INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue5)
Available online at www.ijariit.com

Literature Review on Sentiment Analysis

Venkata Satya Sai Abhishikth Tholana
Department of Computer Science
GITAM University
Visakhapatnam, India
abhishikth.96@gmail.com

Abstract ***Sentiment analysis or Opinion mining is one of the fastest growing fields with its demand and potential benefits increasing every day. With the onset of the internet and modern technology, there has been a vigorous growth in the amount of data. Each individual is able to express his/her own ideas freely on social media. All of this data can be analyzed and used in order to draw benefits and quality information. One such idea is sentiment analysis, here, the sentiment of the subject is considered and necessary information is drawn out whether it be a product review or his/her opinion on anything materialistic. A few of such applications of sentiment analysis and the method in which they are implemented are explained. Furthermore, the possibility of each of these works to effect any future work is considered and explained along with the analysis as to how the previous problems in the same field have been overcome.***

Keywords: Sentiment Analysis, Opinion Mining, Product Review, Drug Related Classification, Forecasting Price.

INTRODUCTION

In this paper, a detailed analysis of various works related to sentiment analysis is taken. The works in various fields such as housing, stock exchange, product review, drug response analysis and others are considered and explained. Sentiment analysis has its applications in numerous fields. Wherever there is a data attached to a subject there is always a sentiment. Sentiment analysis is usually considered to have either a positive, neutral or a negative result. The emotional analysis is sometimes merged with sentiment analysis in which case multiple tuples come into use.

Sentiment analysis basically involves classification, but the content that actually contains the sentiment or opinion must be first identified. The most fundamental problem in sentiment analysis is sentiment polarity categorization. This is tackled in each of the aforementioned works in a unique way so as to improve the efficiency of analysis in each of the cases.

First, an efficient drug event extraction is enabled and its possibility by using sentiment analysis on data sets from Twitter is analysed. Second, the effect of the bias in investors' sentiment on the volatility of stocks is explained. Also, the way in which an individual's opinion on social media affects the price of a stock and the methods to forecast such shocks are analysed. Furthermore, based on the public sentiments, potential livable places are analysed and hence calibrated. Of the works discussed, the analysis of the sentiment of a customer on a product through reviews has been the most approached area in sentiment analysis due to its potential in both increasing accuracy and business benefits.

SENTIMENT ANALYSIS USING PRODUCT REVIEW DATA

Xing Fang and Justin Zhan

In this paper, Xing Fang and Justin Zhan try to discuss the most fundamental problem in sentiment analysis, the sentiment polarity categorization, by considering a dataset containing over 5.1 million product reviews from Amazon.com with the products belonging to four categories: beauty, books, electronics and home.

The previous papers in this area suggested to remove all the objective content in order to conduct sentiment analysis but here, the subjective content is instead extracted for future analysis. The inputs taken are reviews which contain the customer details, review, helpfulness and rating. The rating is considered as the ground truth for more accurate analysis of the review's sentiment.

A max-entropy POS tagger is used in order to classify the words of the sentence into 46 tags. An additional python program is particularly used to speed up this process. As a result, a total of 25 million adjectives, 22 million adverbs and 56 million verbs are

identified, which usually tend to determine the sentiment. The negation words like no, not, and more are included in the adverbs whereas Negation of Adjective and Negation of Verb are specially used to identify the phrases. 21,586 phrases are identified with a total of 0.68 million. The algorithm also makes a list of phrases based on occurrence. The following are the various classification models which are selected for categorization: Naïve Bayesian, Random Forest, and Support Vector Machine.

Although this paper tackles the problem of sentiment polarity categorization it still faces multiple challenges and has its limitations. One such being the curse of dimensionality in feature vector formation which limits the number of dimensions and also forces to have the same number of dimensions.

The performance of this approach is estimated by considering the average F1 score. Therefore future work would be benefited if these limitations considered and thereby the accuracy and performance can be improved.

EXPLORING PUBLIC PLACES FOR LIVEABLE PLACES BASED ON A CROWD-CALIBRATED SENTIMENT ANALYSIS MECHANISM

Linlin You and Bige Tuncer

To fill the vacancy, a sentiment analysis service, called geo-sentiment analysis service is required. Thus, this paper firstly proposes CGSA: a Crowd-calibrated Geo-Sentiment Analysis mechanism, which can 1) start the sentiment analysis process based on the design of CTS (Compound Training Samples), and SSF (Social Sentiment Features), 2) perform three analyses, namely sentiment, clustering and time series analysis on geo-tagged social network messages, and 3) collect crowd-labelled data based on a crowd-sourced calibration service to gradually improve the classification accuracy. SSF has the best accuracy in training sentiment classifiers, and the performance of the calibrated classifier increases gradually and significantly from 74.71% to 80.05% in three calibration cycles. Moreover, as a part of a big project "Liveable Places", "Sentiment in places" service with two visualization modes, namely 2D sentiment dashboard and 3D sentiment map, is implemented to support local authorities, urban designers and city planners better understand the effects of public sentiments regarding place (re)design in the test-bed area: Jurong East, Singapore. In general, three issues are solved, choosing sentiment features, maintaining up-to-date and localized lexicons. The application of social sentiment analysis brings many benefits in various domains, even using existing sentiment analysis tools, e.g. SentiStrength, or very simple analysis methods, which can only provide a very basic analysis without specific optimizations towards an application domain. Through this service, the service users, namely local authorities, urban designers and city planners, can better measure the satisfaction of people, and evaluate the fulfillment of predefined functionalities of facilities in the test-bed. Also, through these analyses, general sentiment patterns can be created, which can be used as baselines to evaluate the influence of changes or events, e.g., the reconstruction work, or the temporarily out of service of MRT (Mass Rapid Transit). Therefore, service users can better understand the reactions of people, and make better decisions.

