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

Koosha Golmohammadi and Osmar R. Zaiane

University of Alberta
Edmonton,Alberta, Canada
{golmoham, zaiane}@ualberta.ca

Abstract. In this paper, We propose a formalized method to improve the performance of Contextual Anomaly Detection (CAD) for detecting stock market manipulation using Big Data techniques. The method aims to improve the CAD algorithm by capturing the expected behaviour of stocks through sentiment analysis of tweets about stocks. The extracted insights are aggregated per day for each stock and transformed to a time series. The time series is used to eliminate false positives from anomalies that are detected by CAD. We present a case study and explore developing sentiment analysis models to improve anomaly detection in the stock market. The experimental results confirm the proposed method is effective in improving CAD through removing irrelevant anomalies by correctly identifying 28% of false positives.

1 Introduction

Market capitalization exceeded $1.5 trillion in Canada and $25 trillion in USA in 2015 [1] (GDP of Canada and USA in 2015 were $1.5 and $17 trillion respectively). Protecting market participants from fraudulent practices and providing a fair and orderly market is a challenging task for regulators. 233 individuals and 117 companies were prosecuted in 2015, resulting in over $138 million in fines, compensation, and disgorgement in Canada. However, the effect of fraudulent activities in securities markets and financial losses caused by such practices is far greater than these numbers suggest as they impact public and market participants trust. Market manipulation and price rigging remain the biggest concerns of investors in today's market, despite fast and strict responses from regulators and exchanges to market participants that pursue such practices. Market manipulation is forbidden in Canada[2] and the United States[3]. We define market manipulation in securities (based on the widely accepted definition in academia and industry) as: *"market manipulation involves intentional attempts to*

[1] http://data.worldbank.org/indicator/CM.MKT.LCAP.CD
[2] Bill C-46 (Criminal Code, RSC 1985, c C-46, s 382, 1985)
[3] Section 9(a)(2) of the Securities Exchange Act (SECURITIES EXCHANGE ACT OF 1934, 2012)

deceive investors by affecting or controlling the price of a security or interfering with the fair market to gain profit."

The industry's existing approach for detecting market manipulation is top-down and is based on a set of known patterns and predefined thresholds. Market data such as price and volume of securities (i.e. the number of shares or contracts that are traded in a security) are monitored using a set of rules and red-flags trigger notifications. Then, transactions associated with the detected periods are investigated further, as they might be associated with fraudulent activities. These methods are based on expert knowledge but suffer from two issues: i) detection of abnormal periods that are not associated with known symptoms (i.e. unknown manipulative schemes), and ii) adaption to changing market conditions whilst the amount of transactional data is exponentially increasing which makes designing new rules and monitoring the vast data challenging. These issues lead to an increase in false negatives (i.e. there is a significant number of abnormal periods that are left out of the investigation). Data mining methods may be used as a bottom-up approach to detect market manipulation by identifying unusual patterns and data points that merit further investigation, as they are potentially associated with fraudulent activities.

A time series $\{ x_t, t \in T_0 \}$ is the realization of a stochastic process $\{ X_t, t \in T_0 \}$. For our purposes, set T (i.e. the set of time points) is a discrete set and the real-valued observations x_t are recorded on fixed time intervals. Though there has been extensive work on anomaly detection [7], the majority of the techniques look for individual objects that are different from normal objects but do not take the temporal aspect of data into consideration. For example, a conventional anomaly detection approach based on values of data points may not capture anomalous data points in the ECG data where a subsequence with values close to the mean does not follow expected motifs. Therefore, the temporal aspect of data should be considered in addition to the amplitude and magnitude values. Time series anomaly detection methods are successfully applied to different domains including management [24], detecting abnormal conditions in ECG data [15], detecting shape anomalies [25], and credit card fraud detection [11]. Contextual anomalies in time series are data points that are anomalous in a "specific context but not otherwise". For example, Edmonton's average temperature during 2013 was 4.03 degrees Celsius, while the same value during January would be an anomaly (i.e. contextual anomaly). A set of anomalous data points creates an anomalous subsequence (motif). The context is defined both in terms of similarity to the neighbourhood data points of each time series and similarity of time series pattern with respect to the rest of time series in the group. Local anomaly detection methods are particularly useful in non-homogeneous datasets and datasets with changing underlying factors such as financial data. The major motivation for studying local anomaly detection is the development of methods for detecting local anomalies/outliers in complex time series that do not follow a seasonal pattern and are non-parametric, meaning it is difficult to fit a polynomial or deterministic function to the time series data. This is a challenging problem in domains with complex time series such as stock market. Market manipulation

periods have been shown to be associated with anomalies in the time series of assets [23], yet the development of effective methods to detect such anomalies remains a challenging problem.

