Information Sciences 285 (2014) 181–203

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Stream-based active learning for sentiment analysis in the financial domain

Jasmina Smailović [a,b,*], Miha Grčar [a,b], Nada Lavrač [a,b,c], Martin Žnidaršič [a,b]

[a] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[c] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

ARTICLE INFO

*Article history:*
Received 22 March 2013
Received in revised form 31 March 2014
Accepted 17 April 2014
Available online 25 April 2014

*Keywords:*
Predictive sentiment analysis
Stream-based active learning
Stock market
Twitter
Positive sentiment probability
Granger causality

ABSTRACT

Studying the relationship between public sentiment and stock prices has been the focus of several studies. This paper analyzes whether the sentiment expressed in Twitter feeds, which discuss selected companies and their products, can indicate their stock price changes. To address this problem, an active learning approach was developed and applied to sentiment analysis of tweet streams in the stock market domain. The paper first presents a static Twitter data analysis problem, explored in order to determine the best Twitter-specific text preprocessing setting for training the Support Vector Machine (SVM) sentiment classifier. In the static setting, the Granger causality test shows that sentiments in stock-related tweets can be used as indicators of stock price movements a few days in advance, where improved results were achieved by adapting the SVM classifier to categorize Twitter posts into three sentiment categories of positive, negative and neutral (instead of positive and negative only). These findings were adopted in the development of a new stream-based active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet streams. To this end, a series of experiments was conducted to determine the best querying strategy for active learning of the SVM classifier adapted to sentiment analysis of financial tweet streams. The experiments in analyzing stock market sentiments of a particular company show that changes in positive sentiment probability can be used as indicators of the changes in stock closing prices.

## 1. Introduction

Predicting the value of stock market assets is a challenge investigated by numerous researchers. One of the reasons for addressing this challenge is the controversy of the efficient market hypothesis [17], which claims that stocks are always traded at their fair value. Based on this market theory, claiming that it is not possible for investors to buy undervalued stocks or sell stocks for overestimated prices, it is impossible for traders to consistently outperform the average market returns. This hypothesis is based on the assumption that financial markets are informationally efficient (i.e., that stock prices always reflect all the relevant information at investment time). The unpredictable nature of stock market prices was first investigated by Regnault [51] and later by Bachelier [4]. Fama [17], who proposed the efficient market hypothesis, also claimed that stock price movement is unpredictable and that past price movements cannot be used to forecast future stock prices.

* Corresponding author at: Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. Tel.: +386 1 4773 143.
*E-mail address:* jasmina.smailovic@ijs.si (J. Smailović).

http://dx.doi.org/10.1016/j.ins.2014.04.034

However, as the efficient market hypothesis is controversial, researchers from various disciplines (including economists, statisticians, finance experts, and data miners) have been investigating the means to predict future stock market prices. The findings vary: from those claiming that stock market prices are not predictable to those presenting opposite conclusions [9,33].

This paper addresses the described challenge in the context of the explosive growth of social media and user-generated content on the Internet. Through blogs, forums, and social networking media, more and more people share their opinions about individuals, companies, movements, or important events. Such opinions both express and evoke sentiments [49]. Recent research indicates that analysis of these online texts can be useful for trend prediction. For example, it was shown that the frequency of blog posts can be used to forecast spikes in online consumer purchasing [23]. Moreover, it was shown by Tong [72] that references to movies in newsgroups are correlated with their sales. Sentiment analysis of weblog data was successfully used to predict the financial success of movies [40]. Twitter[1] posts were also shown to be useful for predicting box-office revenues of movies before their release [3].

Twitter is currently the most popular microblogging platform [46] allowing its users to send and read short messages of up to 140 characters in length, known as *tweets*, via SMS, the Twitter website, or a range of applications for mobile devices. Twitter gained global popularity very quickly with over 500 million active users in 2012, writing over 340 million tweets daily [16,41]. Twitter data (and data from other social network websites) are very interesting because of their large volume, popularity, and capability of near-real-time publishing of individuals' opinions and emotions about any subject. Given that this massive amount of user-generated content became abundant and easily accessible, many researchers became interested in the predictive power of microblogging messages, especially in the domain of stock market prediction, prediction of election results, or prediction of the financial success of movies or books. Many of these studies use *sentiment analysis* [36,75] as a basis for prediction. The term *sentiment*, used in the context of automatic analysis of text and detection of predictive judgments from positively and negatively opinionated texts, first appeared in the papers by Das and Chen [14] and Tong [72], where the authors were interested in analyzing stock market sentiment. Even though there are many studies on predicting the phenomenon of interest using sentiment analysis of online texts, there is still an urge to develop methods and tools for adaptive dynamic sentiment analysis of microblogging posts, which would enable handling changes in such data streams. This field of research is still insufficiently explored and represents a challenge, which is addressed in this work through *active learning* [61].

This work contributes to sentiment analysis and to active learning research, and partly towards better understanding of phenomena in financial stock markets. While sentiment analysis is generally aimed at detecting the author's attitude, emotions or opinions expressed in the text, our study is concerned with the development of an approach to *predictive sentiment analysis*. With this term, we denote an approach in which sentiment analysis is used to predict a specific phenomenon or its changes, postulating that the proposed methodology for predictive sentiment analysis of streams of microblogging messages should be capable of predicting the financial phenomenon of interest. The indication that there may be a relationship between emotions and stock market prices relies on findings in psychological research which indicate that emotions are crucial to rational thinking and social behavior [13], and can influence the choice of actions. Given that the general mood of a society is propagated through social interactions, the collective social mood can be transferred through the investors to the stock market and consequently, the sentiment can be reflected in stock price movements. As a result, the stock market itself can be considered as a measure of social mood [44]. It is, thus, reasonable to expect that the analysis of the public mood can be used to predict price movements in the stock market. We hypothesize that this assumption may hold in situations when people actually express positive or negative opinions about some topic concerning the stock market, whereas in situations when people do not express opinions, but mostly neutral facts, we anticipate finding no correlations. In accordance with this hypothesis, we propose a mechanism for distinguishing opinionated (positive and negative) from non-opinionated (neutral) tweets in Twitter data streams.

In an effort to build an active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet data streams, we first addressed a static Twitter data analysis problem, which was explored in order to determine the best Twitter-specific text preprocessing setting for training the Support Vector Machine (SVM) sentiment classifier. In the static setting, the Granger causality test showed that sentiment in stock-related tweets can be used as an indicator of stock price movements a few days in advance, where improved results were achieved by adapting the SVM classifier to categorize Twitter posts into three sentiment categories of positive, negative and neutral (instead of positive and negative only). These findings were successfully used in the development of a new stream-based active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet data streams.

Using stream data for sentiment analysis makes sense when the information about the changes in the sentiment is time-critical and a proper data flow is available, for example, in the analysis of streams of financial tweets in which people express their opinions about stocks in real time. The main idea of active learning [56,61,65], adapted in this study for continuously updating the sentiment classifier from a tweet stream, is that the algorithm is allowed to select new examples to be labeled by the oracle (e.g., a human annotator) and added to the training set. It aims at maximizing the performance of the algorithm with as little human labeling effort as possible. The main challenge of active learning is the selection of the most suitable

[1] www.twitter.com.

examples for labeling in order to achieve the highest prediction accuracy, while knowing that one cannot afford to label all the examples [86]. For example, query algorithms based on uncertainty sampling select for labeling the examples for which the current learner has the highest uncertainty [35,62,73]. Similarly, algorithms based on query-by-committee use disagreement among an ensemble of learners to select new examples for labeling [19,38,66]. The active learning approach proposed in this paper combines uncertainty and random sampling and was developed by adapting the initial static sentiment analysis approach to deal with changes over time in a tweet stream. On the one hand, the use of active learning is a consequence of the scarcity of labeled tweets available for sentiment analysis, which prevents the use of conventional machine learning methods. It is namely very difficult and costly to obtain large hand-labeled datasets of tweets, especially if they are domain dependent. On the other hand, these datasets and the resulting models change with time and, consequently, soon become outdated. Thus, continuous learning that allows for adaptations to change in the modeled environment is inevitable to keep the models current.

