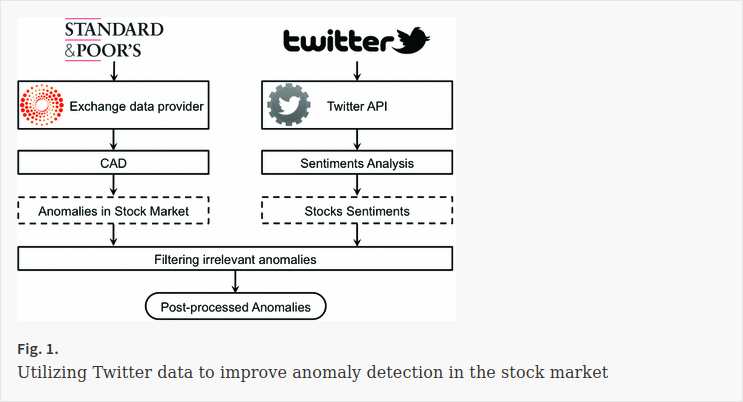
# Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation → Review

**Introduction**

The paper looks to propose a method to improve Contextual Anomaly Detection (CAD) to be used in detecting stock market manipulation. The method looks to capture the expected behavior of stocks through sentiment analysis of tweets. From there the insights from all of the tweets are then aggregated and transformed in to a time series which is used to eliminate false positives from detected anomalies. The goal is to get an improved CAD by removing anomalies that are false positive.

**Model and Method**

Researchers looked at using big data for anomaly detection in the stock market by extracting information from Twitter. They are particularly interested in techniques to extract information from unstructured data from tweets.



The figure above is a high level overview of the process.

1. extract market data

2. predict anomalies from the market data

3. extract tweets about the market from Twitter

4. prep tweet data by labeling them as positive, neutral and negative

5. build sentiment analysis model on labeled tweets

6. predict sentiment of each stock using the model

7. filter irrelevant anomalies

The typical approach to sentiment analysis is using a bag-of-words (BOW) approach, where each document is represented by a vector of words. Documents can be represented as a matrix where each row is a different document. There are a few different methods that use a bag-of-words approach, lexicon based that is unsupervised where each word is scored and the sum of the score for the document ids the sentiment of the document, and machine learning approach where the document or the words in the document are labeled and used as features in classifiers.

Social media is generating massive amounts of data and can provide insight into the behavior of the masses, and twitter data is becoming more and more popular in financial forecasting. The researchers used two data sets in the study, twitter data and stock data. They extracted data for the Oil and Gas sector of the S&P500 between Jun 22 and July 27 of 2016 and the corresponding tweets for each stock over that time frame. Stock data was pulled from Reuters and twitter data was pulled by using twitters search API looking for tweets that use the $ symbol followed by a stock ticker. The data is then preprocessed by tokenizing words, removing symbols and stopwords and noting emojis.

From there 3 different classifiers were used to determine tweet sentiment, Naive Bayes, Maximum Entropy, and Support Vector Machines and the same features are applied to each classifier. From there the sentiment is checked for the previous day when a anomaly is detected in the price of a stock and considered a false positive if the sentiment confirms the change in stock price.

**Results**

The researchers then propose a two step anomaly detection process. Anomalies are predicted on a given time series and then checked using sentiment analysis to remove false positives. They found that using feature selection (the words being used in the BOW approach) matters a lot, in that more features is more predictive up to about 10,000 features when the performance starts to decay. Ultimately they found they could correctly id 28% false positives that were previously missed.

**Citation**

**Predictive Sentiment Analysis of Tweets: A Stock Market Application**, *by: Jasmina Smailović, Miha Grčar, Nada Lavrač, Martin Žnidaršič*

Golmohammadi K., Zaiane O.R. (2017) Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation. In: Bellatreche L., Chakravarthy S. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2017. Lecture Notes in Computer Science, vol 10440. Springer, Cham