www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Rumour identification in social media using time varying network topology

A.Jose Prakash[1], C.Keerthiraja[2], P. Kirubhakaran[3], T. Nithyanandhan[4], M. Senthamarai[5]

[1]UG Students, Department of Computer Science and Engineering, Nandha Engineering College (Autonomous),Erode,Tamilnadu,India

[2]Assistant Professor, Department of Computer Science and Engineering, Nandha Engineering College (Autonomous),Erode,Tamilnadu,India

ABSTRACT

Twitter is an interesting platform for the dissemination of news. The real-time nature and brevity of the tweets are conducive to sharing of information related to important events as they unfold. But, one of the greatest challenges is to find the tweets that we can characterize as news in the ocean of tweets. In this paper, we propose a novel method for detecting and tracking breaking news from Twitter in real-time.We filter the stream of incoming tweets to remove junk tweets using a text classification algorithm. Finally, we rank the news using a dynamic scoring system which also allows us to track the news over a period of time.

Keywords: IBM,SVM,Tweets classification.

INTRODUCTION

The real-time nature and shortness of the tweets encourages user to communicate real-time events using least amount of text. Sakaki et al. used Twitter for early detection of earthquakes in the hope of sending word about them before they even hit. In fact, due to this real-time nature, Twitter can be used as a sensor to gather up-to-date information about the state of the world. The goal of this paper is to design a system to be used for detecting and tracking breaking news in real-time on Twitter.

The paper proposes an approach to detect and track breaking news in presence of noisy data stream without relying on traditional news publishers. We evaluate different algorithms which classify tweets as either news or junk. We also show how a traditional density based clustering algorithm can be used for detecting clusters in a stream of streaming data. We also propose a singular technique to parallelize classification of tweets using RabbitMQ. Finally, the paper also proposes a novel dynamic scoring system for ranking and tracking news

Classification of tweets

Millions of users share opinions on var0ious topics using micro-blogging every day. Twitter is a very popular micro blogging site where users are allowed a limit of 140 characters; this kind of restriction makes the users is concise as well as expressive at the same time. For that reason, it becomes a rich source for sentiment analysis and belief mining. The aim of this paper is to develop such a functional classifier which can correctly and automatically classify the sentiment of an unknown tweet. This project introduce two methods: one of the methods is known as sentiment classification algorithm (SCA) based on k-nearest neighbor (KNN) and the other one is based on support vector machine (SVM). This

Author for correspondence:
Department of Computer Science and Engineering, Nandha Engineering College (Autonomous) ,Erode, Tamilnadu, India

project also evaluate their performance based on real tweets. These days social networks, blogs, and other media produce a huge amount of data on the World Wide Web.

Both techniques work with same dataset and same features. For both SCA and SVM this project calculate weights based on different features. Then in SCA, this project build a pair of tweets by using different features. From that pair, this project measure the Euclidian distance for every tweet with its counterpart. From those distance this project only consider nearest eight tweets label to classify that tweet. On the other hand in SVM, build a matrix from the calculated weights based on different features and by applying PCA (principal component analysis), this project try to find k eigenvector with the largest Eigen values. From this transformed sample datasetthis project try to find the best c and best gamma by using grid search technique to use in SVM. Finally, this project apply SVM to assign the sentiment label of each tweet in the test dataset. In both algorithms, this project use confusion matrix to calculate the accuracy.Later, this project compare our two techniques in respect to an accuracy level of detecting the sentiment accurately. This project found that Sentiment Classifier Algorithm (SCA) performs better than SVM.

METHODOLOGY

Real-world event identificationon twitter

The work proposes User-contributed messages on social media sites such as Twitter have emerged as powerful, real-time means of information sharing on the Web. These short messages tend to reflect a variety of events in real time, making Twitter particularly well suited as a source of real-time event content. Our approach relies on a rich family of aggregate statistics of topically similar message clusters. Large-scale experiments over millions of Twitter messages show the effectiveness of our approach for surfacing real-world event content on TwitterSocial media sites (e.g., Twitter, Facebook, and YouTube) have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events..Twitter messages reflect useful event information for a variety of events of different types and scale. These event messages can provide a set of unique perspectives, regardless of the event, reflecting the points of view of users who are interested or participate in an event.

