IJPREMS

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 359-364

e-ISSN : 2583-1062

Impact Factor : 5.725

A PRECISE SPOTTING OF COUNTERFEIT NEWS USING SENTIMENT CLASSIFICATION

Rajavenkatesswaran K C[1], Nareshkumar V[2], Nithesh Kumar M[3], Praveen Kumar[4], Vairaprakash M[5]

[1]Assistant Professor, Department of Information Technology, Nandha College of Technology, Perundurai 638 052, Tamil Nadu, India.

[2,3,4,5]UG Students - Final Year, Department of Information Technology, Nandha College of Technology, Perundurai 638 052, Tamil Nadu, India.

ABSTRACT

We propose a collaborative multi-Trends sentiment classification approach to train sentiment tweets simultaneously. Specifically, we decompose the sentiment classifier of each trend into two components, a global one and a Trends-specific one. Automatically identifies the important aspects of topics from online consumer reviews. We analyse and experiment with a set of straight forward language-independent features based on the social spread of trends to categorize them into the introduced typology as news, ongoing events, memes and commemoratives. The global model can capture the general sentiment knowledge and is shared by various tweets. The Trends-specific Greedy & Dynamic Blocking Algorithms like model can capture the specific sentiment expressions in each Trend. In addition, we extract Trends-specific sentiment knowledge from both labeled and unlabeled samples in each Trend and use it to enhance the learning of Trends-specific sentiment classifiers. Two kinds of Trends similarity measures are explored, one based on textual content and the other one based on sentiment expressions. Moreover, we introduce two efficient algorithms to solve the model of our approach. Experimental results on benchmark datasets show that our approach can effectively improve the performance of multi-Trends sentiment classification and significantly outperform baseline methods.

Keywords: Tweets, multi-Trends, Sentiment classification, Trends-specific, Topics, Reviews, Greedy & Dynamic Blocking Algorithm.

1. INTRODUCTION

1.1 Web Opinion Data Mining Concept

User-generated content (UGC), such as product reviews, blogs, and microblogs, has seen explosive growth in Web 2.0 websites. Mining sentiment data from a lot of user-generated content can help understand how people feel about a variety of topics, such as brands, disasters, celebrities, and so on, and is useful in a lot of applications. Product review sentiment analysis can assist businesses in improving their products and services and assisting customers in making better-informed choices. User interest mining, personalized recommendation, social advertising, customer relationship management, and crisis management all benefit from sentiment analysis of user-generated content. As a result, sentiment classification is a hot topic for academic and industrial research.

The following are the major contributions of this paper:

- To train sentiment classifiers for multiple tweets simultaneously, we propose a collaborative multi-Trends sentiment classification method (CMSC) based on multi-task learning. It can effectively alleviate the issue of limited labeled data by taking advantage of the sentimental connection between various tweets.

Data mining is a method for finding reliable patterns and/or systematic associations between variables in data. These patterns can then be applied to new subsets of data to verify the conclusion. Predictive data mining is the most common type of data mining and the one used in the majority of direct commerce applications. Prediction is the ultimate goal of data mining. The three stages of data mining are as follows: 1) The preliminary investigation; 2) The construction of models or the identification of patterns with justification or confirmation; and 3) The deployment.

Stage 1: Exploration: This step typically begins with data training, which may include cleanout data, data transformations, selecting subsets of statements, and data sets with a large number of variables. Some beginning attribute selection operations are also performed to transfer the number of variables into a convenient sequence. The initial stage of the data mining process may range from straightforward exploratory analyses employing a wide range of graphical and statistical techniques (Exploratory Data Analysis, or EDA) to more complex options of straightforward predictors for a regression model.

Stage 2: Constructing and validating a model: Allowing for a variety of models and selecting the best one based on their predictive presentation (i.e., explaining the unpredictability at hand and producing stable outcomes across samples) is the goal of this step. It may appear to be a straightforward operation, but in reality, it sometimes involves a very complex procedure.

Stage 3: Deployment: Using the model that was chosen as the best in the first step and putting it to use with new data to make predictions or estimates of the likely outcome is the final step. Excluding Data Mining, which is still based on the conceptual principles of statistics and the conventional Exploratory Data Analysis (EDA) and modeling, which shares some mechanism of its common approaches and explicit techniques, there has been an increased interest in developing innovative analytical techniques specifically designed to address the issues applicable to business Data Mining (e.g., Classification Trees).