In summary, the main contribution of this paper is a new methodology for stream-based active learning for tweet sentiment analysis in finance, which can be used on continuously changing tweet streams. A series of experiments was conducted to determine the best querying strategy for active learning of the SVM classifier, which was adapted to sentiment analysis of streams of financial tweets and applied to predictive stream mining in a financial stock market application. As a side effect, since there is no large labeled dataset of financial tweets publicly available, we have labeled and made publicly available a collection of financial tweets, making it the first large (in the sense of labeling effort) publicly available dataset of its kind. We used the dataset in the simulated active learning setting and in the evaluation of the results of tweet stream analysis.

The paper is structured as follows. Section 2 presents a brief overview of related work. Section 3 discusses Twitter-specific text preprocessing options, and presents the developed SVM tweet sentiment classifier, learned from adequately preprocessed Twitter data. Section 4 presents the dataset of financial tweets, which were collected for the purpose of the study, as well as the method and technology developed for enabling financial market predictions from Twitter data. The approach uses *positive sentiment probability* as a new indicator for predictive sentiment analysis in finance, proposed in our previous work [69]. Furthermore, due to the fact that financial tweets do not necessarily express the sentiment, this section applies sentiment classification using the *neutral zone*, which allows classification of a tweet into the neutral category, thus improving the predictive power of the sentiment classifier compared to the SVM classifier categorizing Twitter posts into positive and negative sentiment categories only. Section 5 introduces incremental learning of the classifier on a stream of financial tweets. The general purpose classifier was incrementally updated in order to adapt to the changes in the data stream by using the active learning approach. The paper concludes with a summary of results and plans for further work in Section 6.

## 2. Related work

In this section, we give an overview of related studies, which are focused on: (i) analyzing sentiment in Twitter data, (ii) sentiment analysis of social media as a predictor of the future stock market indicators, and (iii) active learning on data streams. Although these tasks have been well-studied separately, there is a lack of work which would combine them and propose a dynamic adaptive sentiment analysis methodology for microblogging stream posts, which would be able to handle changes in data streams—our work addresses this issue.

### 2.1. Sentiment analysis and microblogging channels

In recent years, several studies have analyzed sentiments expressed in Twitter data in order to describe its content and study its relation to trends. O'Connor et al. [45] analyzed several surveys on consumer confidence and political opinion, and found a correlation with sentiments in Twitter messages. Furthermore, Thelwall et al. [71] analyzed 30 top events in Twitter over a one-month period and showed that popular events are associated with an increase in average negative sentiment strength. In [28], the authors addressed target-dependent sentiment classification and applied it to English tweets on popular topics. They incorporated target-dependent features and also took related tweets into consideration. Asur et al. [3] constructed a model based on tweet-rate about particular topics for predicting box-office revenues of movies before their release. They further showed how sentiment extracted from Twitter posts can improve their forecasting power. In the context of the 2009 German federal elections, Tumasjan et al. [74] showed that sentiment expressed in Twitter messages closely corresponds to the offline political landscape.

There has also been research exploring whether sentiment analysis of social media can be used to predict future stock market indicators. In [59], the authors analyzed sentiment in messages from the *Yahoo! Finance* website[2] and demonstrated that sentiment and stock values are closely correlated. They also showed that one can use sentiment analysis to make predictions about stock behavior over a short-term period. In [46], the authors analyzed sentiments in postings from stock microblogging channel, Stocktwits.com,[3] over a period of three months and found that stock microblog sentiments may predict future stock price movements. Additionally, they found that pessimistic information has higher predictive value as compared

[2] http://finance.yahoo.com.

[3] http://www.stocktwits.com.

to optimistic information. Zhang et al. [85] measured positive and negative emotions in tweets and analyzed the correlation between these measures and stock market indices such as Dow Jones, S&P 500, NASDAQ, and VIX. The authors indicated that by inspecting Twitter for any kind of emotional outburst gives a predictor of how the stock market will perform the following day. Bollen et al. [7] measured mood in tweets in terms of six dimensions (calm, alert, sure, vital, kind, and happy) and showed that changes in calmness can predict daily up and down changes in the closing values of the Dow Jones Industrial Average Index (DJIA). Furthermore, Chen and Lazer [11] confirmed the results of Bollen et al. [7] and showed that even with much simpler sentiment analysis methods, a correlation between Twitter sentiment data and stock market movements can be observed. Mittal and Goel [39] based their work for finding a correlation between public sentiment and the stock market on the approach of Bollen et al. [7]. Their results [39] are in some agreement with the results of Bollen et al. [7], but they indicate that not only the calm, but also the happy mood dimension has a good correlation with the DJIA values. The authors in [42] calculated daily sentiment of aggregated data from multiple sources (Twitter, 11 online message boards, and *Yahoo! Finance* news stream), where the data was concerned with stocks of the S&P 500 index during a six-month period. In their experiments, they showed that publicly available data in microblogs, forums, and news have predictive power for stock price changes on the following day. Sprenger et al. [70] analyzed about 250,000 stock-related tweets and found that the sentiments in tweets is associated with exceptional stock returns and that message volume predicts next-day trading volume. In addition, the authors showed that users that give above-average investment advice are retweeted more often and have more followers, which shows their influence in microblogging forums. Finally, Yu et al. [83] studied the effect of social and conventional media on firm stock market performance and found that social media has a stronger impact. Nevertheless, the authors found that social and conventional media together do have an effect on the stock market. They also found that the effect of social media varies depending on its type.

The above literature overview confirms that sentiment analysis of social media contains predictive information about future stock market indicators, which is also the topic of this paper. Close to our research is the work of Sprenger et al. [70], which aims at finding associations among various values describing tweets and stocks. Also, a similar idea exists in [42], but the authors were interested in aggregating data from multiple sources, whereas we are specifically interested in adjusting our approach to microblogging data. In our previous studies, we used the volume and sentiments in stock-related tweets to identify important events, as a step towards the prediction of future movements of stock prices [68,69]. This paper substantially extends our previous work.

### 2.2. Stream-based active learning

Active learning has been studied in three different scenarios: (i) membership query synthesis, (ii) pool-based sampling, and (iii) stream-based selective sampling [63]. In the membership query synthesis scenario, the learner may select new examples for labeling from the input space or it can generate new examples itself. In the pool-based scenario, the learner may request labels for any example from a large pool of historical data. Finally, in the stream-based active learning scenario, examples are made available constantly from a data stream and the learner has to decide in real time whether to request a label for a new example or not.

Active learning on data streams has been a subject in many studies. One of the simplest ways to select the examples to be labeled is based on maximizing the expected informativeness of labeled examples. For example, the learner may find the examples with the highest uncertainty to be the most informative and request them to be labeled. Zhu et al. [86] used uncertainty sampling to label instances within a batch of data from the data stream. Žliobaite et al. [88] proposed strategies that extend the fixed uncertainty strategy with dynamic allocation of labeling efforts over time and randomization of the search space. The latter approach was used also in some of our active learning strategies described in Section 5. These newly proposed active learning strategies explicitly handle concept drift and adapt the classifier to data distribution changes in data streams over time. In contrast to our approach, Žliobaite et al. [88] do not consider batches, but perform labeling decisions on every encountered data instance. Also, their labeling budget management is different, as they have a fixed overall budget and dynamically adapt the active learning rate according to the amount of budget left. This can be beneficial for flexible adaptation, but can disperse the labeling effort very unevenly. We opted for a fixed budget per batch, which enables the labeling effort to remain the same in each time period. This was perceived as a favorable approach from the user's point of view, as in our case the labeling cost is measured in human time, which is difficult to provide in unevenly dispersed bursts. Deciding which instances are the most suitable for labeling can be made by a single evolving classifier [88] or by a classifier ensemble [79,86,87]. In classifier-ensemble-based active learning frameworks, a number of classifiers are trained from small portions of stream data. These classifiers construct an ensemble classifier for predictions [84], while our work is concerned with the development of a single evolving sentiment classifier for Twitter posts.

Active learning on stream data for sentiment analysis of tweets in financial domains is still insufficiently explored and represents a significant challenge. Our preliminary work on this topic was presented in [54]. Bifet and Frank [6] discuss the challenges posed by Twitter data streams, focusing on classification problems, and consider these streams for sentiment analysis, but they do not use the active learning approach. On the other hand, Settles [64] has developed an active learning annotation tool, DUALIST; while he showed its potential by applying it to sentiment analysis of general tweets, his tool is not specifically adjusted to tweet analysis.