In this paper, we present a formalized method to improve the performance of Contextual Anomaly Detection algorithm that we proposed in our previous work [12]. CAD utilizes an unsupervised learning approach towards detecting anomalies given a set of similar time series. First, a subset of time series is selected based on the window size parameter, Second, a centroid is calculated representing the expected behaviour of time series of the group. Then, the centroid values are used along with the correlation of each time series with the centroid to predict the values of the time series. The proposed method improves recall from 7% to 33% compared to kNN and random walk without compromising precision. The experiments were on S&P industry sectors over 40 years both using daily and weekly data. However, the precision of CAD, kNN and random walk are 0.5% in these experiments (the baseline is less than 0.04% because the number of anomalies in the data is a tiny percentage of samples). We attempt to address this issue by aggregating data from Twitter to reduce the number of false positives that CAD produces.

2 Methods

We adopted big data techniques to improve the performance of the Contextual Anomaly Detection (CAD) method by eliminating false positives. A formalized method is developed to explore the market participants' expectation for each detected datapoint. This information is used to filter out irrelevant items (false positives). Big data techniques are often used to predict consumer behaviour, primarily using social network services such as Twitter, Facebook, Google+ and Amazon reviews. We utilized big data for a novel application in time series anomaly detection, specifically stock market anomalies, by extracting information from Twitter. This information can be integrated into the anomaly detection process to improve the performance of the proposed anomaly detection by eliminating irrelevant anomalies. Although anomalies that are captured using anomaly detection methods represent anomalous data points and periods, some of them may be irrelevant, because there might be a reasonable cause for the anomaly outside time series of market data (for example a news release about a company before the event may explain the abnormal stock return). Using big data techniques to integrate additional information to improve anomaly detection is particularly challenging in securities fraud detection, which are typical challenges in big data problems - velocity, volume, and variability. We are specifically interested in big data techniques to extract information from unstructured data from tweets.

We developed a case study to investigate sentiment analysis on Twitter to improve anomaly detection in the stock market. Figure 1 describes a high-level overview of the process flow in the case study:

A.1) extracting market data for Oil and Gas stocks of S&P 500,

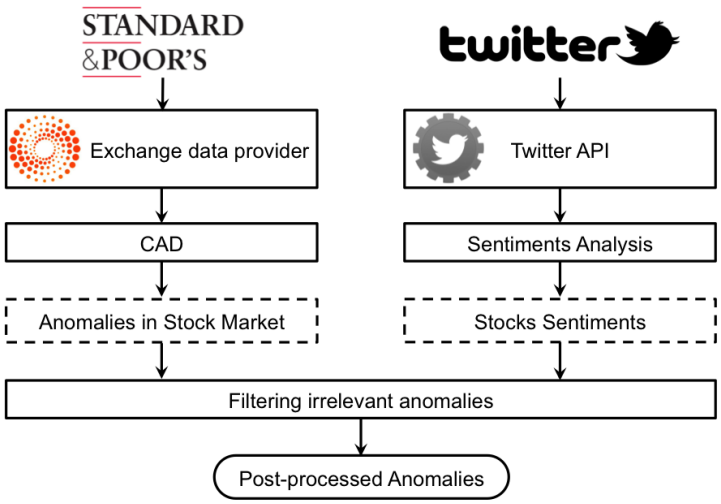

Fig. 1. Utilizing Twitter Data to Improve Anomaly Detection in the Stock Market

A.2) predicting anomalies in the Oil and Gas stocks,

B.1) extracting tweets from Twitter for the Oil and Gas stocks in S&P 500,

B.2) preparing a training dataset by extracting tweets for stocks in the Information Technology sector (this data is manually labelled as negative, neutral and positive by an individual who was not involved in developing the methods to preserve fairness of the study),

B.3) building a model for sentiment analysis of tweets that is trained and tested using tweets on stocks (i.e. labelled tweets on the Information Technology sector),

B.4) predicting sentiment of each stock per day using the sentiment analysis model (this produces a time series of sentiments for any given stock returns time series), and,

C.1) filtering irrelevant anomalies based on the respective sentiment on the previous day of every detected anomaly.

2.1 Sentiment Analysis on Twitter

Sentiment analysis is the process of computationally identifying and categorizing people's opinions towards different matters such as products, events, organizations, etc. [4]. Several experiments confirm prediction capabilities of sentiment analysis of social media content such as predicting the size of markets [6] and unemployment rate [1]. Some research works suggest analyzing news and social media such as blogs, micro-blogs, etc. to extract public sentiments could improve predictions in the financial market [22]. Feldman et al. [10] proposed a hybrid approach for stock sentiment analysis based on news articles of companies.