1.2 Overview of Data Mining

The prediction tool known as data mining is used by large organizations to focus on the most crucial data in their data warehouses. It is a tool for predicting upcoming trends that enables organizations and businesses to make decisions based on knowledge directly. Data mining's computerized prospective analyses outperform those provided by conventional decision support system tools for past measures. That conventional procedure required too much time to answer the business questions. Experts in information discovery may overlook the hidden patterns in the source database because they are outside of their expectations. Data mining tools can look at huge databases to find answers to questions like, "How many clients and which clients the majority to take action my next promotional mailing and why?" and can be used on higher-end client/server or parallel processing computers. The fundamental data mining technologies are introduced in this paper. The current business environment is a starting point for the development of data warehouse architectures to distribute the rate of data mining to end users and has profitable applications.

2. LITERATURE REVIEW

2.1 Opinion Mining and Sentiment Analysis

According to Bo Pang, one essential component of our information-gathering behavior has always been to learn what others think. People can and do now actively use information technologies to seek out and comprehend the opinions of others, which presents new opportunities as well as challenges with the growing availability and popularity of opinion-rich resources like personal blogs and online review sites. Before people knew about the World Wide Web, many of us asked our friends to recommend a car mechanic or tell us who they planned to vote for in local elections. We also asked our coworkers for letters of recommendation about job applicants and looked at Consumer Reports to choose a dishwasher. The Internet and the World Wide Web have now, among other things, made it possible to learn about the experiences and opinions of the vast pool of individuals who are neither our personal acquaintances nor well-known professional critics**. [1]**

2.2 Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena

Johan Bollen has suggested that all tweets published in the second half of 2008 on the microblogging platform Twitter undergo a sentiment analysis. We compute a six-dimensional mood vector for each day in the timeline by utilizing a psychometric instrument to extract six mood states—tension, depression, anger, vigor, fatigue, and confusion—from the compiled Twitter content. Twitter's launch is to blame for the rise of this straightforward but widely used method of online communication. Microblogging is used by members of these online communities to share a variety of information. A recent examination of the Twitter network revealed a diverse set of uses (Java), including a) daily chatter, such as posting one's current activities, b) conversations, such as sending tweets to specific members of one's community of followers, c) information sharing, such as linking to web pages, and d) news reporting, such as providing commentary on news and current affairs. Tweets can convey information about the author's state of mind in either scenario. An explicit "sharing of subjectivity" (Crawford) such as "I am feeling sad" is indicative of mood expressions in the first instance. In other instances, a user's mood can be reflected in their message, even if they aren't specifically microblogging about their personal emotional state. **[2]**

2.3 From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Brendan O'Connor has proposed connecting poll-based public opinion and text-based sentiment measurements. We find a correlation between sentiment word frequencies in contemporaneous Twitter messages and several surveys on consumer confidence and political opinion from 2008 to 2009. Although our outcomes differ across datasets, significant large-scale trends are captured by correlations as high as 80% in several instances. The findings emphasize the potential of text streams to complement and replace conventional polling. If we want to know, for instance, how much people in the United States like or dislike Barack Obama, the obvious thing to do is conduct a poll with a random sample of people. Millions of people have shared their thoughts and opinions on a wide range of topics as a result of the meteoric rise of text-based social media platforms. It is demonstrated in this preliminary work that summary statistics derived from extremely straightforward text analysis techniques can not only predict future polling movements but also correlate with polling data on consumer confidence and political opinion. We discover that temporal smoothing is a crucial component of a successful model **[3].**

2.4 Mining and Summarizing Customer Reviews

Minqing Hu has proposed that sellers of topics on the Internet frequently request customer reviews of the topics and services they have purchased. As online business is turning out to be increasingly famous, the quantity of client surveys that an item gets develops quickly. There may be hundreds or even thousands of reviews for a well-known product. Because of this, it is difficult for potential customers to read them and decide whether or not to buy the product. As opposed to the traditional text summarization, we do not select a subset of the reviews and then rewrite some of the original sentences to convey the main points. There are three stages to our task: 1) analyzing customer feedback about a product's features; 2) identifying each review's opinion sentences and determining whether or not each opinion sentence is favorable or negative; 3) condensing the findings. To complete these tasks, a number of novel methods are proposed in this paper. Our trial results utilizing surveys of various subjects sold online exhibit the adequacy of the procedures. At some large merchant websites, popular topics can receive hundreds of reviews. In addition, many reviews are lengthy and contain only a few sentences of opinions regarding the product. A potential customer won't be able to read them well enough to decide whether or not to buy the product. He or she may have a skewed perspective if they only read a few reviews. **[4]**