**Table 1**
List of emoticons used for labeling the training set.

| Positive emoticons | Negative emoticons |
|---|---|
| :) | :( |
| :-) | :-( |
| : ) | : ( |
| :D | |
| =) | |

## 3. Defining the best parameter setting for tweet preprocessing

Preprocessing is a necessary data preparation step to supervised machine learning when training a sentiment classifier. We describe here the algorithm used in the development of the initial general tweet sentiment classifier, the dataset, different data preprocessing settings, and the experiments that led to the choice of the best tweet preprocessing setting.

In this work, *classification* refers to the process of categorizing a new tweet into one of the two categories or classes: the positive or the negative sentiment of a tweet. The classifier is trained to classify new instances based on a set of class-labeled training instances (tweets), each described by a vector of features (terms, formed from one or several consecutive words) which have been pre-categorized manually or in some other presumably reliable way. Features are all the terms detected in the training dataset. The length of the feature vectors corresponds to the number of features. The approach to tweet preprocessing and classifier training was implemented using the LATINO[4] software library of text processing and data mining algorithms.

### *3.1. The algorithm used for sentiment classification*

There are three common approaches to sentiment classification [48,71]: (i) machine learning, (ii) lexicon-based methods, and (iii) linguistic analysis. Linguistic analysis tends to be computationally demanding for use in a streaming near-real-time setting. Lexicon-based methods are faster, but are unable to adapt to changes in the modeled environment. In the analysis of dynamic concepts, such as public sentiment, this is a serious drawback. Namely, certain terms, such as company names, countries or phrases, can shift with time from one sentiment class to the other. Therefore, we have decided to use a machine learning approach to learn a sentiment classifier from a set of class-labeled examples.

An algorithm, standardly used in document classification, is the linear Support Vector Machine (SVM) [77,78,12]. The SVM algorithm has several advantages, which are important for learning a sentiment classifier from a large Twitter dataset. For example, it is fairly robust to overfitting and it can handle large feature spaces [10,29,58]. Based on a set of training examples, labeled as belonging to one of the two classes, an SVM algorithm represents the examples as points in the space and separates them by a hyperplane. The aim of the SVM is to place this hyperplane in such a way that examples of the two classes are divided by a gap that is as wide as possible. New examples are then mapped into that same space and classified based on the side of the hyperplane in which they reside. For training the tweet sentiment classifier, we used the SVM$^{perf}$ [30–32] implementation of the SVM algorithm.

### *3.2. The data used for initial classifier training*

Since there is no publicly available large hand-labeled data set for sentiment analysis of Twitter data, we have trained the general purpose tweet sentiment classifier on an available large collection of 1,600,000 (800,000 positive and 800,000 negative) tweets collected and prepared by Stanford University [21], where the tweets were labeled based on a presence of positive and negative emoticons. Therefore, the emoticons approximate the actual positive and negative sentiment labels. This approach was proposed by Read [50]. For example, if a tweet contains the ":)" emoticon, it is labeled as positive, and if it contains the ":(" emoticon, it is labeled as negative. Tweets containing both positive and negative emoticons were not taken into account. The full list of emoticons used for labeling can be found in Table 1. Inevitably, this simple approach causes partially correct or noisy labeling. However, in Appendix A, we illustrate that smiley-labeled tweets are still a reasonable approximation for manually-annotated positive/negative sentiments of tweets. In the dataset, the emoticons were already removed from the tweets in order for the classifier to learn from the other features that characterize them. Note that the tweets from this set do not focus on any particular domain.

### *3.3. Data preprocessing*

The data preprocessing step is important in sentiment analysis and with appropriate selection of preprocessing techniques, the classification accuracy can be improved [24]. We apply both Twitter-specific and standard preprocessing on the data set. The Twitter-specific preprocessing is necessary, since the Twitter community has created its own specific

[4] LATINO (Link Analysis and Text Mining Toolbox) is open-source (mostly under the LGPL license) and is available at http://latino.sourceforge.net/.

language to post messages. Therefore, we first explored the unique properties of this language and experimented with the following options [2,21] for Twitter-specific preprocessing to better define the feature space:

- Usernames: mentioning other users in a tweet in the form @TwitterUser was replaced by a single token named *USERNAME*.
- Usage of web links: Web links pointing to different web pages were replaced by a single token named *URL*.
- Letter repetition: repetitive letters with more than two occurrences in a word were replaced by a word with one occurrence of this letter (e.g., word *loooooooove* was replaced by *love*).
- Negations: we replaced negation words (*not, isn't, aren't, wasn't, weren't, hasn't, haven't, hadn't, doesn't, don't, didn't*) with a unique token *NEGATION*. Using this approach, we do not lose information about a negation, but treat all negation expressions in the same way.
- Exclamation and question marks: exclamation marks were replaced by a single token *EXCLAMATION* and question marks by a single token *QUESTION*.

In addition to Twitter-specific text preprocessing, other standard preprocessing steps were performed [18] to define the feature space for tweet feature vector construction. These include text tokenization (text splitting into individual words/terms), removal of stopwords (words carrying no relevant information, e.g., and, or, a, an, the, etc.), stemming (converting words into their root form), and *N*-gram construction (concatenating 1–*N* stemmed words appearing consecutively in the text, where $N = 2$) for feature space reduction. We also added the condition that a given term has to appear at least twice in the entire corpus, either twice in a given tweet or in two different tweets. The resulting terms were used as features in the construction of feature vectors representing the documents (tweets). In our experiments, we did not use a part of speech (POS) tagger, since it was indicated by Go et al. [21] and Pang et al. [47] that POS tags are not useful when using SVMs for sentiment analysis. Moreover, Kouloumpis et al. [34] showed that POS features may not be useful for sentiment analysis in the microblogging domain.

The standard approach to feature vector construction is TF-IDF-based, where TF-IDF stands for the "term frequency-inverse document frequency" feature weighting scheme [29,82]. TF is the term frequency feature weighting scheme, where a weight reflects how often a word is found in a document, while TF-IDF is the term frequency-inverse document frequency feature weighting scheme, where a weight reflects how important a word is to a document in a document collection (TF-IDF increases proportionally to the number of times a word appears in the document, but decreases with respect to the number of documents in which the word occurs). We experimented with both schemes, TF-IDF- and TF-based, where for every document (tweet) TF weights were normalized to a range of [0,1]. As shown in Table 2, the TF-based approach proved to outperform the TF-IDF-based approach to tweet preprocessing, which is expected in a classification setting [37]. The significance of the finding is confirmed using the Wilcoxon's significance test [15,81], which concluded that using TF is statistically significantly better than TF-IDF (with $p < 8.0 \times 10^{-7}$).

### 3.4. Selecting the best preprocessing setting for classifier training

The experiments with different Twitter-specific preprocessing settings were performed to determine the best preprocessing options which were used in addition to the standard text preprocessing steps. The best preprocessing setting for a classifier[5] was chosen according to the *F*-measure (also known as *F*-score or *F*1 score) [76], determined using the ten-fold cross-validation method.[6] The *F*-measure was used for comparison of different preprocessing settings since later, in the active learning experiments, to compare different querying strategies, we calculate the *F*-measure of positive tweets as there is high three-class imbalance in batches from the data stream. In order to be consistent and allow the reader to compare results in our paper, we used the *F*-measure in all our experiments.

The experiments show that the best preprocessing setting is Setting 1 shown in the first row of Table 2. It is TF-based, uses maximum *N*-grams of size 2, words which appear at least two times in the corpus, it replaces links with the *URL* token, and replaces negations with the *NEGATION* token. This tweet preprocessing setting resulted in the construction of 1,288,681 features used for classifier training. Using the unpaired one-tailed homoscedastic *t*-test [52], we investigated whether the best preprocessing setting (Setting 1) is statistically significantly better than the other preprocessing settings. The results show that the best preprocessing setting is not significantly better than Settings 2–12,14, and 16, but it is significantly better than the remaining preprocessing settings (Setting 13, Setting 15, Settings 17–32) with a *p*-value lower than 0.05.

Since in these experiments the original dataset was pre-filtered and did not contain tweets with both positive and negative emoticons, the reported results may be somewhat overoptimistic (i.e., if the data were not pre-filtered and contained also tweets with mixed emoticons, the results in terms of the *F*-measure would probably be somewhat lower). Nevertheless, even if the reported results are overoptimistic, this property of the dataset does not affect the general conclusions concerning the choice of preprocessing settings, given that in all the settings the dataset was preprocessed in the same way.