Twitter is the most popular micro-blogging platform. Twitter's technology and popular brand enable millions of people to share their opinions on a variety of topics such as their well-being, politics, products, social events, market conditions and stock market. The flexible architecture and APIs enable researchers and industry to use Twitter for various prediction purposes. Twitter was used to predict movie ticket sales in their opening week with the accuracy of 97.3% [2]. We utilize Twitter in this paper to identify people's opinions about stocks. Sentiment analysis techniques could be used to automatically analyze unstructured data such as tweets in the neighbourhood time period of a detected anomaly. Textual analysis has a long history in the literature [9], however, categorization through sentiments is more recent [20]. The typical approach for representing text for computational processes is based on a the bag-of-words (BOW) [9] where each document is represented by a vector of words. This bag-of-words is called a collection of unigrams. This approach assumes a euclidean space of unigrams that are independent of each other. Thus, documents can be represented as a matrix where each row represents a document. Sentiment analysis methods can be divided into two groups while both use BOW:

1. **lexicon based method** that is an unsupervised approach where a polarity score is assigned to each unigram in the lexicon and the sum of all polarity scores of the text identifies the overall polarity of the text,
2. **machine learning approach** that is a supervised approach where the unigrams or their combinations (i.e. N-grams) are used as features by classifiers.

Social media are increasingly reflecting and influencing the behaviour of other complex systems such as the stock market. Users interactions in social media are generating massive datasets that could explain the "collective behaviour in a previously unimaginable fashion" [14]. We can identify interests, opinions, concerns and intentions of the global population with respect to various social, political, cultural and economic phenomena. Twitter, the most popular micro-blogging platform on internet, is at the forefront of the public commenting about different phenomena. "Twitter data is becoming an increasingly popular choice for financial forecasting" [13]. Researchers have investigated whether the daily number of tweets predicts the S&P 500 stock return [19]. Ruiz et al. used a graph-based view of Twitter data to study the relationship between Twitter activities and the stock market [21]. Some research works utilize textual analysis on Twitter data to find relationships between mood indicators and the Dow Jones Industrial Average (DJIA) [5]. However, the correlation levels between prices and sentiments on Twitter remains low in empirical studies especially when textual analysis is required. More recently, Bartov et al. found aggregated opinions on Twitter can predict quarterly earnings of a given company [3]. These observations suggest a more complicated relationship between sentiments on Twitter and stock returns. Every day, a huge number of messages are generated on Twitter which provides an unprecedented opportunity to deduce the public opinions for a wide range of applications [16]. We intend to use the polarity of tweets to identify the expected behaviour of stocks in the public eyes. Here are some example tweets upon querying the keyword "$xom".

- $XOM flipped green after a lot of relative weakness early keep an eye on that one she's a big tell.
- #OILALERT $XOM »Oil Rises as Exxon Declares Force Majeure on #Nigeria Exports
- Bullish big oil charts. No voice - the charts do the talking. http://ln.is/www.youtube.com/ODKYG $XLE $XOM $CVX $RDS $SLB @TechnicianApp

The combination of the $ sign along with a company ticker is widely used on Twitter to refer to the stock of the company. As shown, the retrieved tweets may be about Exxon Mobil's stock price, contracts and activities. These messages are often related to people's sentiments about Exxon Mobil Corp., which can reflect its stock trading. We propose using Twitter data to extract collective sentiments about stocks to filter false positives from detected anomalies in stocks. We study the sentiment of stocks at time $t-1$ where t is the timestamp of a detected anomaly. A sentiment that aligns with the stock return at time t confirms the return (i.e. aligns with expected behaviour) thus, indicates the detected anomaly is a false positive. We introduce a formalized method to improve anomaly detection in stock market time series by extracting sentiments from tweets and present empirical results through a case study on stocks of an industry sector of S&P 500.

2.2 Data

We use two datasets in this case study: Twitter data and market data. We extracted tweets on the Oil and Gas industry sector of S&P 500 for 6 weeks (June 22 to July 27 of 2016) using the Twitter search API. Table 1 shows the list of 44 Oil and Gas stocks in S&P 500 and the respective number of tweets constituting 57,806 tweets.

Table 1: Tweets about Oil and Gas industry sector in S&P 500

Ticker	Company	cashtag	Tweets
APC	ANADARKO PETROLEUM	$APC	1052
APA	APACHE	$APA	1062
BHI	BAKER HUGHES	$BHI	1657
COG	CABOT OIL & GAS 'A'	$COG	736
CAM	CAMERON INTERNATIONAL	$CAM	255
CHK	CHESAPEAKE ENERGY	$CHK	4072
CVX	CHEVRON	$CVX	3038
COP	CONOCOPHILLIPS	$COP	1912
CNX	CONSOL EN.	$CNX	1023
DNR	DENBURY RES.	$DNR	1008
DVN	DEVON ENERGY	$DVN	1459
DO	DIAMOND OFFS.DRL.	$DO	1227
ESV	ENSCO CLASS A	$ESV	825
EOG	EOG RES.	$EOG	1149

