www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9503-9506

A Survey on Sentiment Analysis And Summarization For Prediction

Vikrant Hole[1], Mukta Takalikar[2]

[1] Pune Institute Of Computer Technology
Departmentof Computer Engineering,Pune,India
holevikrant@gmail.com

[2] Pune Institute Of Computer Technology
Departmentof Computer Engineering,Pune,India
muktapict@gmail.com

Abstract: ***Social Networking portals are been widely used by people all over the world, they use it for expressing their views (sentiment) on different issues like products, movie review and also for expressing their opinion on different political parties. These opinion or sentiment can be used for predicting or analyzing the view of people on services which could be beneficial for various companies and political parties to understand their customer and take needful steps for improvising their efficiency. Summarization of tweets will allow understanding hidden events and sentiment with respect to events. In this paper, we firstly discuss the realms which can be predicted with current social media Secondly, we will see how summarization of sentiment with respect to events will allow people of interest to take decision easily.***

Keywords: Micro blogs, Sentiment analysis, Social intelligence, Tweet Summarization

.

1. Introduction

Social media has provided a platform which has given people to create and publish their views.. There are multiple social media websites like, Twitter, Facebook, What's up, tumblr amd many more which shows its high development and huge influence on people. Twitter, Facebook are in the top 10 most-visited websites in the world. Facebook has more than 800 million active users [2], and by March 2011, on Twitter, there were about 140 million information pieces created and transferred daily [3]. There is other specialized social media that are focused on entertainment, sports, finance and politics and Ecommerce.

Today, nearly every person uses social media for expressing their views on different topic. So, analyzing their views and summarizing it to understand and take decision can help the people in finance, product marketing, politics. But there are limitation using these social media data for prediction like sometimes it is necessary to take into consideration the time factor. It is necessary for fast retrieval of data and reach a point to take decision. It has attracted many researcher to this subject. Study of the user data of social networks is one of the current trends of the times. Big data can be used for collection of large and unstructured data which is difficult while using traditional database. Handling the complexity of big data it is possible to use it for understanding data pattern and use it for learning to predict. These prediction can be done by human also, but for efficiency and avoiding any person intervention it is necessary to be done automatically..

2. Technical Background

In this section we will briefly have an overview of sentiment analysis and summarization. We will also understand the different prediction methods.

2.1 Sentiment Analysis

Sentiment Analysis is the detection of attitude "enduring affectivity, colored beliefs, disposition towards object or person." In Analyzing sentiment, the message is broken down into three attributes which are holder(source of sentiment in message) , target (object towards which sentiment targets) and type of attitude(beliefs like love, hate, positive, negative, etc). In sentiment analysis mainly three tasks are done. A) Simplest task – Attitude is positive or negative. B)Complex task is ranking attitude text into from 1 to 5 and C)Advanced task is detection of target ,source or complex type attitude.

Sentiment Analysis Baseline Algorithm given by Pang and Lee includes mainly three steps

A) Tokenization: In tokenization phase the message is segmented for proper analysis. The data that is collected from websites for sentiment analysis contain HTML and XML markup, Twitter markups(names, Hashtags), Capitalization, Numbers should also be handled in these phase.

B) Feature Extraction: When the input data is too large to be processed then transforming the input data into the set of feature is called feature extraction. If the features extracted properly than it is expected that will perform the desired task.

C) Classification: In sentiment analysis we have to identify the class of sentiment. There are different classification techniques some of which are explained in section 2.2.

2.2 Prediction methods (Classification)

In this section, we discuss some methods used in prediction with social media which in actual makes classification in deciding the class of sentiment.

Regression method: In statistics, regression methods analyze relationship between the variables, It typically makes one understand how dependent variable changes when one of its independent variable is varied, such as the text classification. Regression model are of two types namely, linear and non-linear. Regression analysis is mainly used for forecasting and prediction. The performance of model depends on the way the data is generated and how its relation is with regression model. The linear regression model gives best result [25].This is the simplest method used for prediction.

Naive Bayes classifier: Naïve Bayes classifier is a probabilistic classifier which uses Bayes theorem. Based upon the priori probability which is obtained during training is used to build model which provides the formula for posterior probability, to calculate the object belongs to the result classes, and then make decision that posterior probability with maximum probability will give result class.

K-nearest neighbor classifier: K-nearest neighbor classifier is one of unsupervised machine learning algorithm, which cluster the objects having similar properties in one cluster and objects having different properties into different cluster. The similarity between two objects is measured using metrics the Euclidean distance and Manhattan distance. For two entities p={$p1$,$p2$,…,pn} and q={$q1$,$q2$,…,qn} with n-dimensional feature vector, the Euclidean distance is computed as:

$$ED(p,q) = ED(q,p) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

The Manhattan distance is calculated as:

$$MD(p,q) = MD(q,p) = \sum_{i=1}^{n}|p_i - q_i|$$

The object falls into the cluster with minimum distance.