[5] Based on our previous experience in [69], the parameters for the SVM$^{perf}$ learner were set to "-c 160000 -e 10".

[6] The *F*-measure is a harmonic mean of precision and recall, and it reaches its best value at 1 and worst at 0. It is calculated as: $F_1 = 2 \times precision \times recall/(precision + recall)$. Precision is the fraction of all examples classified as positive which are correctly classified as positive, while recall is the fraction of all the positive examples that are correctly classified as positive.

**Table 2**
Classifier performance evaluation for various preprocessing settings.

| ID | Usernames to a token | URLs to a token | Remove letter repetition | Negations to a token | Exclamation and question marks to tokens | Avg. *F*-measure ten-fold cross-val. ± std. dev. (TF) | Avg. *F*-measure ten-fold cross-val. ± std. dev. (TF-IDF) |
|---|---|---|---|---|---|---|---|
| 1 | | X | | X | | 0.7937 ± 0.0059 | 0.7763 ± 0.0067 |
| 2 | X | | | X | | 0.7936 ± 0.0041 | 0.7779 ± 0.0061 |
| 3 | | X | X | X | | 0.7933 ± 0.0056 | 0.7756 ± 0.0051 |
| 4 | | | | X | | 0.7923 ± 0.0062 | 0.7801 ± 0.0055 |
| 5 | X | X | | X | | 0.7922 ± 0.0062 | 0.7786 ± 0.0053 |
| 6 | | | X | X | X | 0.7912 ± 0.0065 | 0.7774 ± 0.0033 |
| 7 | | X | | X | X | 0.7910 ± 0.0057 | 0.7755 ± 0.0047 |
| 8 | X | | X | X | X | 0.7907 ± 0.0064 | 0.7756 ± 0.0063 |
| 9 | X | X | X | X | X | 0.7906 ± 0.0055 | 0.7776 ± 0.0042 |
| 10 | X | | | X | X | 0.7904 ± 0.0052 | 0.7741 ± 0.0050 |
| 11 | X | | X | X | | 0.7903 ± 0.0073 | 0.7764 ± 0.0075 |
| 12 | X | X | X | X | | 0.7897 ± 0.0071 | 0.7759 ± 0.0069 |
| 13 | | | X | X | | 0.7897 ± 0.0042 | 0.7763 ± 0.0071 |
| 14 | | | | X | X | 0.7894 ± 0.0058 | 0.7716 ± 0.0055 |
| 15 | | | X | | X | 0.7894 ± 0.0050 | 0.7663 ± 0.0067 |
| 16 | | X | X | X | X | 0.7891 ± 0.0061 | 0.7752 ± 0.0051 |
| 17 | X | X | | | | 0.7885 ± 0.0064 | 0.7694 ± 0.0061 |
| 18 | X | X | X | | | 0.7882 ± 0.0063 | 0.7730 ± 0.0055 |
| 19 | X | X | | X | X | 0.7881 ± 0.0062 | 0.7740 ± 0.0059 |
| 20 | X | | X | | | 0.7878 ± 0.0055 | 0.7739 ± 0.0059 |
| 21 | | X | X | | | 0.7876 ± 0.0061 | 0.7697 ± 0.0072 |
| 22 | X | X | X | | X | 0.7874 ± 0.0046 | 0.7670 ± 0.0081 |
| 23 | | X | X | | X | 0.7862 ± 0.0062 | 0.7716 ± 0.0062 |
| 24 | | X | | | X | 0.7862 ± 0.0036 | 0.7675 ± 0.0069 |
| 25 | | X | | | | 0.7853 ± 0.0071 | 0.7681 ± 0.0050 |
| 26 | X | | | | X | 0.7850 ± 0.0054 | 0.7695 ± 0.0063 |
| 27 | X | | X | | X | 0.7840 ± 0.0063 | 0.7711 ± 0.0065 |
| 28 | X | | | | | 0.7839 ± 0.0057 | 0.7690 ± 0.0077 |
| 29 | | | | | X | 0.7836 ± 0.0067 | 0.7702 ± 0.0089 |
| 30 | | | X | | | 0.7834 ± 0.0054 | 0.7745 ± 0.0061 |
| 31 | X | X | | | X | 0.7830 ± 0.0082 | 0.7657 ± 0.0049 |
| 32 | | | | | | 0.7829 ± 0.0077 | 0.7712 ± 0.0077 |

From Table 2, it follows that replacing negation words is particularly beneficial since almost all settings which perform this replacement are placed in the upper part of the table. Replacing exclamation and question marks with a token does not seem to be helpful, since the five top settings do not employ this replacement. Regarding replacing usernames and URLs with a token, and removing letter repetition, one cannot draw general conclusion, since these preprocessing options are dispersed across the table. Nevertheless, the first setting employs replacing URLs with a token and we used it in the rest of our experiments. Interestingly, the setting which does not apply any of the preprocessing adjustments achieved the lowest performance, leading to the conclusion that, in general, it is beneficial to preprocess Twitter data.

In addition to the SVM algorithm, we also tested the *k*-nearest neighbor (KNN) and Naive Bayes classifiers on the same dataset. In this setting, the standard KNN algorithm proved to be too slow (in ten-fold cross-validation experiments for $K = 5$ and $K = 10$, the onefold experiment took more than 24 hours on a standard desktop computer), and Naive Bayes had lower performance compared to the SVM (the ten-fold cross-validation achieved an *F*-measure of 0.73). We, thus, used the SVM classifier with preprocessing Setting 1 from Table 2 for the rest of the study and analyses.

## 4. Stock market analysis in a static predictive tweet analysis setting

Motivated by the earlier research and observation that the stock market itself can be considered as a measure of social mood [44], this section investigates whether sentiment analysis on Twitter posts can provide predictive information about the value of stock closing prices. We use a supervised machine learning approach to train a sentiment classifier, using a SVM algorithm. By applying the best setting for tweet preprocessing, as explained in Section 3.4, two sets of experiments were performed. In the first set of experiments, tweets were classified into two categories, positive or negative. In the second set of experiments, the SVM classification approach was advanced by taking into account the neutral zone, enabling us to identify neutral tweets (not clearly expressing positive or negative sentiments) as those, positioned a small distance from the SVM model hyperplane.

### *4.1. The data used in the stock market application*

A tweet dataset and stock closing prices of several companies were collected for our experiments. On the one hand, we collected 152,570 tweets discussing relevant stock information concerning eight companies (Apple, Amazon, Baidu, Cisco, Google, Microsoft, Netflix, and RIM)[7] in the nine-month time period from March 11 to December 9, 2011. On the other hand, we collected stock closing prices of these companies for the same time period.

The data source for collecting financial Twitter posts is the Twitter API[8] (i.e., the Twitter Search API), which returns tweets that match a specified query. By informal Twitter conventions, the dollar-sign notation is used for discussing stock symbols. For example, the $BIDU tag indicates that the user discusses Baidu stocks. This convention was used for the retrieval of financial tweets.[9] The stock closing prices of the companies for each day were obtained from the *Yahoo! Finance* website.

The time of tweets in our dataset is presented in UTC (Coordinated Universal Time) since the Twitter API stores and returns dates and times in UTC. On the other hand, Baidu is included in the NASDAQ-100 index, and this stock exchange works in the EST (Eastern Standard Time)/EDT (Eastern Daylight Time) timezone which is four to five hours behind UTC. Therefore, compared to EST/EDT, there is an additional shift of four to five hours; thus, there is more of a time lag between the tweets of a previous day and the stock market activity and closing prices of the current day.

In the entire study, we focused on the analysis of financial tweets on the Chinese web search engine provider, Baidu,[10] in order to investigate relationships between the observed sentiments in the stock-related tweets and the corresponding stock price movements. The collection of Baidu tweets was manually labeled by the domain expert. The data of this Chinese web search engine provider was chosen for hand-labeling since the set of tweets related to Baidu was of a manageable size given the resources available (we collected and labeled approximately 11,000 tweets, compared to, for example, approximately 40,000 tweets that we collected for the Apple company). Even this hand-labeling effort took us over three months to ensure good quality of the labeled data.

