www.ijiarec.com

ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Emerging Twitter Using Text Classification Based on Live Streaming

[1]K.Mohana Priya, [2]R.Bavithra, [3]M.Kanimozhi, [4]V.Meenatchi, [5]D.Janani

[1]Assistant Professor / CSE, [2-5] Final year CSE

Velalar College of Engineering and Technology, Thindal, Erode.

E-mail id: kmohanapriyacse@gmail.com[1],bavithra1411@gmail.com[2],kanimozhim14@gmail.com[3]

meenatchicse44@gmail.com[4],janani.dsj@gmail.com[5]

ABSTRACT

Twitter is one of the communication channels for spreading breaking news. It is an interesting platform for dissemination of news. The nature and brevity of the tweets are conducive way to share information related to important events. But one of the greatest challenges is to find the number of tweets that characterized as breaking news in the ocean of tweets. A novel method is used for detecting and tracking breaking news from Twitter in real-time. Filtering the stream of incoming tweets and it removes junk tweets by using greedy and text classification algorithm. It Compares the performance of different task and clusters the similar tweets. Finally, dynamic scoring system is used to track the news over a period of time. This method is used to collect, group, track and update the breaking news automatically. This provides a convenient way for people to follow the breaking news and stay connected with real-time updates. The domain-specific Naive Bayes model can capture the specific sentiment expressions in each domain. Two kinds of domain similarity measures are explored, one based on textual content and the other one based on sentiment expressions. In this method an efficient way to accurately categorize trending topics without need of external data, enabling news organizations to discover breaking news in real-time, or to quickly identify viral memes that might enrich marketing decisions, among others.

INTRODUCTION

WEB REVIEW ANALAYSIS AND OPINION MINING

The development of Web 2.0 websites, user generated content (UGC), such as product reviews, blogs, microblogs and so on, has been growing explosively. Mining the sentiment information in the massive user generated content can help sense the public's opinions towards various topics, such as products, brands, disasters, events, celebrities and so on, and is useful in many applications.

The major contributions of this paper are as follows:

- We propose a collaborative multi-domain sentiment classification approach (CMSC) based on multi-task learning to train sentiment classifiers for multiple tweets simultaneously. It can exploit the sentiment relatedness between different tweets and effectively alleviate the problem of scarce labeled data.
- We propose to extract domain-specific sentiment knowledge for each domain by propagating the sentiment scores inferred from limited labeled samples along contextual similarities mined from massive unlabeled samples.
- We propose to incorporate the similarities between tweets into the collaborative learning process. In addition, we propose a novel domain similarity measure based on the sentiment expression distributions.

- We introduce an accelerated algorithm based on FISTA to solve our model effectively, and propose a parallel algorithm based on ADMM to further improve its efficiency.
- We evaluate our approach by conducting extensive experiments on the benchmark Twitter product review datasets. The experimental results show our approach can improve the sentiment classification accuracy by 2.74 percent in average compared with the best baseline method.

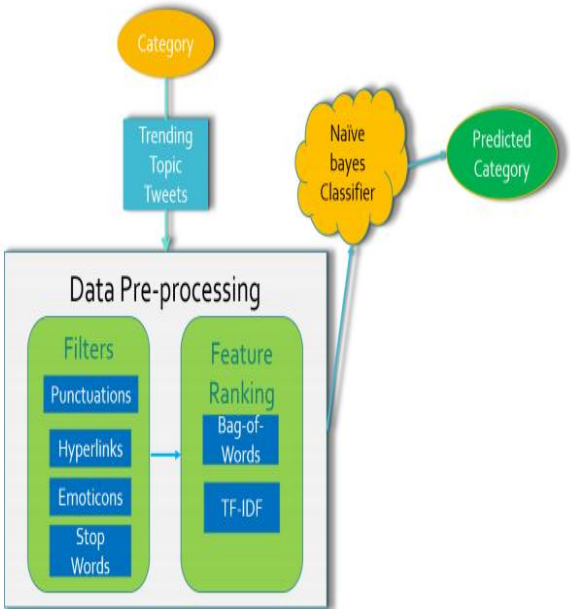
Data Mining is an investigative procedure considered to investigate data in discovery of reliable patterns and/or systematic associations among variables, and to verify the conclusion by applying the detected patterns to novel subsets of data. The ultimate objective of data mining is prediction and predictive data mining is the most numerous type of data mining and one that most direct commerce applications. The process of data mining contain of three stages: (1) The preliminary investigation (2) Model building or pattern identification with justification/confirmation (3) Deployment.