EFFICIENT ADVERSE DRUG EVENT EXTRACTION USING TWITTER SENTIMENT ANALYSIS

Yang Peng, Melody Moh, and Teng-Sheng Moh

In this paper, Yang Peng, Melody Moh, and Teng-Sheng Moh discuss how the advancements of social media are being helpful to extract large datasets by using a drug-related classification and sentiment analysis to extract ADEs on Twitter.

ADEs are adverse drug events. Even though the pharmaceutical companies perform many drug-related tests beforehand, when a drug is released into the market some ADEs will be unidentified. Through the above-mentioned method, a data of four months on Twitter is collected so, as to capture the maximum number of ADEs.

A simple and efficient pipeline is proposed to retrieve data from Twitter. The process of the pipeline is, the tweets from twitter are captured firstly and then the data is pre-processed (cleaned data is the output of data pre-processing). The drug classification is done for the cleaned data and the user opinion data is collected from which the ADEs are extracted. The captured tweets are stored in HIVE. Tweets are in JSON file and can, therefore, be stored in HIVE directly. They used python NLP tool for capturing tweets and Data pre-processing. For storing datasets of drug-related classification and tweets of sentiment analysis WEKA is used. Thus after thorough research on different tweets pipelines are built and they are compared to newly designed ones to extract numerous ADEs. As, a result an average of 5 times of total number of ADEs, among them 20% are new ADEs.

The proposed system may further include streaming more tweets from Twitter by using Topsy API, using more drugs as keywords in the experiments, applying Apache Spark for processing a lot of tweets. The proposed method may be applied to other areas such as daily consumer products for recognizing side effects and user opinions on them.

INVESTOR CLASSIFICATION AND SENTIMENT ANALYSIS

Arjit Chatterjee and Dr. William Perrizo

In this paper, Arijit Chatterjee and Dr. William Perrizo discuss the effect investors' bias has on the volatility of stocks in the market, sentiment analysis was done on tweets of the potential investors and also why they used Microsoft Azure over other sentiment analyser tools.

Twitter is one of the largest social media platforms with over 280 million active users with almost 500 million tweets created every day. Some investors use Twitter to share their opinion on some ticker symbols every day, this paper discusses how these opinions of the investors affect the stock market.

Investors are assumed to be sentiment driven. A top-down approach is used to make sure a stock is not overrated or underrated by the investors. The approach is based on two broad behavioral finance assumptions - sentiment and limits to arbitrage.

Sentiment analysis is done on the tweets pulled from some selected investors' twitter feeds. They assign positive, negative and neutral sentiment scores to the ticker symbols from the pulled tweets by identifying "bad", "not good", "great" words in the tweets.

For processing, the unstructured text in the tweets "Microsoft Azure" analyser tool is used. Because Microsoft Azure gave better results when compared to other analysis tools, this is shown in the four graphs in the paper where Microsoft Azure is compared with Stanford NLP Sentiment Analysis engine and another popular commercial tool.

Through sentiment analysis, a particular user can understand the social sentiment score of a ticker symbol based on the discussions of key investors and make an informed decision about which stock to invest.

FORECASTING PRICE SHOCKS WITH SOCIAL ATTENTION AND SENTIMENT ANALYSIS

Li Zhang, Liang Zhang, Keli Xhao and Qi Liu

In this paper, the data from the Chinese Stock Market – SZSE and SSE are considered along with the social media activities in Weibo.com in order to extend recent studies on financial activities in social media and their impact on the stock market.
What makes this work stand out is the way in which the previous limitations, such as, inability to tackle practical problems in finance, lack of proper knowledge regarding the direction of the price shock and, were overcome by using this implementation.

This work makes use of DSA – Degree of Social Attention, which has been introduced by the previous works to capture stock price shocks. The method involves identifying the price shock as negative, near-zero or positive. These price shocks are essentially the difference between the expected value and actual value. Prior to estimating these price shocks, the social media activities are analysed and the features such as account information, future tracking, and response such as, like, repost, comment are considered. Furthermore, these details are cross-referenced to the account holder's actual activity and its effect on trading.

The efficiency of this work is proved by evaluating the F-measure without DSA and with DSA to show the improvement. The following are the classifiers used: Naïve Bayes, Decision Tree, Radom Forest, Logistic and LibSVM. Out of these classifiers, the Radom Forest has the best performance and SVM the worst. Also, once the price shocks are estimated it is identified that the negative shocks lead to better accuracy than the positive shocks.

CONCLUSION

In this work, the latest developments in sentiment analysis are reviewed and the future possibilities for each of these developments are presented. Sentiment analysis is a field which is catching up in the recent years and its applications are subject to increase to a broader range in near future. This work is an attempt to create a basis with the help of which future works can be improved and also take a note of the challenges this field offers. The effectiveness of various approaches has been evaluated and shown.

REFERENCES

1. Xing Fang and Justin Zhan: Sentiment analysis using product review data
2. Linlin You and Bige Tuncer : Exploring public places for livable places based on a crowd-calibrated sentiment analysis mechanism
3. Yang Peng, Melody Moh, and Teng-Sheng Moh : Efficient Adverse Drug Event extraction using Twitter sentiment analysis
4. Arjit Chatterjee and Dr. William Perrizo : Investor classification and sentiment analysis
5. Li Zhang, Liang Zhang, Keli Xhao and Qi Liu : Forecasting price shocks with social attention and sentiment analysis

