Sentiment Analysis to Analyze Tron ($TRX)

Authors: Jon Li, Annan Miao
Instructor: Christopher Dancy
Course: CSCI 379-01 FA18
Date: 10/31/2018

Thesis

This report discusses the problem of sentiment analysis and the approach to implementing it using statistic method of data training and modeling. It also describes the implementation of sentiment analysis on twitter data for TRON ($TRX), a highly publicized and high market cap cryptocurrency. The investigate the effects of public sentiment on the price of the coin and attempt technical and comparative statistical analysis.

Introduction

Twitter is a vital platforms for people to publicly express their opinions and feelings about different topics. This paper will introduce the opinion mining and sentiment analysis on Twitter data, discuss possible methods and theories, and describe how to build a sentiment analysis classifier from scratch.

This report is structured as follows - the Related Work presents the related work of methods for sentiment analysis. The implementation describes the goal, the overall approach and detailed implementation taken in this research, marked with a chronological steps. The implementation presents the results as a product of consolidating the different work. Finally, the paper will conclude with the ethical implications.

Related Work

Sentiment analysis is the Natural Language Processing work, which involves opinion detection and classification of attitudes in texts. In order to perform different sentiment classification tasks, various sentiment algorithms were developed. Medhat et al. (2014) grouped the SA into two categories: machine learning and lexicon-based approaches.

Machine-learning methods were used to automatically discover sentiment polarity pattern rules in large data in order to learn opinions or emotions of given texts or features. For example, (Rushdi Saleh et al.) applied Support Vector Machines (SVM) to detect whether the opinion expressed is positive or negative about a given topic using several weighting schemes. Balahur (2013) developed an unsupervised method especially for a Twitter data sentiment analysis using the SVM, which normalised Tweet language, including higher order n-grams to spot modifications in sentiment polarity articulated and selecting features using simple heuristics.

Lexicon-based approaches focus on measuring subjectivity and opinions in texts using Semantic orientation (SO), which capture orientations of opinions (positive or negative) and strengths or degrees of orientation. Sentiment lexicons are the key for this type of methods. For example, Paltoglou and Thelwall (2012) proposed a lexicon-based approach to identify whether a text conveys negative or positive attitudes and to estimate the level of emotional intensity of a text in

social media and microblogging environments. They added extensive linguistically (negation/capitalization detection, intensifier/diminisher detection and emoticon/exclamation detection) functionalities to the traditional classifiers such as Support Vector Machines (SVM) and Maximum Entropy classifier.

Implementation

The goal of the research is to build a classifier of sentiment analysis with statistical (machine learning) method, and use the classifier to analyze tweets related to the topic of cryptocurrency Tron. The implementation needs to data collecting from Twitter, and data classification. Data-scraping consisted of two parts: the scraping of the market movement data, historical data, and Twitter data. The market movement data was simply a product of scraping coinmarketcap.com using the HTML xpath so that was very simplistic. Next, the historical data was needed. Although there was some historical CSV files, there simply wasn't any day-to-day transactions that existed at anything else than daily intervals, so this portion was not as efficient or effective as the team would've liked. The Twitter data was scraped, albeit only from the last two weeks, with a Python API called tweepy and with the keywords, "$TRX" and "#TRX" , the hashtags which tweets were filed under. The API was effective but failed to have versatility in data collection by time interval. This affected testing because of the lack of diverse data sets but the advantageous component is the live streaming.

The dataset needs to be transformed into a format suitable for modeling in machine learning algorithm. In this case, the method of vectorization, also called one hot encoding, is feasible to transform each tweet into numeric representation. Basically, it will create one very large matrix with one column for every unique word in the database. Then it will transform each tweet into one row containing 0s and 1s, where 1 means that the word corresponding to that column appears in that tweet. The target of the training data, which is a list of 0s and 1s representing whether each tweet in training data is sentimentally positive or negative, is determined by the build in python library NLTK (Natural Language Toolkit), a leading platform of python to work with human language data.