2.5 Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis

The distribution of polarity ratings across reviews written by various users or evaluated based on various topics is frequently skewed in the real world, as proposed by Tao Chen and Ruifeng Xu. As a result, the task of sentiment classification of reviews would benefit from incorporating information about users and products. Researchers focused on determining the text's polarity using language clues taken from reviews' text. Beyond ratings, many recommendation and review websites provide a wealth of information, such as opinions expressed by users (hereinafter referred to as "users") and target entities (hereinafter referred to as "topics") that received reviews **[5].**

3. EXISTING APPROACH

In the performance of the current system for a brand-new category of data analysis software known as "recommender systems." The issue of providing personalized product recommendations during a live customer interaction is addressed by recommender systems, which make use of knowledge discovery methods. Recommender systems face significant obstacles as a result of the rapid expansion of topics and customers over the past few years. They are: making numerous high-quality recommendations per second for millions of customers and topics. Although matrix factorization is a very expensive process, Singular Value Decomposition (SVD)-based recommendation algorithms can quickly produce high-quality recommendations. A method that has the potential to incrementally build SVD-based models and promises to make the recommender systems highly scalable is proposed and experimentally validated in this work.

4. PROPOSED APPROACH

Greedy & Dynamic Blocking Algorithms recommends tweets by matching users with others who share similar interests, as proposed in our work. It gathers user feedback in the form of ratings for particular tweets and looks for patterns in rating practices among users to identify a group of users who share similar preferences. A list of the most popular terms, also known as "trending topics," is one of the main features on Twitter's homepage at all times. These terms represent the topics that are currently receiving the most attention in the site's frantic stream of tweets. Twitter focuses on topics that are being discussed much more than usual, i.e., topics that have recently experienced an increase in use, so that it became a trend for some reason, in order to avoid topics that are popular frequently (for example, good morning or good night at certain times of the day).Here, a user's preferences, which the user has either explicitly or implicitly provided, are represented by a user profile. Twitter's approach, for instance, uses user ratings and purchase patterns of its users to suggest tweets. Each user has a list of tweets that have been given explicit or implicit ratings. A user-tweets rating matrix with the symbol "R" is created in this manner, displaying the preferences of users regarding tweets. Different methods are used to find missing ratings, such as recommending tweets to new users based on the ratings provided by their closest neighbors and locating the "nearest neighbor."

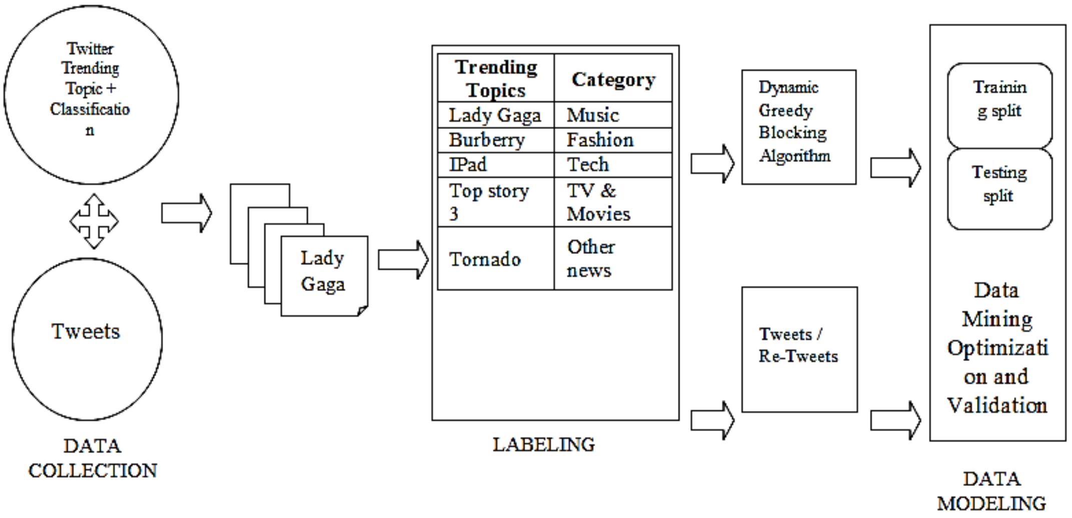

Figure 1. System Flow Diagram

4.1 Preprocessing

The actual ratings dataset is used in the creation of the database for the Twitter asynchronous system. The creation of a database is an essential step because the use rating histories is made available by some websites. This makes it possible to have a sufficient number of highly predicted tweets for each user's recommendations. Using Twitter's API, which is accessible to the public, the data were gathered. Twitter temporarily updates its list of the top ten most popular topics. There is no indication of how a topic is selected to be included on this list or how frequently it is updated. However, for any given trending topic, one can request up to 1500 tweets. In order to collect this data, it had two processes running. One process kept a unique list of Twitter's most popular topics and requested one every 30 seconds. The other process used Twitter's search API to request a list of related tweets whenever it detected a new trending topic. The trending topics were manually annotated into the following four categories following the collection of the data:

- News
- Meme
- Current Event

The three annotators were utilized for the annotation of the trending subjects. To assign a suitable category, they all looked at tweets related to the most popular topics.