Even for planned events (e.g., the 2010 Apple Developers conference), Twitter users often post messages in anticipation of the event. Identifying events in real time on Twitter is a challenging problem, due to the heterogeneity and immense scale of the data. Twitter users post messages with a variety ofcontent types, including personal updates and various bits of information .While much of the content on Twitter is not related to any particular real-world event, informative event messages nevertheless abound. As an additional challenge, Twitter messages, by design, contain little textual information, and often exhibit low quality (e.g., with typos and ungrammatical sentences).

SVM

Text categorization with support vector machines

This explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyzes the particular properties of learning with text data and identifies why SVMs are appropriate for this task. Empirical results support the theoretical findings. SVMs achieve substantial improvements over the currently best performing methods and behave robustly over a variety of different learning tasks. Furthermore, they are fully automatic, eliminating the need for manual parameter tuning.With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the WWW, and to guide a user's search through hypertext.

Since building text classifiers by hand is difficult and time-consuming, it is advantageous to learn classifiers from examples. They are well-founded in terms of computational learning theory and very open to theoretical understanding and analysis. After

reviewing the standard feature vector representation of text, I will identify the particular properties of text in this representation. I will argue that SVMs are very well suited for learning in this setting. The empirical results in will support this claim. Compared to state-of-the-art methods, SVMs show substantial performance gains. Moreover, in contrast to conventional text classification methods SVMs will prove to be very robust, eliminating the need for expensive parameter tuning.

A comparison of event models for naive bayes textclassification

In this paper work [9]Andrew McCallum(2012), has proposed Recent approaches to text classification have used two different first-order probabilistic models for classification, both of which make the naive Bayes assumption. Some use a multi-variate Bernoulli model, that is, a Bayesian Network with no dependencies between words and binary word features (e.g. Larkey and Croft 1996; Koller and Sahami 1997). Others use a multinomial model, that is, a uni-gram language model with integer word counts (e.g. Lewis and Gale 1994; Mitchell 1997). This paper aims to clarify the confusion by describing the differences and details of these two models, and by empirically comparing their classification performance on five text corpora. This paper find that the multivariate Bernoulli performs well with small vocabulary sizes, but that the multinomial performs usually performs even better at larger vocabulary sizes—providing on average a 27% reduction in error over the multivariate Bernoulli model at any vocabulary size.

Simple Bayesian classifiers have been gaining popularity lately, and have been found to perform surprisingly well. These probabilistic approaches make strong assumptions about how the data is generated, and posit a probabilistic model that embodies these assumptions; then they use a collection of labelled training examples to estimate the parameters of the generative model..The naive Bayes classifier is the simplest of these models, in that it assumes that all attributes of the examples are independent of each other given the context of the class. This is the so-called ―naive Bayes assumption.‖ While this assumption is clearly false in most real-world tasks, naive Bayes often performs classification very well. Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes is large.

TOPICAL CLUSTERING OF TWEETS

In this paper the emerging field of micro-blogging and social communication services, users post millions of short messages every day. Keeping track of all the messages posted by your friends and the conversation as a whole can become tedious or even impossible. In this paper presented a study on automatically clustering and classifying Twitter messages, also known as ―tweets‖, into different categories, inspired by the approaches taken by news aggregating services like Google News. Our results suggest that the clusters produced by traditional unsupervised methods can often be incoherent from a topical perspective, but utilizing a supervised methodology that utilize the hash-tags as indicators of topics produce surprisingly good results. This paper also offer a discussion on temporal effects of our methodology and training set size considerations. Lastly, this paper describe a simple method of finding the most representative tweet in a cluster, and provide an analysis of the results.

Recent research efforts in social media analysis and natural language processing have focused on interesting uses of Twitter messages, or ―tweets‖ as they are more colloquially known, and other short socially communicated messages, such as SMS and micro-blogging messages or comments. One interesting problem in tweet analysis is the automatic detection of topics being discussed in tweets. This paper propose that the hash-tags that appear in tweets can be viewed as approximate indicators of a tweets topic. Thefirst discuss past work on tweet and microblogging message analysis. Nextthis paper formulate our approach to Twitter message topic detection, target topics and describe our data set. Then this paper describe a set of experiments and results. Finallythis paper offer a discussion of our results and suggest research future directions.