Artificial Neural network: Taking the nature of brain as model artificial neural network learn recognizes from experience. ANN is usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. For example, in a neural network for handwriting recognition, a set of input neurons may be activated by the pixels of an input image representing a letter or digit. The activations of these neurons are then passed on, weighted and transformed by some function determined by the network's designer, to other neurons, etc., until finally an output neuron is activated that determines which character was read. There are three types of learning in ANN, namely supervised, unsupervised and reinforcement learning. Perhaps the greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that 'learns' from observed data. However, using them is not so straightforward, and a relatively good understanding of the underlying theory is essential.

Decision tree: Decision tree learning uses a decision tree as a prediction model which uses observation to make prediction model. There are mainly two types of classification models in decision tree namely tree model in which leaf represent label of class and branches represent feature that lead to label classes, secondly, regression tree where target variable takes continuous values. Decision tree can be used to visually represent decision.

Support vector machine: Support vector machine is supervised learning method that analyzes data and recognize pattern for classification. Given a set of training, each marked as belonging to one of class, SVM build model from training data and assign new example into one class or other. SVM performs both linear and non-linear classification. In non-linear classification SVM uses kernel trick. Support Vector Machines work very well in practice. The user must choose the kernel function and its parameters, but the rest is automatic.

2.3 Summarization

Summarization produces an abridged version of text that contains information that is important or relevant to the user. Summarization outlines or abstracts the document or simplifying text by compressing sentences. In these mainly two things are done summarizing single document to produce abstract and summarizing multiple document to produce series of news or stories on same event or set of web page about some topic. There are mainly two types of summarization firstly, Generic summarization where we summarize the content of document and secondly query focused in which we summarize with respect to into need expressed in user query. There are two ways of summarization firstly, Extractive Summarization which produce summary from phrases or sentences in source document and secondly, Abstractive Summarization where summary express ideas in source document using different words. Baseline summarization algorithm is to take first sentence as it best represent document. In summarization there mainly three stages A) Content Selection- extracting sentences that we need from document. B) Information Ordering- choose the order to place the sentences in summary and C) Sentence Realization- simplification of sentence in summary.

For prediction we need to analyze the huge amount of data and summarize to take decision. Firstly we identify the classes of sentiment and summarize to find different domains and make decision based on analysis.

2.4 Tweet Summarization

In summarizing twitter message, blogs or short messages are too short. So in summarization we generally find topic or salient words/phrases and provide an abstractive type of summarization as in multi-document summarization. We form summary by detecting topic and find important tweets related to those topics. The goal of topic detection is to identify is to identify the events that took place overtime.

2.4.1 Topic Detection

There two approaches for topic detection firstly, Stream Based topic detection, If the volume in the tweet stream rises or changes suddenly then there are chances of new topic. Tweets within surges are used to represent subtopic. Dehong Gao et al[25] used peak area concept to represent tweets within the time span of each surge. Peak area detection algorithm like offline peak area detection algorithm can be used for detecting subtopics. Secondly, Semantic Based topic detection, Twitter collect twits from different geographical areas and different time zone areas. But stream based topic detection approach not

works well here. however, semantic based approach can be used which employ dynamic topic model to capture different subtopics[28]. Latent Dirichlet Allocation (LDA) is also used in semantic based approach[27].

2.4.2 Summary Generation

Once the topic detection is done then the most significant tweets are extracted for each topic. The ranking strategies employed are firstly, Stream Based Summary Generation which require summary to not only contain tweets related to topic but it should cover the whole topic information. Dehong Gao et al [25] calculated importance using crowding endorsement. Secondly, Semantic Based Summary Generation, in this approach the evaluation of tweet was done by finding distribution of words in tweets and then giving score to each tweet. Tweet with highest score was selected in summary.

3. Related Work

3.1 Related Work for sentiment Analysis

Several research works have been carried out for social network analysis and sentiment analysis for deriving mood of people with respect to some product. From tweet's to poll the co-relation between twitter data and gallup polls is nearly 0.804[1]. Sentiment analysis study has been carried over decades. Is review positive or negative ,these classification is done using only words(tokenization), i.e cool is positive and disappointing is negative using polarity[3]. Sentiment tokenization issues like handling HTML and XML markup, Twitter markup(names, Hashtags), Capitalization, Numbers are difficult to handle, however approaches like utilizing n-grams, Part-of-speech tagging have been employed effectively in [1] and [4] for finding the twitter sentiment using the machine learning techniques and other methodologies.

Earlier only words(tokens) were used consider for sentiment analysis, however these issue is solved using extended target based feature[2]. Different sentiment lexicon providing different classes of positive, negative, strong, pronoun, quantifier and many more have been used[6][7]. Sentiwordnet was enhanced lexical resource which automatically annotated to degree of polarity, i.e positive, negative and objectivity/neutral[5]. How likely is each word to appear in each sentiment class using count gives wrong result so using scaled likelihood removes error[8]. Adjective conjoined by 'and' have same polarity while 'but' do not have been implemented using Turney algorithm[9].

In order to find correct sentiment of sentence it is important for finding aspect or attribute target of sentiment. For example, Food was great but Service was bad ,here there is one sentiment for one attribute while other for other one called micro sentiment. There are two approaches for it. First consist of phrases and rules where we find frequent frequency phrases i.e fish, paneer etc then filter these by rules like occur right after sentiment word i.e great fish. Second find aspect in advance and find dataset related to it[29].Handling multi class classifier have been implemented using linear or regression or specialized models like metric labeling..