After the dataset transforms to readable representation, the program is able to build a classifier for sentiment analysis, and then train and model the classifier with the training data. (In this case, the training database would be the tweets of related topic posted within most recent 3 hours.) The method of Logistic Regression is suitable model for this case because each row of the matrix will be mostly zeros, and thus linear models tend to perform well on sparse datasets like this one. After training the classifier with training data and target, the program can test the classifier with the generated test data, which is the tweets of related topic posted within most recent 1 hour. After operating sentiment analysis on the test data, the program will return a sentiment variable

represented by number of positive tweets divided by total number of tweets, and then post the number to a twitter bot (@TeamBot123) along with the current market direction of Tron (TRX).

Ethics

The ethical framework of the issue revolves around two of the larger issues at hand, intentional and unintentional violations of ethics and the law. There's a matter of utilitarian ethics and Kantian ethics, rooted in financial gain and loss and the effects of trading bots. Then, there's the legality, which can be argued as an extensive of practical ethics. In essence, what can be constituted as market manipulation and what isn't? Utilitarian ethics claims that if the overall profit is positive, then the effect if positive. Then, the bot would violate Kantian ethics, because someone would be negatively affected.

In terms of legality, the usage of Twitter as a market manipulated tool has already been legally challenged by the SEC with Tesla's chairman, Elon Musk, which resulted in him settling for a $50 million dollar sum. Twitter has thus been proven as a platform that counts as public, financial data. So, what happens when the bot develops the end goal of being to make money. If that is the purpose of the bot when programmed by the creator, then the bot theoretically could make some really poor decisions ethically. Given the time and resources, the team could have developed a following and a reputation of the market. With the prevalence and pervasiveness of the bot, what would happen is an effect which is like Elon Musk. The bot could thus continue and extend periods of bear and bull markets, effectively having some sort of control of the market given the amount of high-frequency trading platforms that exist with TRON and other cryptocurrencies. Say the bot decides to tweet false information to spoof the market and then shorts the position, gaining profit. Without implementation of an ethical barrier, the bot would be committing insider trading and spoofing, or tricking, the market, thus violating the 1934 Exchange Act and the 2010 Dodd Frank Act (Wellman).

Conclusion

In implementation of the bot, data limitations were unable to lead the team to the conclusions that were originally planned, but nonetheless, the framework for a sentiment analysis bot exists, which can tweet and iterates through datasets. Ethical issues in finance were investigated but with extreme scenarios that a thorough program can eliminate. Some of the limitations of the twitter API such as the timing constraints and the program timed out after 100 calls. Using the statistical method to build a classifier, and train it with datasets which includes texts with predetermined sentiment tendency. With this approach to implementing it using statistic method of data training and modeling, the team has implemented the first foundation of a sentiment analyzing bot.

References

Aaron Kub. (2018. Jul 31). Sentiment Analysis with Python Retrieved from https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184

Blockchain Engineer. Crypto Trading Bot — Sentiment Analysis Bot with TextBlob and Python. 2018. Medium.com

Song, Z., & Xia, J. (2016). Spatial and Temporal Sentiment Analysis of Twitter data. In Capineri C., Haklay M., Huang H., Antoniou V., Kettunen J., Ostermann F., et al. (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 205-222). London: Ubiquity Press. Retrieved from http://www.jstor.org/stable/j.ctv3t5r09.20

Stojanovski, D., Chorbev, I., Dimitrovski, I., & Madjarov, G. (2016). Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots. In Capineri C., Haklay M., Huang H., Antoniou V., Kettunen J., Ostermann F., et al. (Eds.), *European Handbook of Crowdsourced Geographic Information* (pp. 223-236). London: Ubiquity Press. Retrieved from http://www.jstor.org/stable/j.ctv3t5r09.21

Wellman, Michael P. and Rajan, Uday. *Ethical Issues for Autonomous Trading Agents*. Minds and Machines. 2017. Volume 27. Number 4.