

1P.G. Preethi and 2V. Uma

1M.Tech. Student Department of Computer Science, Pondicherry University, Puducherry, 605014, India
2Assistant Professor, Department of Computer Science, Pondicherry University, Puducherry, 605014, India

ABSTRACT

Sentiment analysis or opinion mining identifies and classifies opinions in text by determining the polarity as positive, negative or neutral. In Social media like Twitter, people express their opinions. Analysis of tweets done using the aspect keywords helps in identifying the sentiments over time. The temporal information involved in the tweets can be mined and sequential rules can be generated. Sequential rules aid in defining the order of occurrence of events. The sequential rule discovery is an effective method to predict the sequence of events and sentiments of upcoming events. In this work, mining of sentiments and temporal relations between events is done and sequential rules are discovered to predict the event order and the sentiments of the future events. The sentiment analysis phase is evaluated using the performance measures precision and recall. The accuracy of sequential rule prediction is calculated using the measure Percent_Similarity.

© 2015 AENSI Publisher All rights reserved.


INTRODUCTION

Social media can be considered as a reflection of real world and can be used to analyze the opinions of current society. Social media is a combination of traditional media like newspaper and non-traditional media such as Facebook, Twitter etc. Popularity of social media continues to grow exponentially every second.

Large quantity of user data is generated in social media. This trend will continue in future also and thus result in large quantities of user data. Since social media contains large amount of user data, it is important for producers, consumers, and service providers to manage and utilize user-generated data. Social media can indirectly become a useful source of user information and helps in identifying different ways to improve the user experience (P. Gundecha et al., 2012). Finding these details from social media is a big challenge and these details can provide an opportunity for data miners to develop different algorithms and methods to leverage social media data. Social media mining can be used in various fields such as business, politics, medicine and education.

Social media sentiment analysis involves analyzing opinions that appear in blog posts, comments, reviews or tweets. The field of sentimental analysis or opinion mining analyses opinions, sentiments, evaluations, attitudes, and emotions towards some entities such as goods, facilities, groups, individuals, actions etc. Liu (2012). There are many challenges in sentimental analysis. First, the opinion word may be considered positive in one situation and negative in another and it depends on the context. Second, people express their opinion in different ways in different situations. The third being, most reviews contain both positive and negative words and hence analyzing and categorizing it in either positive or negative light is a difficult task Vinodhini et al. (2012).

Sentiment analysis is useful in several fields like marketing, advertisement, politics, education, medicine etc. Nowadays many people use social media for sharing their opinions and thoughts about individuals, companies, products, movements, important events etc. Mining social media can help in prediction of events.

Temporal expressions are used in describing events and the temporal relation between events assist in temporal ordering of events. The temporal order of events helps in identifying the sequential relation between events and thus help in predicting the occurrence of events Mirza (2014).

The proposed work introduces a prediction system based on sentiment analysis to identify the sentiments and uses the temporal information to predict the event sentiment and order of events. The proposed work uses
Naive Bayes method for sentiment classification and the sequential rule is found using support and confidence measures.

The remainder of the paper is structured as follows. Section 2 briefs the survey on sentiment analysis. Section 3 discusses the relation between sentiment analysis and temporal relations. Section 4 discusses about sequential rule discovery and sentiment analysis. Section 5 provides the overview of the proposed system and in Section 6 conclusion and future work are provided.

**Literature Survey:**

This section presents the survey on various methods for sentiment analysis and sentiment classification

**Sentiment Analysis:**

The phrase sentiment analysis first appeared in Nasukawa and Yi. (2003) and the term opinion mining appeared in Hu and Liu (2006). Instead of classifying the entire document as positive or negative Nasukawa and Yi (2003), it concentrates on extracting the sentiment associated with the text by assigning positive or negative polarities for specific subjects in a document.

Research on sentiment and opinion mining started in early 2000's Huettner and Subasic (2001) Turney (2002). Although the research on sentiment analysis started during the early 2000’s, there were some earlier works on sentiment adjectives, subjectivity, viewpoints, and affects Hatzivassiloglou and McKeown (1997) Wiebe et al. (1999)

**Different Levels of Analysis:**

Sentimental analysis can be conducted in Document Level, Sentence Level and Entity and Aspect Level. Document Level sentiment analysis classifies the entire document into either positive or negative Pang et al. (2002). Sentence level classification classifies the sentence into positive, negative or neutral category. In Tan et al. (2011) polarity prediction model for sentence level sentiment classification was introduced. Entity and Aspect level also known as feature level sentimental analysis gives the summary about which feature of a product does user like or dislike Wiebe et al. (1999). Bafna et al. (2013) introduced a feature based summarization method for online products.