EQT	EQT	$EQT	669
XOM	EXXON MOBIL	$XOM	5613
FTI	FMC TECHNOLOGIES	$FTI	511
HAL	HALLIBURTON	$HAL	2389
HP	HELMERICH & PAYNE	$HP	838
HES	HESS	$HES	917
KMI	KINDER MORGAN	$KMI	2138
MRO	MARATHON OIL	$MRO	2063
MPC	MARATHON PETROLEUM	$MPC	950
MUR	MURPHY OIL	$MUR	689
NBR	NABORS INDS.	$NBR	384
NOV	NATIONAL OILWELL VARCO	$NOV	827
NFX	NEWFIELD EXPLORATION	$NFX	779
NE	NOBLE	$NE	1102
NBL	NOBLE ENERGY	$NBL	583
OXY	OCCIDENTAL PTL.	$OXY	671
OKE	ONEOK	$OKE	651
BTU	PEABODY ENERGY	$BTU	186
PSX	PHILLIPS 66	$PSX	1205
PXD	PIONEER NTRL.RES.	$PXD	955
QEP	QEP RESOURCES	$QEP	713
RRC	RANGE RES.	$RRC	860
RDC	ROWAN COMPANIES CL.A	$RDC	476
SLB	SCHLUMBERGER	$SLB	1962
SWN	SOUTHWESTERN ENERGY	$SWN	1912
SE	SPECTRA ENERGY	$SE	421
TSO	TESORO	$TSO	1086
RIG	TRANSOCEAN	$RIG	1846
VLO	VALERO ENERGY	$VLO	1464
WMB	WILLIAMS COS.	$WMB	2471

There are two options for collecting tweets from Twitter: the Streaming API and the Search API. The Streaming API provides a real-time access to tweets through a query. It requires a connection to the server for a stream of tweets. The free version of Streaming API and the Search API provide access to a random sampling of about 1% of all tweets [4]. While the syntax of responses for the two APIs is very similar, there are some differences such as the limitation on language specification on queries in Streaming API. We used the Search API to query recent English tweets for each stock in the Oil and Gas industry sector of S&P 500 using its cashtag. Twitter unveiled the cashtag feature in 2012 enabling users to click on a $ followed by a stock ticker to retrieve tweets about the stock. The feature has been widely adopted by users when tweeting about equities. We

[4] The firehose access on Streaming API provides access to all tweets. This is very expensive and available upon case-by-case requests from Twitter.

account for the search API rate limits by sending many requests for each stock with 10-second delays. The batch process runs daily to extract tweets and store them in a database.

The market data for stocks in Oil and Gas industry sector is extracted from Thompson Reuters. The stock returns are calculated as $R_t = (P_t - P_{t-1}/P_{t-1})$ where R_t, is the stock return and P_t and P_{t-1} are the stock price on days t and $t-1$ respectively.

2.3 Data Preprocessing

The JSON response for a search query on Twitter APIs (e.g. $msft) includes several pieces of information such as username, time, location, retweets, etc. [5] For our purposes, we focus on the timestamp and tweet text. We store tweets in a mongoDB database ensuring each unique tweet is recorded once. mongoDB is an open source NoSQL database which greatly simplifies tweet storage, search, and recall eliminating the need of a tweet parser. Tweets often include words and text that are not useful and potentially misleading in sentiment analysis. We remove URLs usernames and irrelevant texts and symbols. Our preprocessing includes three processes:

- **Tokenization** that involves extracting a list of individual words (i.e. bag of words) by splitting the text by spaces. These words are later used as features for the classifier.
- **Removing Twitter Symbols** which involves filtering irrelevant text out such as the immediate word after @ symbol, arrow, exclamation mark, etc.
- **Removing Stopwords** that involves removing words such as “the”, “to”, “in”, “also”, etc. by running each word against a dictionary.
- **Recording smiley faces** which involves translating smiley and sad faces to a positive and negative expression in the bag of words.

2.4 Modelling

We adopted three classifiers for determining sentiment of tweets including Naive Bayes, Maximum Entropy and Support Vector Machines. The same features are applied to all classifiers. The anomalous time series of $\{ \eta_k, 0 \leq k \leq n \}$ for the time series $\{ x_1, x_2, \ldots, x_n \}$ where η_k represents an anomaly on day k in the time series (i.e. stock) X. We check sentiment of the stock on day $k-1$ given η_k. We consider the detected anomaly as a false positive, if the sentiment confirms the change in stock return on day k, however, a sentiment that is in disagreement with the return on the next day implies unexpected stock behaviour, thus anomaly. We study the proposed method for filtering out false positives within detected anomalies by first, running Contextual Anomaly Detection (CAD) method on an anomaly-free dataset, second, removing detected anomalies in the first step that do not conform with their respective sentiment on Twitter. Figure 2 describes

[5] https://dev.twitter.com/rest/reference/get/search/tweets

an example of stock sentiments on Twitter and anomalies that are detected on XOM (Exxon Mobil). The figure shows 4 anomalies (represented by red circles) that are detected on XOM along with the stock sentiment on Twitter for each day (days with no bars have the neutral sentiment). The data points on June 24 and July 21 are declared irrelevant because the stock's sentiments on the day before these dates confirm the change direction on the next day. However, other two anomalies (July 7 and 15) remain relevant because the sentiments on the day before the anomalies do not confirm the change direction in the stock return.