4.2 Tweets Rating Prediction

Techniques for twitter asynchronous systems, such as greedy and dynamic blocking algorithms, are proposed in this module: The greedy algorithm, which is based on live content, suggests tweets that are similar to those that users have previously favored. The dynamic greedy strategy recommends tweets that users who share similar preferences have liked in the past. It can combine collaborative filtering with content-based filtering. The greedy and dynamic blocking algorithms approach is utilized in the proposed system. The Twitter asynchronous system performs the two tasks listed below while providing suggestions to each user. First, a recommendation algorithm is used to predict the ratings of unrated tweets based on the information that is available. To make the process of selecting features easier, a variety of feature ranking algorithms, including TF-IDF and bag-of-words, are utilized. This helps bring the most important features to light, reduces feature space, and speeds up the classification process. Four Eager and Dynamic Impeding text classifiers (one for each class), upheld by these modern component positioning and element determination procedures, are utilized to effectively order Twitter patterns. Our research shows that the bag-of-words and TF-IDF rankings provide a class precision improvement of 33.14% and 28.67%, respectively, over the current methods. Second, the system finds relevant tweets and recommends them to the user based on predicted ratings.

4.3 Greedy & Dynamic Blocking Algorithms Tweet Based Collaborative Filtering

This module takes the set of tweets that the active user has rated, calculates how similar they are to the target tweets, and then chooses N tweets that are the most similar. The corresponding similarities between tweets are also calculated. The prediction is calculated using the tweets with the greatest similarity. Movies from the movie database are actually retrieved and selected by the information filtering module. The process of filtering information is carried out using the information gleaned from the learning module. The standardized ratings that the user provides are saved in the rating database following the completion of the user knowledge test. The following steps are used to recommend a movie to the user ui based on the information in the rating database: M is the total number of users, N is the total number of movies, and n is the total number of movies that have not been rated by the user.

- Determine the correlation between each of the other (N-1) films for each film F n that has not been rated by user ui.
- Using the values of the correlation coefficients, select S films that are most closely associated with F to form a group of S films that are similar to F.
- Using the ratings that each user gave to those similar films, determine the correlation of all users with the current user interface. Select X users who are most correlated with user based on the correlation coefficient values. As a result, a group of X users similar to user will emerge.

4.4 Tweet Similarity Computation

Before calculating the similarity between two tweets a (target tweets) and b in this module, the users who have rated both tweets must first be identified. The calculation of similarity can be done in a variety of ways. The adjusted cosine similarity method, which subtracts the corresponding user average from each co-rated pair, is more advantageous in the proposed system. It is stated that tweets a and b share similarities.

4.5 Prediction Computation Module

The weighted sum method is used to get predictions in this module. By summarizing the user's ratings for tweets that are similar to the target tweets, weighted sum calculates the prediction of target tweets for user u. Expectation on an tweets a for client u is given Substance based method The utility for client u of tweets I is assessed in view of the utilities relegated by client u to set of all tweets like tweets. Tweets that are highly similar to the preferences of users will only be recommended.

4.6 Trending Tweets Result Analysis Module

Information about users, movies, and ratings has been stored in various tables in the movie database creation module. As a result, the system can accurately retrieve data from the database and obtain explicit user ratings for movies. The tweets similarity computation and prediction computation modules have been incorporated into the tweets based collaborative filtering technique. On movies that the login user has not purchased, recommended lists are generated. Therefore, we have calculated system predicted ratings for all login user-owned movies. We used a weighted sum approach to compute the rating prediction for the target movie before obtaining the five tweets that were the most similar to each other to determine the system predicted rating. According to the 5-star size of rating, anticipated esteem lies between 1 to 5. The accuracy of this module's predicted ratings, as depicted in the graph, was evaluated with the help of the accuracy metric known as Mean Absolute Error (MAE).

