AN IMPROVED MACHINE LEARNING FOR TWITTER BREAKING NEWS EXTRACTION BASED ON TREND TOPICS

Dr.R. Nithya M.E.,Ph.D.,Assistant Professor, Department of Information Technology, K.S. Rangasamy College of Technology
Priyadharshini K, Department of Information Technology, K.S.Rangasamy College of Technology
Swathi Krishna S, Department of Information Technology, K.S.Rangasamy College of Technology
Swetha A V, Department of Information Technology, K.S.Rangasamy College of Technology

---

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 can characterize as news in the ocean of tweets. A novel method for detecting and tracking breaking news from Twitter in real-time.*

The filter the stream of incoming tweets to remove junk tweets using a text classification algorithm. The performance of different supervised text classification algorithms for this task are compared. Then cluster similar tweets, so that, tweets in the same cluster relate to the same real-life event and can be termed as a breaking news. Finally, rank the news using a dynamic scoring system which also allows us to track the news over a period of time.

Keywords - Brevity, Junk tweets, cluster, real-life event, dynamic scoring

1. 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. They evaluate different algorithms which classify tweets as either news or junk. A traditional density based clustering algorithm can be used for detecting clusters in a stream of streaming data are shown. They also propose a singular technique to parallelize classification of tweets using RabbitMQ. Finally, a novel dynamic scoring system for ranking and tracking news .

TWITTER

Twitter (/'twɪtər/) is an online news and social networking service where users post and interact with messages, known as "tweets." These messages were originally restricted to 140 characters, but on November 7, 2017, the limit was doubled to 280 characters for all languages except Japanese, Korean and Chinese. Registered users can post tweets, but those who are unregistered can only read them. Users access Twitter through its website interface, Short Message Service (SMS) or mobile device application software ("app").Twitter, Inc. is based in San Francisco, California, United States, and has more than 25 offices around the world.

Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and launched in July of that year. The service rapidly gained worldwide popularity. In 2012, more than 100 million users posted 340 million tweets a day, and the service handled an average of 1.6 billion search queries per day. In 2013, it was one of the ten most-visited websites and has been described as "the SMS of the Internet". As of 2016, Twitter had more than 319 million monthly active users. On the day of the 2016 U.S. presidential election, Twitter proved to be the largest source of breaking news, with 40 million election-related tweets sent by 10 p.m. (Eastern Time) that day.

NEWS DETECTION AND TRACKING

Twitter has been used as one of the communication channels for spreading breaking news. They propose a method to collect, group, rank and track breaking news in Twitter. Since short length messages make similarity comparison difficult, boost scores on proper nouns to improve the grouping results. Each group is ranked based on popularity and reliability factors. Current detection method is limited to facts part of messages.Developed an application called "Hotstream" based on the proposed method. Users can discover breaking news from the Twitter timeline. Each story is provided with the information of message originator, story development and activity chart. This provides a convenient way for people to follow breaking news and stay informed with real-time updates.

Twitter is a social networking service that allows users to share information, which is described by Twitter as "What's happening?" in a form of short texts (140 characters). Main characters of Twitter are: brevity—contents are in short length and simultaneousness—contents are updated frequently. Twitter has transformed the way people convey information especially in the areas of news.

In June 2009, Twitter has played an important role in delivering user-generated contents from the Iranian citizen in the Iran election. They see that people with technology played a role of journalists in the situation where news reporting in a conventional way has been made difficult. Anyone who is not associated to the media industry can also deliver news. Thus, Twitter presents a highly effective way to discover what is happening around the world.

Breaking news is defined by Wiktionary as "news that has either just happened or is currently happening. Breaking news may contain incomplete information, factual error or poor editing because of rush." With this definition Twitter can fit the needs of breaking news delivery.

