Bitcoin Sentiment Analysis from Twitter Data

Osama Zahid

CS182021

Department of Computer Science DHA SUFFA University Karachi,Pakistan

Nihal Sarwar

CS182026

Department of Computer Science DHA SUFFA University Karachi,Pakistan

Furqan Ahmed

CS182002

Department of Computer Science DHA SUFFA University Karachi,Pakistan

**Abstract**

Cryptocurrencies such as Bitcoin (BTC) attracted a lot of attention in recent months due to their unprecedented price fluctuations. This paper aims to propose a new method for predicting the direction of BTC price with sentiment analysis algorithm.

# INTRODUCTION

# 

# In any financial market sentiment gets its own place and is used as a strong indicator to predict whether the market will outperform or not. As crypto-currencies are emerging passively in the financial markets, our goal is to determine what sentiment individuals and organizations have. As bitcoin leads the market so our research will be based on it to predict the sentimental analysis for the overall market. It has three major footsteps to follow. Firstly, to accumulate data from twitter by using twitter developer Api , secondly we have to clean the data first in order to achieve the relevant data. Lastly and the most significant thing of the model is applying the sentimental algorithm

# which will compare the threshold of sentiments among our data.

# In this study, we examine Twitter, a popular microblog, and develop models for categorising "tweets" into positive, negative, and neutral emotion. We create models for categorising sentiment into positive and negative categories, as well as a three-way task that categorises sentiment into positive, negative, and neutral categories. On the test data, including news sentiment resulted in the maximum forecast accuracy of 0.585, which is better than a random guess. Within their respective model classifiers, the model with asset specific (news sentiment and asset specific) input features scored best, implying that both BTC news sentiment and asset specific are important elements in predicting tomorrow's price direction.

# Our feature-based model with only 100 features achieves the same level of accuracy as the model with over 10,000 features. We give a detailed feature analysis of the 100 features we suggest in this study. Our tests demonstrate that attributes related to Twitter-specific features (emoticons, hashtags, etc.) add little to the classifier's usefulness. For both classification tasks, features that integrate prior polarity of words with their parts-of-speech tags are particularly relevant. As a result, we can observe that typical natural language processing methods can be beneficial in genres other than the one in which they were taught (newswire).

# LITERATURE SURVEY

Sentiment analysis has been tackled as a Natural Language Processing problem at several levels of granularity. More recent and distinct issues are posed by microblog information such as Twitter, where users post real-time comments to and critiques of everything. We collect tweets that end with emoticons in the same way for subjective data. They progressively move twitter money owing of famous publications like "New York Times," "Washington Posts," and others for goal information. They re-port the POS, and bigrams assist each other. Both approaches, however, are entirely reliant on anagram models. Furthermore, the information they use for instruction and experimentation is gathered through seek questions. We, on the other hand, give talents that provide a significant improvement over a unigram baseline. In addition, we uncover a novel method of information encoding and demonstrate significant progress over unigram models. In contrast to information gathered through the use of unique queries, our data is a random pattern of streaming tweets. Because of the amount of our hand-classified data, we can do cross validation tests to see if there is any variation in the classifier's overall performance over folds.

Another excellent attempt at sentiment classification on Twitter data is to utilise polarity predictions from three websites as noisy labels to teach a version and to use one thousand manually classified tweets for tuning and another thousand manually classified tweets for testing. Use of twitter syntax features such as retweet, hashtags, links, punctuation, and exclamation marks, as well as features such as previous polarity of phrases and POS of phrases. We improve on their method by integrating past polarity with POS and using actual valued previous polarity. Our findings show that the capabilities that improve our classifiers' overall performance the most are those that incorporate previous polarity of sentences with their speech components.

# DATA DESCRIPTION

Twitter is a social networking and microblogging website that allows users to send and receive real-time messages known as tweets. The length of a tweet is more likely to be limited to 140 characters. One of the distinguishing features of this blogging service is that humans utilise acronyms, emoticons, and other characters to express themselves in a unique way. Emoticons are a representation of a facial expression, such as a grin, or a combination of keyboard symbols that are intended to avoid misunderstandings. Hashtags are commonly used by users to mark topics. This is in general carried out to boom the visibility in their tweets. We gather the information to our came in the real-time stream no language facility another sort of restrict turned into made for the duration of the streaming process in fact their series include towards in overseas language they use Google Translate to transform 8 into English earlier than the annotation process East it is categorized through a human annotator as positive impartial or junk label method that it can no longer be understood through a human annotator usually junk advised that a lot of those tweets had been people who had been now no longer translated properly the use of Google Translate.