In tweet labeling, we were faced with the problem of choosing a labeling strategy. Having discussed this issue at length with stock market financial experts, we opted for manual labeling of the tweets from the point of view of a particular company and not mainly on the sentiment-carrying words used. The reason for this decision is that our long-term intention is to construct classifiers that should distinguish between sentiments of tweets of different companies; hence, a company-focused view is a necessity. The labels were given to instances according to their financial sentiment; that is, their impact on the perception of the company, its products, or its stock. For example, a tweet: “I just love shorting CompanyX. What a nice day of profits, first of many…” would be labeled as negative, since shorting means betting that the value of the stock will drop. Despite many positive sentiment words, such a tweet would be providing a message of a negative financial prospect for CompanyX. Another issue was that in the dataset there are many tweets that do not discuss Baidu stocks, although they do contain the $BIDU tag. These tweets may actually express an opinion about another company, such as the tweet, “Apple is great $BIDU”, and reflect a positive tweet sentiment, but do not discuss the Baidu company at all. Again, these kinds of tweets were labeled from the point of view of the Baidu company, and not mainly on the sentiment-carrying words used. Therefore, the mentioned tweet would be labeled as neutral.

Therefore, in Baidu sentiment labeling, the annotator was instructed to focus on the following question:

- “What would someone who knows what Baidu is and shares in general, think of Baidu and its shares after he sees this tweet?”, or in other words,
- “Is this tweet positive, negative, or neutral concerning Baidu and/or the price of its shares?”

The resulting hand-labeled dataset consists of 11389 Baidu financial tweets (4861 positive, 1856 negative, and 4672 neutral tweets).[11] In this dataset, neutral tweets are those that contain no sentiment about Baidu, contain both positive and negative sentiments about Baidu, as well as those that do not discuss Baidu even if they are positively or negatively oriented (as discussed above).

### 4.2. Correlation between tweet sentiment and stock closing price

Given the time series of tweet sentiments and the time series of stock closing prices, the question addressed is whether one time series is useful in forecasting another. We applied a statistical test to determine whether sentiments expressed in tweets contain predictive information about the future values of stock closing prices. To this end, we performed a Granger causality analysis test, which is a statistical hypothesis test for discovering whether one time series is effective for forecasting another time series [22]. Since we have the tweets time series on the one hand and the stock closing price time series on the other hand, this test suits our needs to check whether there is a predictive relationship between sentiments in tweets and stock closing prices. If time series $X$ is said to Granger-cause $Y$, then the information in past values of $X$ helps predict values of $Y$ better than only the information in past values of $Y$. Therefore, the lagged values of $X$ will have a statistically significant correlation with $Y$.

[7] Tweet IDs of our datasets are available on: http://streammining.ijs.si/TwitterStockSentimentDataset/TwitterStockSentimentDataset.zip.

[8] https://dev.twitter.com/.

[9] To deal with spam (writing nearly identical messages from different accounts), we employed the algorithm based on the work of Broder et al. [8] to discard tweets that were detected as near duplicates.

[10] www.baidu.com.

[11] The Baidu tweet IDs and manual labels are publicly available on: http://streammining.ijs.si/TwitterStockSentimentDataset/LabeledTwitterDataset.zip, file BIDU.txt.

Granger causality analysis is based on linear regression modeling of stochastic processes and it is usually done using a series of $t$-tests and $F$-tests on lagged values of $X$ (combined also with lagged values of $Y$). The test expects that the time series data is covariance stationary and that it can be represented by a linear model. Complex implementations for nonlinear cases exist; nevertheless, they are often more challenging to apply in practice [60].

The output of the Granger causality test is the $p$-value, which takes values in the $[0,1]$ interval. In statistical hypothesis testing, the $p$-value is a measure of how much evidence we have against the null hypothesis [55]. If the $p$-value is lower than the selected significance level, for example 5% ($p < 0.05$), the null hypothesis is rejected and the result is statistically significant. On the other hand, a large $p$-value represents weak evidence against the null hypothesis; thus, the null hypothesis cannot be rejected. The Granger causality test that we used is based on Free Statistics Software [80].

To enable in-depth analysis, we calculated a sentiment indicator for predictive sentiment analysis in finance, named *positive sentiment probability*: $p_{sp}$, which was proposed in our previous work [69]. Positive sentiment probability is computed for a day $d$ of a time series by dividing the number of positive tweets $N_{pos}$ by the number of all tweets on that day $N_d$.

$$p_{sp}(d) = N_{pos}(d)/N_d(d) \tag{1}$$

This ratio is used to estimate the probability that the sentiment of a randomly selected tweet on a given day is positive.

To test whether one time series is useful in forecasting another, using the Granger causality test, we first calculated positive sentiment probability for each day when the stock market was open. We then calculated two ratios[12] to meet the Granger causality test condition that the time series data needs to be stationary:

- Daily change of the positive sentiment probability $D_{sent}$: *positive sentiment probability today−positive sentiment probability yesterday.*

$$D_{sent}(d) = p_{sp}(d) - p_{sp}(d-1) \tag{2}$$

- Daily return in stock closing price $D_{price}$: *(closing price today−closing price yesterday)/closing price yesterday.*[13]

$$D_{price}(d) = \frac{price(d) - price(d-1)}{price(d-1)} \tag{3}$$

We applied the Granger causality test to test the following null hypothesis:

- "sentiment in tweets does not predict stock closing prices" (when rejected, meaning that the sentiment in tweets Granger-causes the values of stock closing prices).

We performed tests on the entire nine-month time period (from March 11 to December 9, 2011), as well as on individual three-month periods (corresponding approximately to March–May, June–August, and September–November). Results for Baidu are shown in the first column of Table 3. In Granger causality testing, we considered lagged values of time series for one, two, and three days.

Since in our experiments we compute the $p$-value repetitively and do multiple comparisons of $p$-values for different experimental settings, we used the Bonferroni correction [1] to neutralize the problem of multiple comparisons. This correction is considered very conservative. It makes adjustments to a critical $p$-value by dividing it by the number of comparisons being made. In our case, we divided the $p$-value of 0.1 by 4, as this is the number of time periods (whole nine months and three three-month periods) which we consider to be a family of tests. We compare the $p$-values which came from the Granger causality test with 0.1/4 = 0.025 and reject the null hypothesis if the value is lower than 0.025. After applying the Bonferroni correction, the results of the Granger analysis indicated that in this particular setting there are no significant results.

### 4.3. Experiments in a three-class setting with the neutral zone

In the previous section, we classified financial tweets into one of the two categories, positive or negative, and therefore assumed that every tweet contains an opinion. This is, however, sometimes an unrealistic assumption, since a tweet can be objective and without any opinion about a given company (i.e., without expressed sentiment). Considering this, a tweet should also have the possibility of being classified as either neutral or weakly opinionated. In this section, we address a three class problem of classifying tweets into the positive, negative, and neutral categories.

Our training data does not contain any neutral tweets for the classifier to learn from. Therefore, we define a tweet, which is projected into an area close to the SVM model's hyperplane, as neutral. We define this area as the neutral zone, which is parameterized by value $t$, where $t$ represents the positive and $-t$ the negative border of the neutral zone. If a tweet $x$ is projected into this zone, that is, $-t < d(x) < t$, then rather than being assigned to one of the two sentiment classes, it is assumed

[12] The ratios were defined in collaboration with the domain experts from the Stuttgart Stock Exchange (see Acknowledgments).

[13] The same transformation of the price time series was used in [53].