**Sentiment Classification Method:**

Different approaches or methods such as supervised learning, unsupervised learning and semi-supervised learning can be used for sentiment classification

**Supervised Learning:**

This is a type of machine learning technique in which the training data includes both input and desired result. Since this method is fast and accurate, many researchers use supervised learning method for sentiment analysis. In supervised learning, any of the supervised learning methods such as Naïve Bayes classification, Support Vector Machines (SVM) can be used. Vaithyanathan et al. (2002), Qiang Ye et al. (2009) and Deng et al.(2014) applied supervised machine learning methods and Naive Bayes for sentiment classification. Advantage of Supervised method is that, by selecting the features the classifier can be trained well and hence classification can be performed efficiently over test data.

**Unsupervised Learning:**

In unsupervised learning method there is no supervision and so system tries to adapt itself to the situation and learns based on various criteria. There are many sentiment analysis works based on unsupervised methods. Turney (2002) introduced a simple unsupervised method for rating the review as positive (thumbs up) or negative (thumbs down). Bagheri1 et al. (2013) concentrated on identifying the aspects and on the basis of the aspects found the overall sentiment. Gaoyan Ou et al. (2014) introduced a new unsupervised approach called Content and Link Unsupervised Sentiment Model (CLUSM) which focused on the problem of micro blog sentiment classification incorporating link information.

**Semi – Supervised Method:**

Semi – supervised learning uses large amount of unlabeled data along with labeled data in constructing classification boundary. In McDonald and Täckström (2011) two simple methods of combining fully labeled and coarsely labeled data for sentence-level sentiment analysis was introduced. To address the problem of insufficient availability of labeled data Zhou et al. (2013) proposed a new semi supervised method called ADN (Active Deep Network). ADN is constructed by Restricted Boltzmann Machines (RBM) with unsupervised learning based on labeled reviews and large number of unlabeled reviews. Lima et al. (2014) proposed a unique system PERSOMA, for prediction of specific personality traits present in groups of Tweets based on the Big Five Model.
The proposed work intents to use twitter data for sentiment analysis. Naive Bayes classifier is used for the sentiment classification. It is a probabilistic model which calculates the probability of tweet belonging to a particular class. In this work, Naive Bayes classifier is used as it has low computational overhead, yet performing well in many NLP tasks, as indicated by Lake (2011).

**Sentiment Analysis and Temporal Relation Extraction:**

The kind of states which change and thus might need to be located in time are referred as events in the present context. The event entities can be represented in different ways such as finite clauses, nonfinite clauses, nominalizations, event-referring nouns, adjectives, adverbial clauses etc. Generally, events are described in newspapers, stories, micro blogs etc. Fukuhara et al. (2007) proposed a method in which the web-blogs and news articles are given as input and two graphs (topic graph and sentiment graph) are produced as its output. Topic graph shows temporal change of topics associated with a sentiment and the sentiment graph shows the temporal change of sentiments associated with a topic. Das et al. (2011) proposed a Machine learning method based on Conditional Random Field (CRF) for identifying the event-event relation by considering sentiment as a feature of an event. As the first step, Ekman’s six basic universal emotions are assigned to each event and then the temporal relations between the events were analyzed in order to track the event sentiments. Ho et al. (2012) proposed an information extraction framework for mining future spatiotemporal events from web. The mining process involves two steps. The recognition phase identifies and resolves toponyms and future temporal patterns. Spatiotemporal disambiguation, de-duplication, pairing, and sentiment classification were performed in the matching phase. By attaching the sentiment tag to each event, this framework can provide more useful future event guidance. Usually event sentiment over time is calculated from web content like tweets, blogs, normal news article sites etc. These methods can easily summarize the events based on the time and overall sentiment. The proposed work uses both sentiment analysis and temporal relations for predicting the events and the order of occurrence of events using twitter.

**Sentiment Analysis and Sequential Rules:**

Sequential rules identify the order of occurrence of events. By identifying the sequential rule prediction can be performed. Jiang1 et al. (2011) discusses the topic of sentiment change analysis on web documents. First phase of this work is mining opinion on a certain topic and second step identifies significant changes to the sentiment of the opinions on that particular topic and possible reasons for such change. This work identified hot events which are possible causes of a sentiment change. This work is closely related to our proposed work but the main differences are that the existing work is implemented in web domain and identifies the sequence between two events with no predictions being made. But the proposed work uses twitter domain and the sequential rule for predicting events and identifies the possible order of events. Siganos et al. (2014) examines the relationship between the daily sentiment and the trading behavior within 20 international markets using Facebook’s Gross National Happiness Index. It was observed that a positive relation existed between sentiment on Facebook and stock market returns whereas a negative sentiment relates trading volume and volatility. Mirza (2014) proposed a system for automatic extraction of two event relation types, i.e. temporal and causal relations, from natural language text. The interaction between temporal and causal relations was investigated in the context of events and an annotation scheme that identified different types of causality between events, techniques for extracting such relations were presented. The analysis of interconnection between temporal and causal relation was done. Smailovic et al. (2013, 2014) collected the tweet talks based on the some stock companies and evaluated the effect of talks on stock market conversation domain while this proposed work is a generalized approach and can be implemented in any domain. The proposed work identifies sequence between events using the twitter conversation. Dehkharghani et al. (2014) proposed a new concept called sentimental causal rule which is a combined approach that uses sentiment analysis and the causal rule detection. They also proposed a methodology for extracting sentimental causal rules from textual data sources. This concept is very useful in several fields like marketing, policy making etc. The work detected sentiment causal rule but no prediction was made. The proposed work includes temporal factor in sentiment sequential rule discovery and uses it for better event order prediction.