# PRIOR POLARITY SCORING

# Some of our highlights depend on the earlier polarity of words. For acquiring the earlier extremity of words, we take inspiration from work by Agarwal et al. (2009). We use the Dictionary of effect in Language (DAL) (Whissel, 1989) and expand it utilizing WordNet. This word reference of around 8000 English language words appoints each word a charming score ( R) between 1 (Negative) - 3 (Positive). We initially standardize the scores by jumping each score my the scale (which is equivalent to 3). We consider words with extremity under 0.5 as negative, higher than 0.8 as certain and the rest as nonpartisan. On the off chance that a word isn't straightforwardly found in the word reference, we recover all synonyms from Wordnet. We then, at that point, search for every one of the equivalents in DAL. In the event that any equivalent is found in DAL, we allot the first word a similar enjoyableness score as its equivalent word. Assuming none of the equivalents is available in DAL, the word isn't related to any earlier extremity. For the given information we straightforwardly found an earlier extremity of 81.1% of the words. We find the polarity of the other 7.8% of the words by utilizing WordNet. So we find an earlier extremity of around 88.9% of English language words.

# PROPOSED SYSTEM

# 

# This diagram shows the complete process of our model in which we collect data from twitter by using its developer API. Following the data collection in the data model, we must clean the data using pre-processing to obtain categorized tweets..

# The first subjectivity function will be applied to the categorized tweets to break the review sentence obtained from the tweets after which polarity will be applied to determine the negativity, positivity and neutral bias of the statement.

# Sentimental analysis algorithm is the main back-bone of this project for which we have used built-in algorithm by python IDE to obtain our desired result which is to check the bias of sentiments on the twitter platform related to bitcoin.

# OUR FEATURES

# We have endorsed a set of features listed in table 4 for our experiments. The data has been collected from the specific term search. So, we have calculated the features for the entire tweet and for the last one-third. Overall, we got 3 features. Throughout the research paper, we call these features as senti-features.

# All the features that we have listed are divided into three different classifications. The first category includes the type in which the feature primarily belongs to a natural number ∈ N. The second type comprises features that belong to Real numbers ∈ R which have been retrieved from DAL. The third category contains features that belong to Boolean ∈ B which is mainly a group of capitalized letters, exclamation marks, etc. These features are further classified substantially into type types. Polar features and Non-Polar features. Polar features can be defined as those features in which we can calculate the polarity directly by looking up at its DAL (extended through WordNet) or by finding it in the emoticon dictionary. While the non-polar polar features are not associated with prior polarity. These polar and non-polar features are further apportioned into two categories; POS and other. POS deals with features in which the statistics require the use of parts of speech and the other type includes all the multifarious types of features. With reference to table 4, row F1 belongs to the Polar category and hints at negative and positive parts of speech (POS). f2, f3, f4 rows belong to other category that alludes a number of refutation words, a bag of words that comprises positive and negative polarity, count of emoticons, hashtags, capitalized words, and words with exclamation marks that are completely linked with prior polarity. The row F5 belongs to NON- polar POS which signifies to the presence of multiple types of parts of speech tags.

# The row f6, f7 belongs to the NON-polar category which essentially includes latin alphabets, slang words and some words without polarity. Moreover, it also includes a count of hashtags, URLs, targets and newlines. The other row F8 is included in the Polar POS category with polarity scores of words such as JJ, RB, VB and NN. The row f8 is classified in the Polar other category which assists in the calculation of prior polarity scores of all words. The row f10 is added in the NON-Polar other category which mainly calculates the percentage of capitalized tweets.

# The last row f11 is added in the polar other category which comprises features that are marked by the presence of exclamation and capitalized words.