Stage 1: **Exploration:** In this step frequently starts with data training which may reside in cleanout data, data transformations, selecting subsets of statement and data sets with enormous information of variables the stage some beginning attribute selection operations to transmit the number of variables to convenient sequence. Then, depending on the situation of the analytic difficulty, this is the beginning stage of the process in data mining may connect someplace among a simple option of easy predictors for a regression model, to complicated investigative analyses using a wide selection of graphical and statistical methods (Exploratory Data Analysis (EDA)) in order to recognize the most related variables and decide the involvedness and the general nature of models can be taken into explanation in the next step.

Stage 2: **Model building and validation:** in this step involves allowing for a variety of models and choosing the best one based on their predictive presentation (i.e., explaining the unpredictability in question and producing stable consequences across samples). It may sound similar to a simple operation, except in details, it sometimes involves a very complicated procedure.

Stage 3: **Deployment**: The last step involves using the model selected as most excellent in the earlier stage and applying it to novel data in order to produce predictions or estimates of the probable outcome. The conception of Data Mining is becoming more and more popular as an industry information organization tool where it is probable to reveal knowledge structures that can lead decisions in situation of restricted certainty. In recent times, it has been enlarged interest in developing innovative analytic techniques particularly designed to address the issues applicable to business Data Mining (e.g., Classification Trees), excluding Data Mining is still based on the conceptual principles of statistics as well as the conventional Exploratory Data Analysis (EDA) and modeling and it shares with them both some mechanism of its common approaches and explicit techniques.

OVERALL ARCHITECTURE DIAGRAM



MODULE DESCRIPTION

PREPROCESSING:

Creation of database for recommender system, dataset of ratings i.e. actual ratings is used. Validity of results is based on the use of dataset, so creation of database is one important step. Some websites provides the available datasets which include users and tweets with significant rating history, which makes it possible to

have sufficient number of highly predicted tweets for recommendations to each user.

The data was gathered using twitter's publicly available API. Twitter momentarily updates its top ten trending topic list. There is no information as to how a topic gets chosen to appear in this list or how often this list gets updated. However, one can request up to 1500 tweets for a given trending topic.

It had two processes running to collect this data. One process requested a list of trending topics from twitter every 30 seconds and maintained a unique list. Whenever there was a new trending topic detected, the other process requested a list of related tweets from twitter using its search API. After the data was collected, the trending topics were manually annotated into the following three categories:

1. News

2. Meme

3. Ongoing Event

The three annotators were used to annotate the trending topics. Each one of them looked at the tweets related to the trending topics to assign a suitable category.

TWEETS RATING PREDICTION

In this module there are Naive Bayes recommender system techniques Proposed: content based, collaborative and hybrid approaches. Content based approach recommends tweets similar to the user preferred in the past. Dynamic Collaborative filtering approach suggests tweets that users with similar preferences have liked in the past. It can combine both content based and collaborative filtering approaches. The proposed system uses Naive Bayes approach. While giving suggestions to each user, recommender system performs the following two tasks.

First, based on the available information the ratings of unrated tweets are predicted using some recommendation algorithm. a new approach for classifying Twitter trends by adding a layer of feature selection and feature ranking. A variety of feature ranking algorithms, such as TF-IDF and bag-of-words, are used to facilitate the feature selection process. This helps in surfacing the important features, while reducing the feature space and making the classification process more efficient. Four Naïve Bayes text classifiers (one for each class), backed by these sophisticated feature ranking and feature selection techniques, are used to successfully categorize Twitter trends. Using the bag-of-words and TF-IDF rankings, our research provides an average class precision improvement, over the current methodologies, of 33.14% and 28.67% correspondingly

And second, based on the result of predicted ratings the system finds relevant tweets and recommends them to the user.

NAIVE BAYES TWEET BASED COLLABORATIVE FILTERING

In this module uses the set of tweets the active user has rated and calculates the similarity between these tweets and target tweets and then selects N most similar tweets. Tweets's corresponding similarities are also computed. Using the most similar tweets, the prediction is computed. The information filtering module is responsible for actual retrieval and selection of movies from the movie database. Based on the knowledge gathered from the learning module, information filtering process is done.

After passing out the test of user knowledge, the standardized ratings provided by the user are stored in the rating database. Based on the data in the rating database, a film is recommended to the user ui using the following steps Assume M = Total number of users N = Total numbers of films n = Total number of films not rated by user.

1) For each film F ȯ n not rated by user ui , find the correlation with each of the other (N-1) films.