**Table 3**
Statistical significance ($p$-values) of Granger causality correlation between positive sentiment probability and closing stock price for Baidu, while changing the size of the neutral zone (i.e., the $t$ value from 0 to 1). Values which are lower than a $p$-value of 0.1, after applying the Bonferroni correction, are marked in bold.

| Size of the neutral zone ($t$ value) | | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time period | Lag | | | | | | | | | | | |
| 9 Months | 1 | 0.066 | 0.373 | 0.417 | 0.228 | 0.318 | 0.504 | 0.213 | 0.217 | 0.549 | 0.585 | 0.905 |
| March–May | 1 | 0.403 | 0.306 | 0.463 | 0.359 | 0.439 | 0.367 | 0.542 | 0.864 | 0.970 | 0.896 | 0.614 |
| June–August | 1 | 0.069 | 0.124 | 0.203 | 0.070 | 0.164 | 0.392 | 0.157 | 0.068 | 0.240 | 0.376 | 0.340 |
| September–November | 1 | 0.061 | 0.399 | 0.420 | 0.377 | 0.372 | 0.243 | 0.193 | 0.673 | 0.838 | 0.792 | 0.969 |
| 9 Months | 2 | 0.047 | 0.301 | 0.337 | 0.262 | 0.249 | 0.383 | 0.299 | 0.388 | 0.540 | 0.518 | 0.830 |
| March–May | 2 | 0.470 | 0.403 | 0.534 | 0.414 | 0.380 | 0.133 | 0.033 | 0.041 | 0.357 | 0.163 | 0.107 |
| June–August | 2 | 0.050 | 0.038 | 0.033 | 0.039 | **0.020** | **0.012** | **0.004** | **0.010** | 0.122 | 0.140 | 0.311 |
| September–November | 2 | 0.069 | 0.591 | 0.641 | 0.572 | 0.593 | 0.445 | 0.383 | 0.864 | 0.759 | 0.639 | 0.742 |
| 9 Months | 3 | 0.041 | 0.295 | 0.386 | 0.299 | 0.323 | 0.424 | 0.359 | 0.367 | 0.543 | 0.515 | 0.830 |
| March–May | 3 | 0.664 | 0.613 | 0.756 | 0.661 | 0.568 | 0.203 | 0.050 | 0.104 | 0.511 | 0.340 | 0.197 |
| June–August | 3 | 0.098 | 0.076 | 0.059 | 0.100 | 0.050 | **0.021** | **0.011** | 0.029 | 0.199 | 0.171 | 0.293 |
| September–November | 3 | 0.028 | 0.277 | 0.341 | 0.264 | 0.343 | 0.437 | 0.471 | 0.790 | 0.877 | 0.805 | 0.898 |

to be neutral. Note that our "neutral zone" does not denote only the "neutral tweets", such as tweets which would be labeled as neutral by a human annotator. Instead, the neutral zone contains also the tweets which are either positive or negative but close to the SVM hyperplane which separates the positives from the negatives. Thus, the neutral zone includes tweets containing mixed sentiments, weakly opinionated positive/negative tweets, as well as tweets containing terms which were not observed during the training phase (if human annotated neutral tweets were available, they would have been included in the neutral zone as well). For a greater $t$ (i.e., greater size of the neutral zone), the classifier is more confident in its classification decision for positive and negative tweets. Our definition of the neutral zone is simple, but allows fast computation.

We repeated our experiments on classifying financial tweets, but now also took into account the neutral zone. Our aim was to investigate whether the introduction of the neutral zone would improve the predictive capabilities of tweets. Therefore, every tweet which mentioned the Baidu company was classified into one of the three categories: positive, negative, or neutral. Then, we applied the same processing of data as before (count the number of positive, negative, and neutral tweets, calculate positive sentiment probability, calculate daily changes of the positive sentiment probability and the daily return of the stocks' closing price) and performed the Granger analysis test. We varied the $t$ value from 0 to 1 (where $t = 0$ corresponds to classification without the neutral zone) and again calculated the $p$-value for the separate day lags (1, 2, and 3). The results are shown in Table 3. The first column, where the size of the neutral zone is 0, represents the classification without the neutral zone, where financial tweets were classified into one of the two categories, positive or negative. All the remaining columns contain $p$-values for various sizes of the neutral zone. In Appendix B, we also report the results of the Granger causality correlation between positive sentiment probability and closing stock price for the rest of the companies (Apple, Amazon, Cisco, Google, Microsoft, Netflix, and RIM), whose tweets we collected. The results show that for several other companies, the learned classifier has the potential to be useful for stock price prediction in terms of Granger causality.

Values which are lower than a $p$-value of 0.1, after applying the Bonferroni correction, are marked in bold in Table 3. The highest number of significant values was obtained with $t$ values of 0.5 and 0.6 for the border distance of the neutral zone from the SVM hyperplane. Therefore, by introducing the neutral zone, we improved the predictive power of our classifier.

From Table 3 it follows that for the June–August time period we achieved the best results and relationships between sentiments in tweets and stock closing prices. Therefore, we investigated in more detail the Baidu data and public web media from this time period to find possible reasons for this. Fig. 1 shows a screenshot from the Google Finance[14] web page displaying stock price and news media coverage for Baidu in 2011. From the figure, it can be observed that most of the key events in 2011 happened in the period from June to August. Note that this period is also characterized by the highest number of press releases[15] for Baidu in 2011. We hypothesize that this resulted in higher media exposure and, consequently, enabled speculations about price movements in social media. However, further studies are required to confirm or reject this claim.

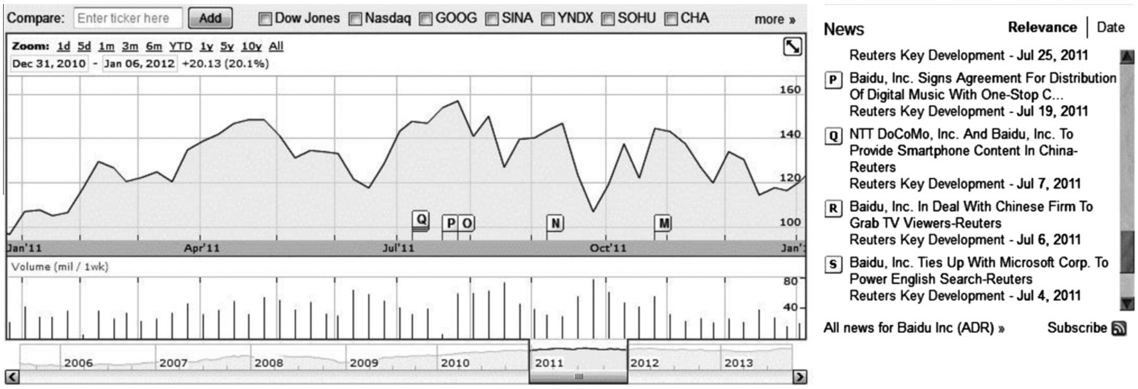
In addition, we explored whether there is evidence for the reversed causality (that the price movements may influence the public sentiment). The results show that, after making adjustments to the critical $p$-value by applying the Bonferroni correction, no significant results were left for the reverse direction.

### *4.4. Summary of the proposed methodology for static predictive tweet analysis*

We proposed a new methodology for determining the correlation between sentiments in tweets that discuss a company's relevant stock information and stock closing prices of the company to determine whether tweet sentiment contains predictive information about the value of the stock closing price. The methodological steps are presented in Fig. 2.

[14] https://www.google.com/finance.

[15] http://phx.corporate-ir.net/phoenix.zhtml?c=188488&p=irol-news&nyo=2.



Reuters Key Development - Jul 25, 2011

P Baidu, Inc. Signs Agreement For Distribution Of Digital Music With One-Stop C...
Reuters Key Development - Jul 19, 2011

Q NTT DoCoMo, Inc. And Baidu, Inc. To Provide Smartphone Content In China-Reuters
Reuters Key Development - Jul 7, 2011

R Baidu, Inc. In Deal With Chinese Firm To Grab TV Viewers-Reuters
Reuters Key Development - Jul 6, 2011

S Baidu, Inc. Ties Up With Microsoft Corp. To Power English Search-Reuters
Reuters Key Development - Jul 4, 2011

All news for Baidu Inc (ADR) » Subscribe

**Fig. 1.** Screenshot from the Google Finance web page showing stock prices and key events. It can be observed that most of the key events in 2011 happened in the period from June to August. We hypothesize that this resulted in a higher media exposure and, consequently, enabled speculations about price movements in social media.