# EVALUATION OF SENTIMENT CLASSIFICATION

The performance of sentiment classification will be evaluated by using four indexes calculated because the following equations:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

In which TP, FN, FP and TN refer respectively to the amount of true positive instances, the quantity of false negative example, the quantity of false positive instances and therefore the number of true negative example

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive | Predicted Negative |
| Actual Positive | TP | FN |
| Actual Negative | FP | TN |

# RESULTS AND DISCUSSION

Analyses was done on this labeled datasets using various feature extraction technique. We used the framework where the pre-processor is applied to the raw sentences which make it more appropriate to grasp. Further, the various machine learning techniques trains the dataset with feature vectors so the semantic analysis offers an outsized set of synonyms and similarity which provides the polarity of the content.

**9. Baseline Algorithm**

The baseline algorithm used is Naïve Bayes without pre-processed data and unigram model. Following table shows the accuracy obtained at different sizes for the baseline algorithm.

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| 10 | 0.46475 |
| 50 | 0.53332 |
| 100 | 0.54744 |
| 500 | 0.61237 |
| 1000 | 0.65230 |
| 5000 | 0.69740 |
| 10000 | 0.71292 |

**9.1. Naive Bayes Algorithm:**

When Naïve Bayes(Baseline) was run, it gave an accuracy of 73.65 percent, which is taken into account because the baseline result. the subsequent thing used as removal of stop-word. When stop-words were removed and Naïve Bayes was run, it gave an accuracy of 74.56 percent. It shows the accuracy obtained at different sizes for the Naïve Bayes with stop-words removed and using pre-processed data and supported unigram model.

# 10. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

10.1. IDENTIFYING SUBJECTIVE PARTS OF TEXT

Subjective parts represent sentiment-bearing content. the identical word will be treated as subjective in one case, or an objective in another. This makes it difficult to spot the subjective portions of text.

10.2. DOMAIN DEPENDENCE

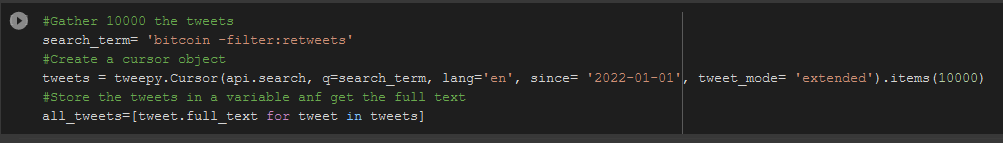
The identical sentence or phrase can have different meanings in several domains. for instance, the word “unpredictable‟ is positive within the domain of BTC next move, but if the identical word is employed within the context of a “unpredictable sell offs” then it's a negative opinion.

10.3. SARCASM DETECTION:

Sarcastic sentences express negative opinion a couple of target using positive words in unique way for example: People can purchase more Bitcoin to form themselves poor.

**11. Project Outcome**

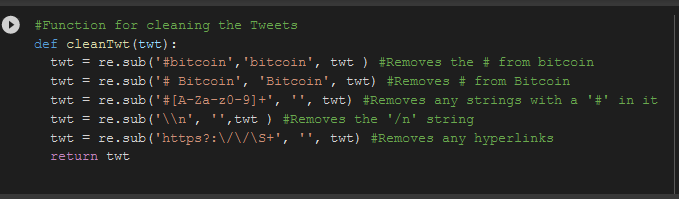
1. Extracting 10,000 tweets using Twitter API