2) Based on the correlation coefficient values select S films, which is mostly closely correlated with F. This will form a group of S similar films with F.

3) Find the correlation of all users with the current user ui based on the rating given by every user to those similar films. Based on the correlation coefficient values, select X users,

which are most closely correlated with user. Thus it will form a group of X users similar with user .

TWEET SIMILARITY COMPUTATION:

In this module the similarity computation between two tweets a (target tweets) and b is to first find the users who have rated both of these tweets. There are number of different ways to compute similarity. The proposed system uses adjusted cosine similarity method which is more beneficial due to the subtracting the corresponding user average from each co-rated pair. Similarity between tweets a and b is given.

PREDICTION COMPUTATION MODULE:

In this modules to obtain the predictions weighted sum approach is used. Weighted sum computes the prediction of target tweets for a user u by computing the sum of ratings given by the user on the tweets similar to target tweets. Prediction on an tweets a for user u is given Content based technique The utility for user u of tweets i is estimated based on the utilities assigned by user u to set of all tweets similar to tweets. Only the tweets with high degree of similarity to user's preferences are would get recommended.

TRENDING TWEETS RESULT ANALYSIS MODULE:

In movie database creation module, information related to user, movies and ratings has been stored in different tables. Thus system can retrieve the data properly from database and also get movie ratings explicitly from the users. In tweets based collaborative filtering technique, tweets similarity computation and prediction computation modules have been implemented. Recommended lists are generated on non purchased movies of login user. So we have computed system predicted ratings for all non purchased movies of login user. To calculate system predicted rating of target movie, first we have obtained 5 most similar tweets and then used weighted sum approach for rating prediction computation. As per the 5-star scale of rating, predicted value lies between 1 to 5. We have used Mean Absolute Error (MAE) accuracy metric to evaluate the accuracy of predicted ratings by this module shown in graph.

CONCLUSION

In the last few decades, recommender systems have been used, among the many available solutions, in order to mitigate information and cognitive overload problem by suggesting related and relevant tweets to the users. In this regards, numerous advances have been made to get a high-quality and fine-tuned recommender system. Nevertheless, designers face several prominent issues and challenges.

In this work, we have touched variety of topics like natural Language Processing, Text Classification, Feature selection, Feature ranking, etc. Each one of these topics was used to leverage the massive information flowing through twitter. Understanding twitter was as important as knowing the topics in question**.** The results of the previous experiments, led us to the conclusion that feature selection is an absolutely necessity in a text classification system. This was proved when we compared our results with a system that uses the exact same dataset without feature selection. We were able to achieve 33.14% and 28.67% improvement with bag-of-words and TF-IDF scoring techniques correspondingly**.**

We also mentioned recognition and some opportunities that our work provides in the fields of news media, marketing and businesses in general. We hope that our work can provide a good foundation to the future of text classification in social media and to the opportunities that comes with it.

REFERENCES

[1] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc. Int. AAAI Conf. Weblogs Social Media, 2011, pp. 17–21.

[2] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proc. Int. AAAI Conf. Weblogs Social Media, 2010, pp. 122–129.

[3] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, "Learning user and product distributed representations

using a sequence model for sentiment analysis," IEEE Comput. Intell. Mag., vol. 11, no. 3, pp. 34–44, Aug. 2016.

[4] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "OpinionFlow: Visual analysis of opinion diffusion on social media," IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1763–1772, Dec. 2014.

[5] A.Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford Univ., Stanford, CA, USA, Project Rep. CS224N, pp. 1–12, 2009.

[6] F. Wu, Y. Song, and Y. Huang, "Microblog sentiment classification with contextual knowledge regularization," in Proc. 29th AAAI Conf. Artif. Intell., 2015, pp. 2332–2338.

[7] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 513–520.

[8] S.-S. Li, C.-R. Huang, and C.-Q. Zong, "Multi-domain sentiment classification with classifier combination," J. Comput. Sci. Technol., vol. 26, no. 1, pp. 25–33, 2011.

[9] L. Li, X. Jin, S. J. Pan, and J.-T. Sun, "Multi-domain active learning for text classification," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1086–1094.

[10] A.Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imaging Sci., vol. 2, no. 1, pp. 183–202, 2009.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1–122, 2011.

[12] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 751–760.

[13] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, "Automatic construction of a context-aware sentiment lexicon: An optimization approach," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 347356.

[14] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[15] Y. He, C. Lin, and H. Alani, "Automatically extracting polaritybearing topics for cross-domain sentiment classification," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, 2011, pp. 123–131.