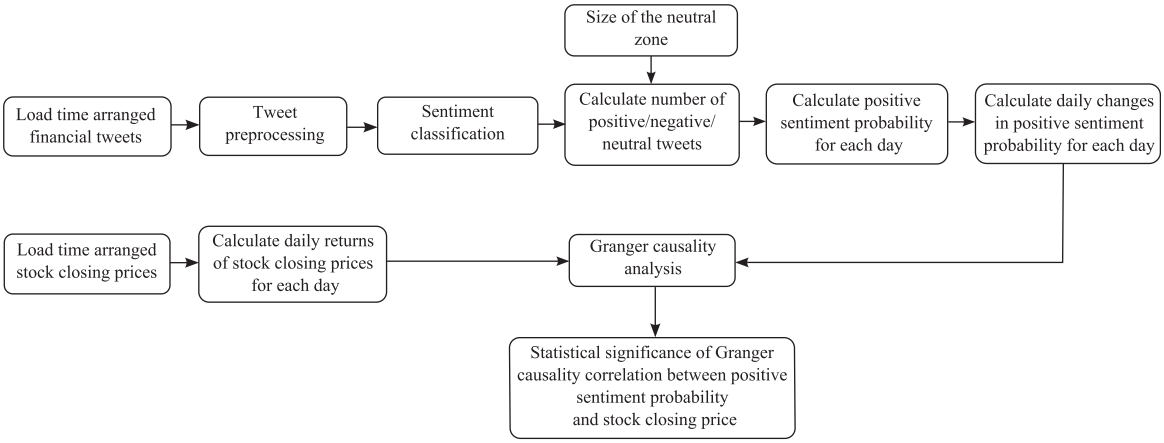

Size of the neutral zone
Load time arranged financial tweets
Tweet preprocessing
Sentiment classification
Calculate number of positive/negative/ neutral tweets
Calculate positive sentiment probability for each day
Calculate daily changes in positive sentiment probability for each day
Load time arranged stock closing prices
Calculate daily returns of stock closing prices for each day
Granger causality analysis
Statistical significance of Granger causality correlation between positive sentiment probability and stock closing price

**Fig. 2.** Methodological steps for predictive sentiment analysis applied to determine the correlation between tweets sentiment and stock closing prices.

As follows from Fig. 2, one should first provide a time series collection of tweets discussing relevant stock information concerning a company of interest. Tweets are then adequately preprocessed (see Section 3). Furthermore, tweets are classified as being positive, negative, or neutral, where tweet labels depend both on the output of the classifier and the size of the neutral zone. Moreover, for every day of a time series, positive sentiment probability is computed by dividing the number of positive tweets per day by the total number of tweets in that day. Last, daily changes of the positive sentiment probability are calculated.

On the other hand, the stock closing prices of the selected company for each day should be collected, which is publicly available information. Daily returns in the stock closing price are then calculated.

Given the daily changes of the positive sentiment probability time series and the daily returns in the stock closing price time series, Granger causality analysis is performed (considering lagged values of time series for one, two, and three days) to test whether tweet sentiment is useful for forecasting the movement of prices on the stock market and how significant the results are.

## 5. Active learning on financial tweet streams for stock market analysis

In the previous section, we classified financial tweets by using a static classifier, which was learned from smiley-labeled general purpose tweets. A significant correlation between the sentiment in financial tweets in the static tweet analysis setting motivated further advances. We focused on three goals: to make the classifier more domain-specific in order to better classify financial tweets; to extend the approach with a capability of continuous updating of the classifier in order to adapt to sentiment vocabulary changes in a data stream; and, in addition to smiley-labeled tweets, to also use hand-labeled financial tweets in the training phase. The crucial element of addressing these challenges was the use of the active learning approach.

In active learning, the learning algorithm interactively queries for manual labels of selected data items. Typically, the active learning algorithm first learns from a small labeled dataset. According to this initial model and the characteristics of the newly observed unlabeled data instances, the algorithm selects a set of new instances to query an expert for their manual labels. This process is repeated until some threshold (e.g., time limit, labeling quota, or target performance) is reached or, as it is the case in stream data, it continues as long as the application is active. This approach largely decreases the number of data instances that need to be manually labeled.

In our experiments, the active learning algorithm first learns from the Stanford smiley-labeled dataset. According to this initial model, the algorithm selects a set of financial tweets from a first batch of data from a data stream to query for their manual labels. Based on these hand-labeled financial tweets, the model is updated and the process is repeated for the next batch of financial tweets. This process is repeated until the end of our simulated data stream is reached. In this way, with time and by updating the model with hand-labeled financial tweets, the sentiment classifier is improved and made more domain specific. Since we use the machine learning approach, sentiment discovery based on tweets may change over time. If we used an approach based on a sentiment lexicon, incremental active learning would not make sense since sentiment-bearing words (senti-words) typically do not change over time (e.g., the word "excellent" surely would not change its sentiment over time). However, taking as a basis the Bag-of-Words document representation, our approach is different, and not based solely on words that explicitly bear sentiment. The classifier takes into account all the terms appearing in tweets including those representing names of people, products, technologies, countries, etc., whose impact on sentiment may change over time and can even completely shift their sentiment polarity (consider terms like "Ireland," whose sentiment has changed in recent history due to the developments of the current financial crisis).

Since it is very difficult and costly to obtain hand-labeled datasets of tweets, especially if they are domain dependent, an active learning approach is highly suitable for our task. The learning algorithm is able to interactively query the expert to obtain the manual labels as new financial tweets come from a data stream. Consequently, the sentiment classifier is more domain specific, it is updated in time in order to detect the changes in sentiment and handle concept drift, and it is improved using highly reliable hand-labeled tweets.

### *5.1. Experimental setting*

To address the challenging task of stream-based sentiment analysis, we employed and tested a selection of active learning strategies and settings. The proposed learning algorithm interactively queries the user to obtain the labels of the tweets which are the closest to the boundaries of the neutral zone. With this approach, with time, the classifier learns how to better distinguish between the neutral and the opinionated (positive/negative) tweets. Note that when the size of the neutral zone is 0, the querying algorithm represents the standard uncertainty sampling approach. We test this hypothesis against the random strategy, and also combine these two strategies. In all the experiments, we varied the size of the neutral zone in order to find the best one.

In our implementation of the active learning approach, we used the Pegasos SVM [67] learning algorithm from the sofia-ml (Suite of Fast Incremental Algorithms for Machine Learning) library [57]. We adjusted this learning algorithm to our active learning experiments. To construct the initial sentiment model for active learning experiments, we used 1,600,000 Stanford general purpose smiley tweets [21].

For the evaluation of the algorithms, we used the holdout evaluation approach [5], where classifier evaluation is performed using a separate unseen holdout set of test examples (in our case: tweets) [25,26]. In dynamic environments, where new examples come from a data stream and concept drift is assumed, an algorithm can collect a batch of examples from the data stream and use them to evaluate the model. The examples in the batch have not yet been used for training. The evaluation is repeated periodically for new batches of examples which come from the stream. After testing of a batch is complete, the algorithm selects the most suitable examples from the batch and asks an oracle to label them. The labeled examples are then used for additional training of the algorithm.

We evaluated our classification model in two different settings: after every 50 and 100 tweets which come from the data stream and represent a batch. After the testing of a batch is completed, the algorithm selects 10 tweets for hand-labeling. Then, only positive and negative labeled tweets are used for additional training, while neutral ones are discarded. We implemented the following strategies for the selection of tweets for labeling as part of the active learning process:

- Active closest to the neutral zone: The algorithm selects 10 tweets that are closest to the boundaries of the neutral zone to be labeled by the human annotator. Out of the selected 10 tweets, at most, five are positive and five are negative, based on positive/negative labeling by the classifier.
- Active combination: This strategy combines the other two strategies in order to better explore the SVM space. We experimented with two combinations: 80% "Closest to the neutral zone" and 20% random strategy (Active combination 20% random) and 50% "Closest to the neutral zone" and 50% random strategy (Active combination 50% random). In the combination strategies, the maximum number of positive/negative tweets selected for hand-labeling from a batch with "Closest to the neutral zone" is five.
- Active random 100%: The algorithm randomly selects 10 tweets from a batch of tweets from the data stream.

To compare different querying strategies, for every batch in the stream, we calculated the *F*-measure of positive tweets. The reason for using the *F*-measure is the unbalanced class distribution in batches.

### 5.2. Determining the best active learning setting

To identify the best active learning setting in terms of the active learning strategy, the batch size, and the size of the neutral zone, we calculated the average *F*-measure for every strategy at different sizes of the neutral zone (see Table 4). The three active learning strategies introduced in Section 5.1 were compared together with the algorithm which did not use the active learning approach (i.e., it did not update the sentiment classifier with time).