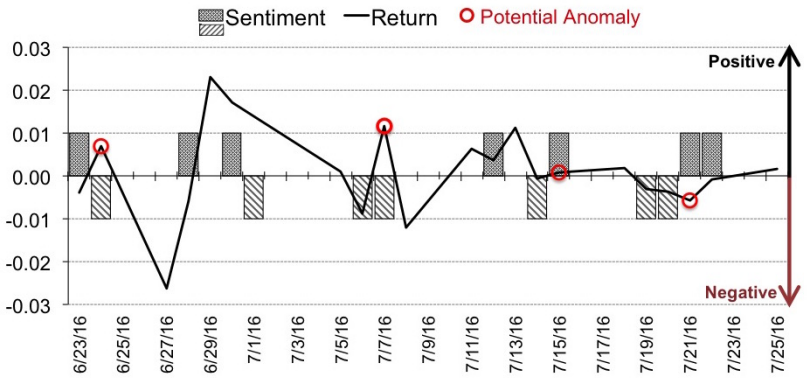

Fig. 2. Identifying false positives in detected anomalies on Exxon Mobil (XOM)

We found through our preliminary experiments that sentiment analysis using classifiers that are trained on movie reviews or generic tweets that are widely used in literature perform poorly for stock tweets. This is due to different corpus and linguistics that are specific to stock market. We developed a training dataset that is labelled manually to address this issue. This dataset includes 66 stocks in the Information Technology industry sector of S&P 500 and respective tweets constituting over 6,000 tweets. We manually labelled over 2,000 tweets from this dataset. We also used StockTwits [6], a widely popular social media platform that is designed for sharing ideas between investors and traders, to extract messages that are labelled by stock market participants. We developed a tool to query StockTwits for a given stock and extract relevant messages. Then, messages that are labelled by their poster as *Bearish* and *Bullish* are mapped to negative and positive sentiments in our code. The training data is labelled manually with three sentiment labels: negative, neutral and positive. This data is used to train the classifiers that we used. The testing dataset is tweets about the Oil and Gas industry sector of S&P 500. Table 1 shows the list of 44 Oil and Gas stocks in the testing dataset S&P 500 and the respective number of tweets constituting 57,706 tweets in total.

[6] http://stocktwits.com/

2.5 Feature Selection

Feature selection is a technique that is often used in text analysis to improve performance of results by selecting the most informative features (i.e. words). Features that are common across all classes contribute little information to the classifier. This is particularly important as the number of features grow rapidly with increasing number of documents. The objective is using the words that have the highest information gain. Information gain is defined as the frequency of the word in each class compared to its frequency in other classes. For example, a word that appears in the positive class often but rarely in the neutral and negative classes is a high information word. Chi-square is widely used as a measure of information gain by testing the independence of a word occurrence and a specific class:

$$\frac{N(O_{w_p c_p} * O_{w_n c_n} - O_{w_n c_p} * O_{w_p c_n})^2}{O_{w_p} * O_{w_n} * O_{c_p} * O_{c_n}} \tag{1}$$

Where $O_{w_p c_p}$ is the number of observations of the word w in the class c and $O_{w_p c_n}$ is the number of observations of the word w in other classes (i.e. class negative). This score is calculated for each word (i.e. feature) and used for ranking them. High scores indicate the null hypothesis H_0 of independence should be rejected. In other words, the occurrence of the word w and class c are dependent thus the word (i.e. feature) should be selected for classification. It should be noted that Chi-square feature selection is slightly inaccurate from statistical perspective due to the one degree of freedom. Yates correction could be used to address the issue, however, it would make it difficult to reach statistical significance. This means a small number of features out of the total selected features would be independent of the class. Manning et al. showed these features do not affect the performance of the classifier [18].

2.6 Classifiers

Naive Bayes: A Naive Bayes classifier is a probabilistic classifier based on the Bayes Rule $P(c|\tau) = \frac{P(\tau|c)P(c)}{P(\tau)}$ where $P(c|\tau)$ is the probability of class c being negative, neutral or positive given the tweet τ. The best class is the class that maximizes the probability given tweet τ:

$$C_{MAP} = \arg\max_{c \in C} P(\tau|c)P(c) \tag{2}$$

where $P(\tau|c)$ can be calculated using the bag of words as features resulting in

$$C_{MAP} = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c) \tag{3}$$

$P(c)$ can be calculated based on the relative frequency of each class in the corpus or dataset. There are two simplifying assumption in Naive Bayes which make calculating $P(x_1, x_2, \ldots, x_n|c)$ straightforward, i) position of the words do

not matter, and ii) the feature probabilities $P(x_i|c_j)$ are independent given the class c:

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) \bullet P(x_2|c) \bullet \cdots \bullet P(x_n|c) \tag{4}$$

in other words, we have the Multinomial Naive Bayes equation as

$$C_{NB} = \arg\max_{c \in C} P(c_j) \prod_{x \in X} P(x|c) \tag{5}$$

Maximum Entropy: MaxEnt eliminates the independence assumptions between features and in some problems outperforms Naive Bayes. MaxEnt is a probabilistic classifier based on the Principle of Maximum Entropy. Each feature corresponds to a constraint in a maximum entropy model. MaxEnt classifier computes the maximum entropy value from all the models that satisfy the constraints of the features for the given training data, and selects the one with the largest entropy. The MaxEnt probability estimation is computed using

$$P(c|f) = \frac{1}{Z(f)} \; exp\left(\sum_i \lambda_{i,c} \, F_{i,c}(f, c)\right) \tag{6}$$

where $Z(f)$ is a normalization function and $F_{i,c}$ is a binary function that takes the input feature f for the class c. λ is a vector of weight parameters that is updated iteratively to satisfy the tweets feature while continuing to maximize the entropy of the model [8]. The iterations eventually converge the model to a maximum entropy for the probability distribution. The binary function $F_{i,c}$ is only triggered when a certain feature exists and the sentiment is hypothesized in a certain class:

$$F_{i,c}(f, c') = \begin{cases} 1 & \text{if } n(f) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Support Vector Machines: SVM is a linear classification algorithm which tries to find a hyperplane that separates the data in two classes as optimally as possible. the objective is maximizing the number of correctly classified instances by the hyperplane while the margin of the hyperplane is maximized. The hyperplane representing the decision boundary in SVM is calculated by

$$(\boldsymbol{w} \cdot \boldsymbol{x}) + b = \sum_i y_i \alpha_i (\boldsymbol{x_i} \cdot \boldsymbol{x}) + b = 0 \tag{8}$$

where weight vector $\boldsymbol{w} = (w_1, w_2, .., w_n)$ which is the normal vector defining the hyperplane is calculated using the n-dimensional input vector $\boldsymbol{x_i} = (x_{i1}, x_{i2}, .., x_{in})$, outputting the value y_i. α_i terms are the Lagrangian multipliers. Calculating w using the training data gives the hyperplane which can be used to classify the input data instance $\boldsymbol{x_i}$. If $\boldsymbol{w} \cdot \boldsymbol{x_i} + b \geq 0$ then the input data instance is labelled positive (the class we are interested in), otherwise it belongs to the negative class (all of the other classes). It should be noted that although SVM is

a linear classifier (as Naive Bayes and Maximum Entropy are) it is a powerful tool to classify text because text documents are typically considered as a linear dataset. It is possible to use Kernel functions for datasets that are not linearly separable. The Kernel is used to map the dataset to a higher dimensional space where the data could be separated by a hyperplane using classical SVM.

There are two approaches for adopting SVM for a classification problem with multiple classes such as sentiment analysis with the classes negative, neutral and positive: i) one-vs-all where an SVM classifier is built for each class, and ii) one-vs-one where an SVM classifier is built for each pair of classes resulting in $M(M-1)/2$ for M classes. We used the latter for classifying sentiments using SVM. In the one-vs-all approach, the classifier labels data instances positive for the class that we are interested in and the rest of instances are labelled negative. A given input data instance is classified with classifier only if it is positive for that class and negative for all other classes. This approach could perform poorly in datasets that are not clustered as many data instances that are predicted positive for more than one class, will be unclassified. The one-vs-one approach is not sensitive to this issue as a data instance is categorized in the class with the most data instances, however, the number of classes can grow rapidly for problems with many classes (i.e. higher numbers of M).

2.7 Classifier Evaluation

We trained classifiers using specifically stock tweets that are carefully labelled manually. We asked a person who has not been involved with training data to label the testing dataset. The testing data includes 1332 stock tweets that are manually labelled. We used 5-fold cross validation for training the classifiers that is sampling the data into 5 folds and using 4 folds for training and 1 fold for testing. This process is repeated 5 times and the performance results are averaged. We used precision and recall for each class in addition to classification accuracy as performance measures to evaluate the classifiers. The precision of a classifier [7] for a given class represents the fraction of the classified tweets that belong to the class, while recall [8] represents the fraction of tweets that belong to the class out of all tweets that belong to the class. The precision for a class measures the exactness or quality, whereas recall measures the completeness or quantity. The classifier with the highest performance is used to predict sentiment of stocks in Oil and Gas industry sector (see Table 1 for the list of stocks in the Oil and Gas sector).