5. RESULTS

We utilized well-known instruments like WEKA and SPSS modeler for our experiments. WEKA is a popular machine learning tool that supports a variety of modeling algorithms for feature selection, clustering, regression, data preprocessing, and classification. The popular data mining software SPSS modeler features a distinctive graphical user interface and high prediction accuracy. It is utilized extensively in national security, resource planning, medical research, law enforcement, and business marketing. The accuracy of the classification was evaluated through 10-fold cross-validation in each and every experiment. After testing our model with a variety of K values, we discovered the K value at which the system provides the highest accuracy. The Zero-R classifier was used to obtain the baseline accuracy, which simply predicts the majority of classes. This model's classification accuracy on the test set was approximately 79 percent.

6. CONCLUSION

Twitter asynchronous systems are one of many solutions that have been used over the past few decades to alleviate the problem of information and cognitive overload by recommending related tweets to users. In this regard, numerous advancements have been made toward developing a Twitter asynchronous system of high quality and refinement. However, designers face a number of significant issues and difficulties. Natural Language Processing, Text Classification, Feature Selection, Feature Ranking, and other related topics have all been covered in this work. Each of these subjects was used to make use of the vast amount of information being shared on Twitter. Knowing the subjects at hand was just as important as comprehending Twitter. The consequences of the past trials, drove us to the end that highlight choice is a totally need in a text grouping framework. This was demonstrated when we contrasted our outcomes and a framework that utilizes precisely the same dataset without highlight determination. With the bag-of-words and TF-IDF scoring methods, we were able to achieve improvements of 33.14% and 28.67%, respectively. We also talked about some of the opportunities and recognition that our work gives businesses, marketing, and news media in general. We hope that our work can serve as a solid foundation for the future of social media text classification and the associated opportunities.

7. REFERENCES

[1] "Opinion mining and sentiment analysis," Found, by B. Pang and L. Lee. Inf. Trends Vol. of retrieval 2, no. 1/2, pp. 1–135, 2020.

[2] "Modeling public mood and emotion:" by J. Bollen, H. Mao, and A. Pepe Proc., "Twitter sentiment and socioeconomic phenomena" Int. Conf. of AAAI Social Media Blogs, 2020, pages 17–21.

[3] "From tweets to polls:" by B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. Linking public opinion time series to text sentiment," in Proc. Int. Conf. of AAAI Social Media Blogs, 2019, pages 122–129.

[4] "Mining and summarizing customer reviews," by M. Hu and B. Liu in Proc. 10th ACM SIGKDD International Conf. Knowl. Pages from Discovery Data Mining, 2019, 168–177.

[5] IEEE Computer, "Learning user and product distributed representations using a sequence model for sentiment analysis," by T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang. Intell. Mag., vol. 11, no. 3, pp. 34–44, Aug. 2019.

[6] "OpinionFlow: Visual analysis of opinion diffusion on social media," by Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, IEEE Trans. Vis. Comput. Graph, vol. 20, no. 12, pp. 1763–1772, Dec. 2017.

[7] "Thumbs up?: Sentiment classification using machine learning techniques," by B. Pang, L. Lee, and S. Vaithyanathan, in Proc. ACL Conf. Empirical Methods Natural Language Process., 2018, pp. 79–86.

[8] "Twitter sentiment classification using distant supervision," by A.Go, R. Bhayani, and L. Huang, Stanford Univ., Stanford, CA, USA, Project Rep. CS224N, pp. 1–12, 2019.

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

[10] "Biographies, bollywood, boom-boxes and blenders: Trends adaptation for sentiment classification," by J. Blitzer, M. Dredze, and F. Pereira, in Proc. 45th Annu. Meeting Assoc. Comput. Linguistics, 2017, vol. 7, pp. 440–447.

[11] "Trends adaptation for large-scale sentiment classification: A deep learning approach," by X. Glorot, A. Bordes, and Y. Bengio, in Proc. 28th Int. Conf. Mach. Learn., 2015, pp. 513–520.

[12] "Multi-Trends sentiment classification with classifier combination," by S.-S. Li, C.-R. Huang, and C.-Q. Zong, J. Comput. Sci. Technol., vol. 26, no. 1, pp. 25–33, 2016.

[13] "Multi-Trends active learning for text classification," by L. Li, X. Jin, S. J. Pan, and J.-T. Sun, in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 1086–1094.

[14] "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," by A.Beck and M. Teboulle, SIAM J. Imaging Sci., vol. 2, no. 1, pp. 183–202, 2015.

[15] "Distributed optimization and statistical learning via the alternating direction method of multipliers," by S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1–122, 2016.