MODULE DESCRIPTION

Preprocessing

The classical division of sentiments into positive and negative is inappropriate, because diseases are generally classified as negative. Positive emotions could arise as a result of relief about an epidemic subsiding, but this project ignore this possibility. use ‗‗Negative‘‘ as the name of the first category and ‗‗Non-Negative‘‘ for the second one. The problem reduces to a two-class classification problem, and a Trendstweet can either be a Negative tweet or a Non-Negative tweet. Twitter messages were transformed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity. This project use ‗‗Negative‘‘ as the name of the first category and ‗‗Non-Negative‘‘ for the second one. Thus, the problem reduces to a two-class word alignment problem, and a Trendsreview can either be a Negative review or a Non-Negative review.

Clue-based review labeli1ng

The clue-based classifier parses each review into a set of tokens and matches them with a corpus of Trendsclues. There is no available corpus of clues for Trendsversus News classification. The MPQA corpus contains a total of 8221 words, including 3250 adjectives, 329 adverbs, 1146 any-position words, 2167 nouns, and 1322 verbs. As for the sentiment polarity, among all 8221 words, 4912 are negatives, 570 are neutrals, 2718 are positives, and 21 can be both negative and positive. In terms of strength of subjectivity, among all words, 5569 are strongly subjective words, and the other 2652 are weakly subjective words . Social media users tend to express their trendsopinions in a more casual way compared with other documents, such as News, online reviews, and article comments. It is expected that the existence of any profanity might lead to the conclusion that the review is a Trendsreview

Machine learning classifiers for trends reviewclassification

This combined the high precision of clue-based classification with Machine Learning-based classification in the Trends vs. News classification. The two classes of data T0p and T0n from the clue-based labellingare used as training datasets to train the Machine Learning models. Used three popular models: Tri Model, and polynomial-kernel Support Vector Machine. After the Trends vs. News classifier is trained, the classifier is used to make predictions which is the pre-processed tweets.

Dataset of low recall in the clue-based approach, this project combined the high precision of clue-based classification with Machine Learning-based classification in the Trends vs. News classification. After the Trends vs. News classifier is trained, the classifier is used to make predictions on each twitter in T0, which is the pre-processed reviews dataset. The goal of Trends vs News classification is obtain the Separate Labels.

Topic classifiaction and identity tweets

The topic classification the well-known Bag-of-Words approach for text classification andnetwork-based classification. In text-based classification method, this project construct word vectors with trending topic definition and tweets, and the commonly used TF-IDF weights are used to classify the topics using a Tri-Model Multinomial classifier. In network-based classification method, this project identify top 5 similar topics for a given topic based on the number of common influential users.

The categories of the similar topics and the number of common influential users between the given topic and its similar topics are used to classify the given topic using a C5.0 decision tree learner. Experiments on a database of randomly selected 768 trending topics (over 18 classes) show that classification accuracy of up to 65% and 70% can be achieved using text-based and network-based classification modeling respectively.