2.8 Calculating Polarity for each Stock

The Twitter sentiments of stocks are predicted using an SVM classifier that is trained using labelled tweets about stocks. First, the classifier predicts sentiment of each tweet (i.e. negative, neutral and positive). Then, the polarity for each stock is computed using time series of negative, neutral and positive tweets:

[7] $TP/(TP+FP)$
[8] $TP/(TP+FN)$

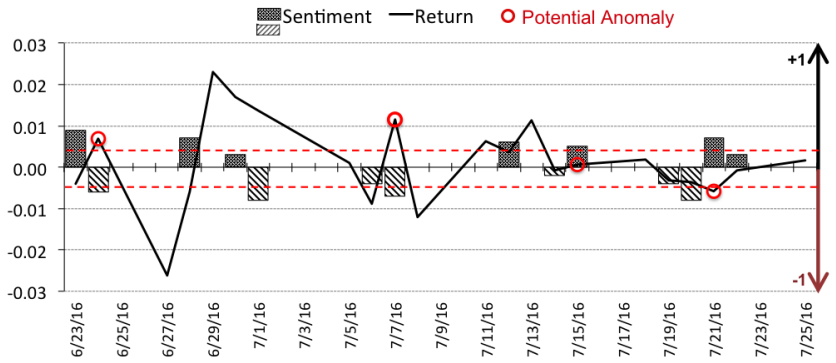

Fig. 3. Polarity of Exxon Mobil stock per day along with potential anomalies that CAD produces

- Negative tweets, tw_d^-: the number of negative tweets on day d
- Neutral tweets, tw_d^0: the number of neutral tweets on day d
- Positive tweets, tw_d^+: the number of positive tweets on day d

The polarity for each stock on a given day is the difference between the number of positive and negative tweets as a fraction of non-neutral tweets. More formally

$$P_{s_d} = \frac{tw_d^+ - tw_d^-}{tw_d^+ + tw_d^-} \tag{9}$$

where P_{s_d} is the polarity of stock s on day d. Figure 3 shows the aggregated polarity of Exxon Mobil. The red dashed lines represent the parameter *sentThreshold* that we define to control for the minimum magnitude of polarity that is required for declaring a potential anomaly a false positive. For example, the method would not include the polarity of Exxon Mobil on July 14 as an indicator to accept or reject the potential anomaly on July 15 as a false positive because its value is below the threshold. This parameter can be set during preliminary tests by trying a grid on *sentThreshold* (e.g. 0.2, 0.3, etc.).

3 Results and Discussion

We propose a two-step anomaly detection process. First, the anomalies are predicted on a given set of time series (i.e. stocks in an industry sector) using Contextual Anomaly Detection (CAD). Second, the anomalies are vetted using sentiment analysis by incorporating data in addition to market data. This process gives a list of anomalies that are filtered using data on Twitter. The first step is based on the unsupervised learning algorithm CAD and the second step, relies on state-of-the-art supervised learning algorithms for sentiment analysis on

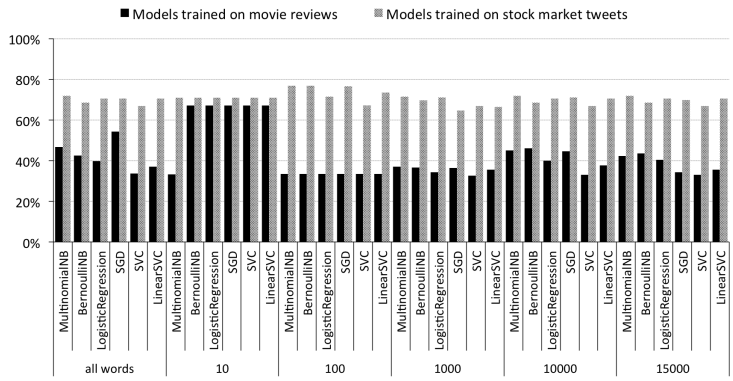

Fig. 4. Accuracy of sentiment analysis models using training datasets in movie reviews and stock market

unstructured data on Twitter. We developed a set of experiments for this case study on the Oil and Gas sector of S&P 500 for the period of June 22 to July 27. The correlation of stocks during this 6-week period for the case study is quite high as we expect within an industry sector (70% stocks within the sector have a correlation higher than 0.7)

We studied several other classifiers in addition to the three classifiers that we introduced in Section 2.4 (i.e. Multinomial Naive Bayes, MaxEnt, also known as Logistic Regression, and SVM) to build a sentiment analysis model including Bernoulli Naive Bayes, Stochastic Gradient Descent (SGD) and C-Support Vector (SVC). Furthermore, we investigated the performance of sentiment analysis models using different number of features (i.e. 10, 100, 1000 etc. words).

Movie reviews data is typically used for sentiment analysis of short reviews as well as tweets [17]. This dataset includes movie reviews that are collected from IMDB [9]. Our experiments show that sentiment analysis models for stock tweets that are trained using this standard dataset perform poorly (see Figure 4). The results confirm our hypothesis that training data that is out of context is inappropriate for sentiment analysis of short text samples, particularly on Twitter. We developed a tool to extract labelled data from StockTwits [10] to address this issue (see Section 2.3 for more information on data). Figure 4 illustrates that these models outperform models which are trained on movie review data consistently.