However, news posted in Twitter requires an effort to discover it. Firstly, users often have problems of deciding which users to follow. That is, to find users with interesting tweets. Secondly, users need to read through status updates and follow links to obtain further information. To ease these problems and to deliver breaking news effectively,propose a method to collect, group, rank and track breaking news in Twitter. This work is a contribution to the area of Topic Detection and Tracking (TDT). The tasks focus are first story detection, cluster detection, and tracking.

ONLINE CLUSTERING

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this metodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, diferentiating it's properties and results. These models are distinguished by their organization and type of relantionship between them. The most important ones are:

- Centralized: each cluster is represented by a single vector mean, and a object value is compared to thus mean values
- Distributed: the cluster is build using statistical distribution.

- Connectivity – he connectivity on these models is based on a distance function between elements.
- Group: Algorithms have only group information.
- Graph: cluster organization and relationship between members is defined by a graph linked structure.
- Density – members of the cluster are grouped by regions where observations are dense and similar.

Clustering Algorithms in Data Mining

Based on the recently described cluster models, there are a lot of clustering that can be applied to a data set in order to partitionate the information. In this article will briefly describe the most important ones. It is important to mention that every method has it's advantages and cons. The choice of algorithm will always depende on the characteristics of the data set and what want to do with it.

Centroid-based

In this type os grouping method, every cluster is referenced by a vector of values. Each object is part of the cluster whose value difference is minimal, comparing to other clusters. The number of clusters should be pre-defined, and this is the biggest problem of this kind of algorithms. This methodology is the most close to the classification subject and are vastly used for optimization problems.

Distributed-based

Related to pre-defined statistical modelks, the distributed methodology combines objects whose values belogs to the same distribution. Because of its random nature of value generation, this process needs a well defined and complex model to interact in a btter way with real data. However these processes can achived a optimal solution and calculate correlations and dependencies.

Connectivity-based

On this type of algorithm, every object is related to its neighbours, depending the degree of that relationship on the distance between them. Based on this assumption, clusters are created with nearby objects, and can be described as a maximum distance limit. With this relationship between members, these clusters have hierarquical representations. The distance function varies on the focus of the analysis.

Density-based

These algorithms create clusters according to the high density of members of a data set, in a determined location. It aggregates some distance notion to a density standard level to group members in clusters. These kind of processes may have less performance on detecting the limit areas of the group.

Cluster Analysis main Applications

Since this is a very valuable data analysis technique, it has several different applications in the sciences world. Every large data set of information can be processed by this kind of analysis, producing great results with many distinct types of data.

One of the most importat application is related to image processing. detecting distinct kinds of pattern in image data. This can be very effective in biology research, distinguishing objects and identifying patterns. Another use is the classification of medical exams.

The personal data combined with shopping, location, interest, actions and a infinite number of indicators, can be analysed with this methodology, providing very important information and trends. Examples of this are the market research, marketing strategies, web analytics, and a lot of others.

Other types of applications based on clustering algorithms are climatology, robotics,

recommender systems, mathematical and statistical analysis, providing a broad spectrum of utilization.

CLASSIFICATION OF TWEETS

Millions of users share opinions on various topics using micro-blogging every day. Twitter is a very popular microblogging 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. In our work,they propose techniques to classify the sentiment label accurately. They 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).They 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. This huge amount of data contains crucial opinion related information that can be used to benefit for businesses and other aspects of commercial and scientific industries. Manual tracking and extracting this useful information from this massive amount of data is almost impossible. Sentiment analysis of user posts is required to help taking business decisions. It is a process which extracts sentiments or opinions from reviews which are given by users over a particular subject, area or product in online.

They can categorize the sentiment into two types: 1) positive or 2) negative that determine the general attitude of the people to a particular topic. Our principal goal is to correctly detect sentiment of tweets as more as possible. This paper has two main parts: the first one is to classify sentiment of tweets by using some feature and in the second one they use machine learning algorithm SVM. In both the cases, they use five-fold cross validation method to determine the accuracy. They propose two approaches for sentiment analysis. One of the technique facilitates KNN and the other uses SVM.