Johan Bollen used POMS score to establish the sentiment values classified data as positive or negative in moods category using a syntactic, term-based approach, in order to detect sentiment[10]. The success depends more on the relationship of the users who publish the tweets for the key-words related to case-study.

Support Vector Machine (SVM), NaIve Bayes (NB) and Maximum Entropy (MaxEnt) classifiers are well discussed in many literatures such as Pang and Lee [4][11][12], where as Artificial Neural Networks (ANN) have been discussed very limited number of times. Rodrigo Moraes et al. [11] discussed the comparative features of ANN and SVM in detail for the document level sentiment classification.

Malhar Anjarie found that using SVM the efficiency of sentiment classifier have raised to 80% which was better than NaIve Bayes (NB) and Maximum Entropy (MaxEnt). Peng Zhang applied RLSC to various dataset and commented on its efficiency comparable with SVM[13]

3.2 Related Work for summarization.

The development of web 2.0 resulted in enhancement of social network, which provides facility of text message like tweets and blogs. Previous on twitter summarization [15] provided topics of discussion using list of significant terms.[16] and [17] provided one line summary for each identified topic.[20] provided nearly 250 words of summary for each topic on basis of selecting tweets with high ranking first.[18]and [21] was used for sports match to detect sub events using spike.[25] provided two summary of which second provided only new information. Several work on identifying events using twitter [14], flicker[22] has been done.[23] has proposed news processing system called, twitterstand. [24] Provided with facility to identify real world events and non event message.[19] has used transmission control protocol congestion detection algorithm to identify peaks(spikes).

4. CONCLUSION

In this paper, we presented a survey on analysis of social media data for sentiment analysis and summarization of tweets. We also gave an overview of different approaches of sentiment analysis and twitter summarization. It is a new area of research and have relatively good accuracy, it has created a new way collect information from crowd in an effective manner with low cost.

.

References

[1] Brendon O'Connor and Balasubramanyan et al,From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series , Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

[2] Malhar Anjaria, Ram Mahana Reddy Guddeti ," Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning" , IEEE 2014.

[3] Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In proceedings of ACL 2002.

[4] Bo Pang. Lilliam Lee, "Seeing Stars: Exploiting class relationships for sentiment categorization with respect to rating scales", 2002.

[5] Stefano Baccianella ,"Enhanced lexical Resource For Sentiment Analysis & Opinion Minning".

[6] Phillip j.Stone ," Sentiment lexicon –General Inquirer: A Competetive Approach to content Analysis",2007

[7] Pennebaker , "Linguistic Inquiry and Word Count",2007.

[8] Potts, Christopher ,"On the negativity of negation",2006.

[10] Johan Bollen, Alberto Pepe, and Huina Mao. 20 II. "Modeling public mood and emotion: Twitter sentiment and

socioeconomic phenomena" . In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain.
[11] Cozma, R., and Chen, K., "Congressional Candidates" Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?" 6lst Annual Conference of the International Communication Association, 2011
[12] Pew Research Center, "Parsing Election Day Media: How the Midterms Message Varied by Platform", Pew, 2010.
[13] Peng Zhang and Jing Peng , "SVM vs Regularized Least Squares Classification", Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
[14] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in Proc. WWW-10, 2010, pp. 851–860.
[15] B. O'connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory search and topic summarization for Twitter," in Proc. ICWSM-10., 2010.
[16] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Automatic summarization of Twitter topics," in Proc. National Workshop Design Anal. Algorithms, 2010.
[17] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in microblog summarization," in Proc. SOCIALCOM-10, 2010, pp. 49–56.
[18] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. AAAI-11, 2011.
[19] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: Aggregating and visualizing microblogs for event exploration," in Proc. CHI-11, 2011, pp. 227–236.
[20] S. Harabagiu and A. Hickl, "Relevance modeling for microblog summarization," in Proc. AAAI-11 Symp. Weblogs Social Media, 2011.
[21] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using Twitter," in Proc. IUI-12, 2012.
[22] L. Chen and Roy, "Event detection from Flickr data through waveletbased spatial analysis," in Proc. CIKM-09, 2009.
[23] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: News in tweets," in Proc. SIGSPATIAL-09,
2009, pp. 42–51.
[24] H. Becker,M.Naaman, and L. Gravano, "Selecting quality Twitter content for events," in Proc. AAAI-11 Symp. Weblogs and Social Media, 2011.
[25] L. Liviu, "Predicting Product Performance with Social Media," Informatics in education, vol. 15, no. 2, pp. 46-56, 2011.
[26] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang, and You Ouyang, "Sequential Summarization: A Full View of Twitter Trending Topics" IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, no. 2, february 2014
[27] D.M. Blei, N. Y. Andrew, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., pp. 993–1022, 2003.
[28] D.M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. ICML, 2006, pp. 113–120.
[29] M. Hu, B. Liu "Mining and summarizing customer review", In Proc. Of International Conference On Knowledge Discovery And Data Mining (KDD), 2004.