We observe that the number of features is an important parameter in the performance of sentiment analysis models. The results show using more features improves the performance results. However, performance of the models decays after hitting a threshold of about 10,000 features. This reiterates our hypothesis on utilizing feature selection to improve sentiment analysis on Twitter. We studied the impact of proposed method in filtering false positives of CAD by first, running

[9] http://www.imdb.com/reviews/
[10] http://stocktwits.com/

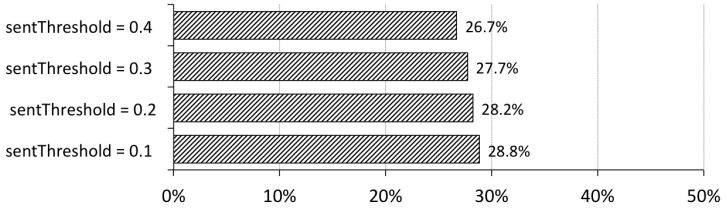

Fig. 5. Filtering irrelevant anomalies using sentiment analysis on Oil and Gas sector

CAD on returns of Oil and Gas stocks during June 22 to July 27 of 2016 with no injected anomalies. The predicted anomalies would be false positives because the S&P 500 data is anomaly-free as we explained. Then, using the proposed method we measured how many of false positives are filtered. CAD predicts 261 data points as anomalous given stock market data for the case study (out of 1,092 data points). Figure 5 describes the percentage of irrelevant anomalies that are correctly filtered by sentiment analysis of tweets about stocks. *sentThreshold* is a parameter we use when comparing the aggregated sentiment values for a given stock per day. The results confirm that the proposed method is effective in improving CAD through removing irrelevant anomalies by correctly identifying 28% of false positives.

References

1. Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., Shapiro, M.D.: Using social media to measure labor market flows. Tech. rep., National Bureau of Economic Research (2014)
2. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. vol. 1, pp. 492–499. IEEE (2010)
3. Bartov, E., Faurel, L., Mohanram, P.S.: Can twitter help predict firm-level earnings and stock returns? Available at SSRN 2782236 (2016)
4. Bing, L.: Sentiment analysis: A fascinating problem. Morgan and Claypool Publishers pp. 7–143 (2012)
5. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM 11, 450–453 (2011)
6. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
7. Chandola, V., Banerjee, a., Kumar, V.: Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering 24(5), 823–839 (May 2012), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5645624
8. Daumé III, H.: Notes on cg and lm-bfgs optimization of logistic regression. Paper available at https://www.umiacs.umd.edu/ hal/docs/daume04cg-bfgs.pdf pp. 1–7 (2004)
9. Dillon, M.: Introduction to modern information retrieval: G. salton and m. mcgill (1983)

10. Feldman, R., Rosenfeld, B., Bar-Haim, R., Fresko, M.: The stock sonarâĂŤsentiment analysis of stocks based on a hybrid approach. In: Twenty-Third IAAI Conference. pp. 1642–1647 (2011)
11. Ferdousi, Z., Maeda, A.: Unsupervised Outlier Detection in Time Series Data, p. 121. IEEE (2006), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1623916
12. Golmohammadi, K., Zaiane, O.R.: Time series contextual anomaly detection for detecting market manipulation in stock market. In: The 2015 Data Science and Advanced Analytics (DSAA'2015). pp. 1–10. IEEE (2015)
13. Graham, M., Hale, S.A., Gaffney, D.: Where in the world are you? geolocation and language identification in twitter. The Professional Geographer 66(4), 568–578 (2014)
14. King, G.: Ensuring the data-rich future of the social sciences. science 331(6018), 719–721 (2011)
15. Lin, J., Keogh, E., Fu, A., Herle, H.: Approximations to Magic: Finding Unusual Medical Time Series, pp. 329–334. IEEE (2005)
16. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1–167 (2012)
17. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-1015
18. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to information retrievalâĂŰ cambridge university press, 2008. Ch 20, 405–416
19. Mao, Y., Wei, W., Wang, B., Liu, B.: Correlating s&p 500 stocks with twitter data. In: Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research. pp. 69–72. ACM (2012)
20. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 341–349. ACM (2002)
21. Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time series with micro-blogging activity. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 513–522. ACM (2012)
22. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Transactions on Information Systems (TOIS) 27(2), 12 (2009)
23. Song, Y., Cao, L., Wu, X., Wei, G., Ye, W., Ding, W.: Coupled behavior analysis for capturing coupling relationships in group-based market manipulations. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 976–984. ACM (2012)
24. Sriastava, A., et al.: Discovering system health anomalies using data mining techniques pp. 1–7 (2005)
25. Wei, L., Keogh, E., Xi, X.: Saxually explicit images: finding unusual shapes. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. pp. 711–720. IEEE (2006)