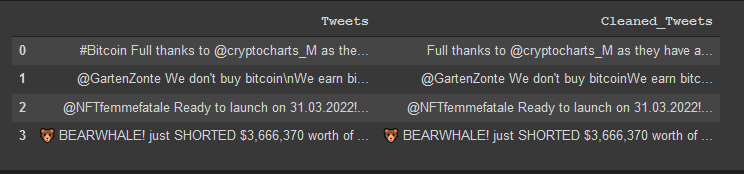


1. Fetched 10,000 tweets

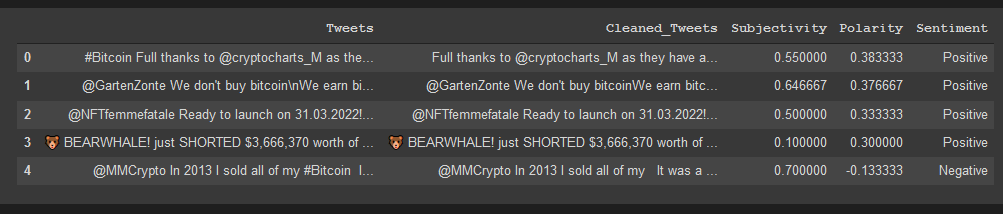


1. Cleaning the extra symbols and hashtags

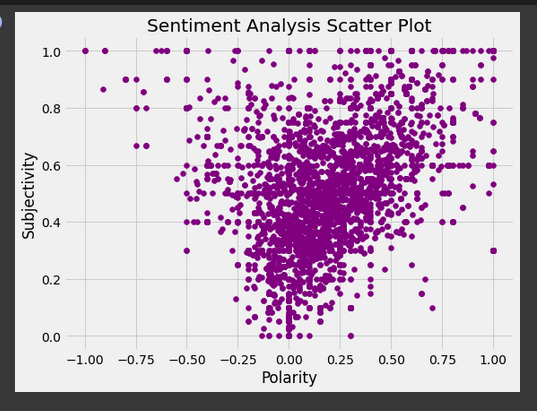




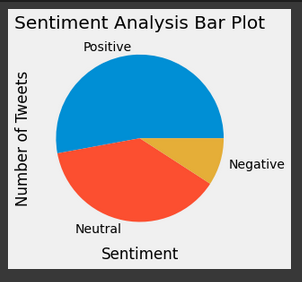
1. Measuring subjectivity, polarity and sentiment of tweets



1. Scatter plot of sentiments of the 10,000 tweets



1. Pie chart to present the biased on BTC



# 12. TWITTER SENTIMENT ANALYSIS

Today Internet contains a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus there's a requirement for a sentiment analysis system which will extract sentiments about market movement of the Bitcoin. it'll help retailors and investors to form clear decision on their investments.

Since social networks, especially Twitter, contains small texts and other people may use different words and abbreviations which are difficult to extract their sentiment by current linguistic communication processing system easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text.

A sentiment predictor system may be helpful in recommender systems still. In online communication, we stumble upon abusive language and other negative elements. These will be detected just by identifying a highly negative sentiment.

12.1 Designing

This technical document details the implementation of Twitter sentiment analysis using Twitter's APIs. There are great works and tools that specialize in text mining on social networks. The following is the method for extracting sentiment from tweets:

1. Begin by downloading and saving the sentiment dictionary to your computer.

2. Get the Twitter testing data sets and enter them into the software.

3. Remove the stop words from the tweets.

4. Tokenize each word within the dataset and enclose to the program.

5. for every word, compare it with positive sentiments and negative sentiments word. Then increment positive count or negative count.

6. Finally, supported the positive count and negative count, we are able to get result percentage about sentiment to make your mind up the polarity.

12.2 Implementation

# We utilised Python to implement emotional analysis in this paper. Tweepy and textblob are two examples of packages that have been used. The following commands will be used to install the desired libraries:

# • pip install tweepy

# • pip install textblob

# 13. CONCLUSION

# In this report, we discussed the importance of social network analysis and its applications on Bitcoin market. We concentrated on Twitter and utilised a Python tool to perform sentimental analysis. We showed the results on different daily sentiments over Bitcoin price. We realized that the neutral sentiments are significantly high which shows there's a necessity to boost Twitter sentiment analysis. we will conclude that more the cleaner data, more accurate results is obtained.

# 14. References

# [1] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter.

# [2] Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44. Adam Bermingham and Alan Smeaton.

# [3] ACM, pages 1833–1836.Fellbaum Wordnet, an electronic lexical database. MIT Press. Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis.

# [4] Proceedings of the 20th international conference on Computational Linguistics. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford. David Haussler. 1999.

# [5] Evolution kernels on discrete Technical report, University of California at Santa Cruz. M Hu and B Liu. 2004. Mining and summarizing customer reviews.

# [6] KDD. S M Kim and E Hovy. 2004. Determining the sentiment of opinions. Coling. Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pages 423–430.

# [7] Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In Proceedings of the 17th European Conference on Machine Learning. Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

# [8] B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

# [9] P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews.

# [10] ACL. C M Whissel. 1989. The dictionary of Affect in Language. Emotion: theory research and experience, Acad press London. T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. ACL