FIGURE

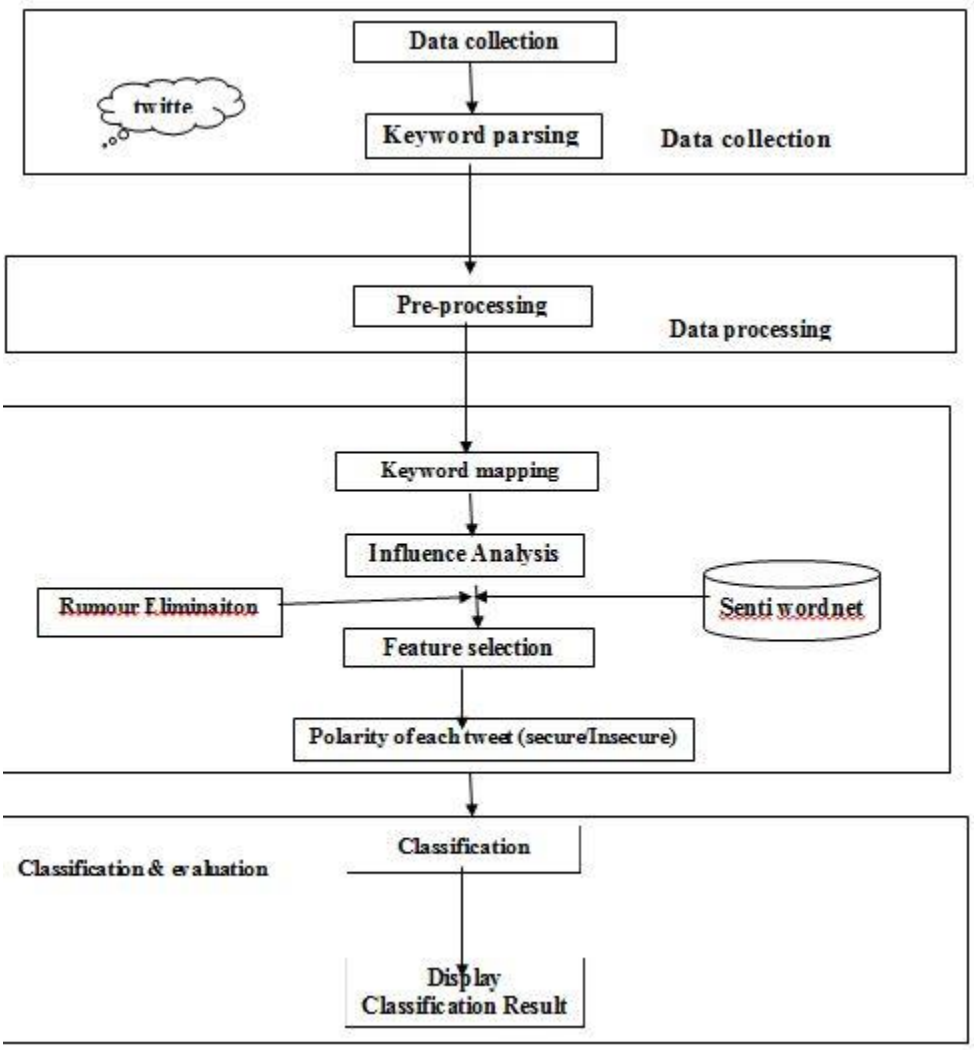

Fig 4.1: If necessary, the images can be extended both columns

Figure Explanation

In this figure the process that took in the module is shown clearly where a hastag datataset is collected in the work module which is preprocessed and all the data is checked with influence after polarity check the result is been displayed.

CONCLUSIONS

The proposed project is to monitor the public health concern from the reviews and them as positive and negative opinion. To find accuracy a two-step sentiment classification approach is implemented: In the first step, classify health reviews into Personal disease inference reviews versus News reviews. It uses a subjective clue-based lexicon and News stop words to automatically extract training datasets labelling Personal disease inference disease inference reviews and News reviews.These auto-generated training datasets are then used to train Machine Learning models to classify whether a review is Personal disease inference disease inference or News.

In the second step, utilized an emotion-oriented clue-based method to automatically extract training

datasets and generate another classifier to predict whether a Personal disease inference review is Negative or Non-Negative. In sentiment classification, by combining a clue-based method with a machine learning method, good accuracy can be achieved. This overcomes the drawbacks of the clue-based method and the Machine Learning methods when used separately

REFERENCES

[1] Becker.H, Naaman.M, and GravanoL. Beyond trending topics: Real-world event identification on twitter‖. ICWSM, 11: 2011, 438–441.

[2] McCallum,Nigam.K, et al(2012).‖ A comparison of event models for naive bayes text classification‖. In AAAI-98 workshop on learning for text categorization, 752, 41–48.

[3] Joachims.T. Text categorization with support vector machines: Learning with many relevant features‖. In European conference on machine learning, 2010, 137–142.

[4] Rosa.K.D,Shah.R,Lin.B,Gershman.A, andFrederking.R―Topical clustering of tweets‖. Proceedings of the ACM SIGIR: SWSM, 2011.