**Proposed Work:**

A brief idea about the proposed work is illustrated in this section. The proposed work mainly consists of five phases. The Fig.1 shows the detailed architecture of the proposed system

**Step1: Keyword Extraction:** The first step of the proposed work is extracting aspect keywords from tweets. These aspect keywords are used for sentiment analysis and sequential rule discovery. After extracting the aspect keywords, the next step is identifying the temporal relation between them and analyzing the sentiment of the aspect keywords.
Step 2: Temporal Relation Extraction: The temporal relation is extracted between the identified aspect keywords. The relation extraction is done using the temporal information present in the tweets such as the time of the tweet and temporal relation related keywords like ‘before’.

Step 3: Sentiment Analysis: The main objective of this phase is determining the sentiment of the aspect keyword. The method used for sentiment analysis is to be evaluated using the performance measures Precision and Recall. Precision and Recall can be calculated Q. Ye et al (2009) using Eqns. (1) (2) (3) (4) and accuracy can be calculated using Eqn(5)

\[
\text{Precision (pos)} = \frac{tp}{tp + fp} \quad (1)
\]
\[
\text{Precision (neg)} = \frac{tn}{tn + fn} \quad (2)
\]
\[
\text{Recall (POS)} = \frac{tp}{tp + fn} \quad (3)
\]
\[
\text{Recall (neg)} = \frac{tn}{tn + fp} \quad (4)
\]
\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (5)
\]

Where tp, tn are true positive and negative review for positive polarity prediction and fp, fn are false

Step 4: Sequential rule discovery: The next step is identifying the sequential rule between events using aspect keywords. The sequential rule is detected using Support and Confidence Dehkarghani et al.(2014)using Eqns. (6) and (7).

\[
\text{Support} (I_1) = \frac{\text{Number of tweets that support } I_1}{\text{Number of tweets during time period } T} \quad (6)
\]
\[
\text{Confidence} (I_1 \text{ before } I_2) = \frac{\text{Number of tweets where } I_1 \text{ occurs before } I_2}{\text{Number of tweets in which } I_1 \text{ occurs}} \quad (7)
\]

Where I1 and I2 are aspect keywords denoting events in tweets.

Step 5: Linear Order Prediction: The final step is prediction based on the sequential rule identified in the previous step. Using time based analysis of tweets and sequential rule, the upcoming events and their sentiments are predicted. Accuracy of prediction is evaluated using Percent_Similarity Hall(2004) as performance measure.

\[
\text{Percent_Similarity} = \frac{\text{Length of common event sequence}}{\text{Length of timeline sequence}} \quad (8)
\]

Length of common event sequence represents the length of the sequence that is common in both original time line and the generated sequence.
Discussion and Result:
R language is widely used among statisticians and data miners for data analysis. R language is used for implementing the Naive Bayes classifier and the result of the classification is illustrated in Fig. 2. The aspect keywords with the score value below zero is considered as negative, score value above zero is considered as positive and the score value equal to zero consider as neutral. After obtaining the positive, negative and neutral tweet then remove the neutral tweet and find the Actual and predicted positive and negative value. Positive and negative review for negative polarity prediction nearly 1000 tweets related to the aspect keyword ‘Narendra Modi’ was retrieved by using Twitter API. The Naïve Bayes classifier has been trained and used to classify the sentiments and precision, accuracy can be calculated using the matrix shown in Table 1. The binomial test is an appropriate measure of accuracy. We run a binomial test to assess the confined interval for the result. Binomial test shows that within a 95% confidence interval, the population mean of the percent classification system would be between 85.28% and 94.13%. Accuracy is calculated as 92.6% based on the data available in Table 1. The sequence of Aspect Keyword is found using the tweet time and the support and confidence are used to find the sequential rule.

Fig. 2: Sentiment Classification Result.

Table 1: Precision and Accuracy Evaluation.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>4</td>
<td>23</td>
</tr>
</tbody>
</table>

Conclusion:
In this proposed work, sentiment analysis is done on tweet data and the polarity of sentiments is mined. The temporal relations mined between the aspects keywords representing the events are used in discovering the sequential rules. The sequential rules discovered are useful for event order prediction. Temporal sequence of events helps in identifying the sentiment sequence. Future work is to evaluate the sequential rule discovery system using the given performance measure and prove the efficiency.

REFERENCES
Fukuhara, T., H. Nakagawa and T. Nishida, 2007. ‘Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events.’ in ICWSM.
Thomas Lake, 2011. Twitter Sentiment Analysis, Western Michigan University, Kalamazoo, MI, for client William Fitzgerald, April.