Both techniques work with same dataset and same features. For both SCA and SVM they calculate weights based on different features. Then in SCA, they build a pair of tweets by using different features. From that pair,measure the Euclidian distance for every tweet with its counterpart. From those distance they 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), they try to find k eigenvector with the largest Eigen values. From this transformed sample dataset they try to find the best c and best gamma by using grid search technique to use in SVM. Finally, they apply SVM to assign the sentiment label of each tweet in the test dataset. In both algorithms, they use confusion matrix to calculate the accuracy.

Later, they compare our two techniques in respect to an accuracy level of detecting the sentiment accurately. They found that Sentiment Classifier Algorithm (SCA) performs better than SVM.

2. METHODOLOGY

They regard extracting opinion targets/words as a co-ranking process. They assume that all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words, which are widely adopted by previous method.The given data is possibly of any modality such as texts or images, while it can be treated as a collection of documents.SUBJECT wise and TOPIC wise Opinion analysis is also possible.They formulate opinion relation identification as a word alignment process. They employ the word-based alignment model to perform monolingual word alignment, which has been widely used in many tasks such as collocation extraction and tag suggestion.

Tri-Model learning(Naïve Bayes,IBK SVM) an ensemble method that starts out with a base classifier that is prepared on the training data. A second classifier is then created behind it to focus on the instances in the training data that the first classifier got wrong. The process continues to add classifiers until a limit is reached in the number of models or accuracy.

ADVANTAGES:

- Tri-Model(Naive Bayes,IBK,SVM) Learning accommodates for both objective and subjective identification on any modalities (e.g., texts and images)
- High classification result.
- word alignment model for opinion relation identification high accuracy.
- The overall performance improved because of the use of partial supervision.

Less time to progress the report by using large dataset.

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 they 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**.** They 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.

Some features need to be removed or replaced. They first deleted the reviews starting with ‘‘RT’’, which indicates that they are re-reviews without comments to avoid duplications. For the remaining reviews, the special characters were removed. The URLs in Social media were replaced by the string ‘‘url’’. Social media’s special character ‘‘@’’ was replaced by ‘‘tag’’. For punctuations, ‘‘!’’ and ‘‘?’’ were substituted by ‘‘excl’’ and ‘‘ques’’, respectively, and any of ‘‘.,:;-|?=/’’ were replaced by ‘‘symb’’. Social media messages were transformed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity.

CLUE-BASED REVIEW LABLING

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 LERNING CLASSIFIER FOR TREND REVIEW CLASSIFICATION

They 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 labeling are 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 preprocessed tweets. Dataset of low recall in the clue-based approach, they 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 preprocessed reviews dataset. The goal of Trends vs News classification is obtain the Separate Labels.

TOPIC CLASSIFIACTION AND RESULT EXTRACTION:

The topic classification;

(i) the well-known Bag-of-Words approach for text classification and

(ii) network-based classification. In text-based classification method, they 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, they 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.

- Improved accuracy guaranteed.

- Better decision support system could increase business.

3. 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 labeling 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.

4. REFERENCES

[1] H. B¨acklund, A. Hedblom, and N. Neijman. A density-based spatial clustering of application with noise. Data Mining TNM033, pages 11–30, 2011.

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

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, pages 137–142. Springer, 1998.

[4] McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on

learning for text categorization, volume 752, pages 41–48. Citeseer, 1998.

[5] S. Petrovi´c, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 181–189. Association for Computational Linguistics, 2010.

[6] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM, 2011.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, pages 851–860. ACM, 2010.

[8] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pages 42–51. ACM, 2009.

[9] S. Van Canneyt, M. Feys, S. Schockaert, T. Demeester, C. Develder, and B. Dhoedt. Detecting newsworthy topics in twitter. In Data Challenge (SNOW 2014), pages 1–8, 2014.