In terms of exploring the best size of the neutral zone, we experimented with all the active learning strategies. The results in Table 4 indicate that, in general, the active learning approach improves the performance of the classifier compared to the strategy, which does not use active learning. The table allows for some other observations as well. For example, in all the strategies, the results show that in terms of the *F*-measure, it is better to have no neutral zone or a very small one (0.0001). Moreover, one can observe that the random component of the querying strategies has mixed effects.

To test the significance of the differences between the multiple settings, we followed the procedure recommended in [15]. Namely, we first used the Friedman test [20] with the Iman-Davenport improvement [27] to check whether the difference in performance is statistically significant, and then the Nemenyi post hoc test [43] to search where the significant differences appear.

The Friedman test ranks the algorithms for each dataset separately, where the best performing algorithm gets the rank of 1, the second best rank 2, etc. In the case of ties (we used the *F*-measure computed to a precision of three decimal points), average ranks are assigned. The Friedman test then compares the average ranks of the algorithms. The null hypothesis states that all the performance of the algorithms is equal and, thus, their ranks should be equal. If the null hypothesis is rejected, we can proceed with a post hoc test. The Nemenyi test [43] needs to be used since all the settings must be compared to each other. The Nemenyi test computes the critical distance between the different strategies, and concludes that the differences between the *F*-measures are statistically significant if the corresponding average ranks of the corresponding algorithms differ by at least the critical distance.

The results of the significance post hoc tests are graphically represented using critical diagrams. Fig. 3 shows the results of the analysis of the *F*-measures from Table 4. On the axis of each diagram, we plot the average rank of the settings. The lowest (best) ranks are to the right. We show the critical distance on the top, and connect the settings that are not significantly different. From the results we can draw several conclusions. "Select 10 of 100" batch selection is better than "Select 10 of 50" batch selection, but not significantly better. Settings without the active learning approach showed poor performance compared to the settings with the active learning approach. Overall, the best setting for active learning is to choose 10 tweets in each batch of 100 tweets and use the querying strategy "Closest to the neutral zone". This setting is significantly better than "Select 10 of 50" batch selection without active learning.

Next, we applied the Friedman test with the Iman-Davenport improvement [27] and its corresponding post hoc Nemenyi test [43] on individual batch selection strategies. In Fig. 4, the results of the test on the case "Select 10 of 50" batch selection can be seen. Similarly, Fig. 5 shows results of the "Select 10 of 100" batch selection. From both figures it follows that strategies with the active learning approach are significantly better than the strategy without the active learning approach.

Additionally, we performed another experiment where incremental active learning was performed from Baidu data only; that is, the active learning algorithm first learns from the 100 positive and 100 negative tweets chosen from the first 1000 hand-labeled financial tweets from the Baidu dataset. According to this initial model, the algorithm selects a set of financial tweets from a first batch of data from the Baidu tweet data stream to query for their labels. Based on these hand-labeled financial tweets, the model is updated and the process is repeated for the next batch of Baidu tweets. This process is repeated until we reach the end of our simulated data stream. For this experiment, we used the "Combination 50% random" active learning approach with the size of the neutral zone 0.001 and "Select 10 of 100" batch selections. The results indicate that the classifier learned on such a small initial dataset, although hand-labeled and specific for the financial domain, is highly unstable. The sentiment classifier learned on this dataset classified all tweets at the beginning of the data stream as negative.

**Table 4**
Values of average *F*-measure ± std. deviation for different strategies, while changing the size of the neutral zone (i.e., the *t* value).

| Size of the neutral zone (*t* value) | 0.0000 | 0.0001 | 0.001 | 0.01 |
|---|---|---|---|---|
| *Select 10 of 100* | | | | |
| Active closest to the neutral zone | 0.5410 ± 0.1106 | 0.5403 ± 0.1104 | 0.5256 ± 0.1076 | 0.4314 ± 0.1016 |
| Active combination 20% random | 0.5413 ± 0.1108 | 0.5402 ± 0.1106 | 0.5249 ± 0.1076 | 0.4312 ± 0.1015 |
| Active combination 50% random | 0.5412 ± 0.1109 | 0.5396 ± 0.1104 | 0.5243 ± 0.1074 | 0.4309 ± 0.1013 |
| Active random 100% | 0.5411 ± 0.1109 | 0.5392 ± 0.1096 | 0.5246 ± 0.1079 | 0.4317 ± 0.1013 |
| No active learning | 0.5351 ± 0.1102 | 0.5334 ± 0.1098 | 0.5200 ± 0.1089 | 0.4284 ± 0.1011 |
| *Select 10 of 50* | | | | |
| Active closest to the neutral zone | 0.5394 ± 0.1325 | 0.5380 ± 0.1330 | 0.5276 ± 0.1310 | 0.4310 ± 0.1299 |
| Active combination 20% random | 0.5390 ± 0.1327 | 0.5380 ± 0.1330 | 0.5273 ± 0.1314 | 0.4310 ± 0.1299 |
| Active combination 50% random | 0.5387 ± 0.1325 | 0.5369 ± 0.1328 | 0.5258 ± 0.1309 | 0.4313 ± 0.1294 |
| Active random 100% | 0.5372 ± 0.1327 | 0.5368 ± 0.1326 | 0.5248 ± 0.1310 | 0.4304 ± 0.1297 |
| No active learning | 0.5298 ± 0.1331 | 0.5280 ± 0.1329 | 0.5141 ± 0.1345 | 0.4232 ± 0.1302 |

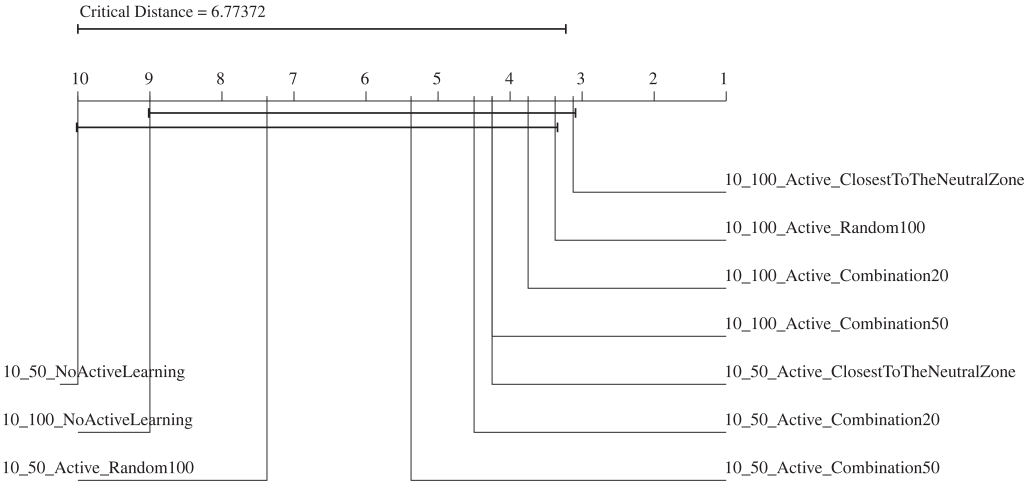


**Fig. 3.** Visualisation of Nemenyi post hoc tests for the active learning strategies on data from Table 4.

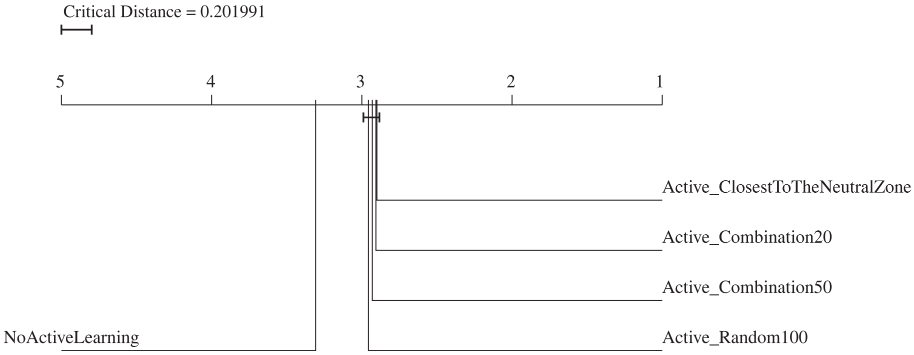


**Fig. 4.** Visualization of Nemenyi post hoc tests for the "Select 10 of 50" batch selection.

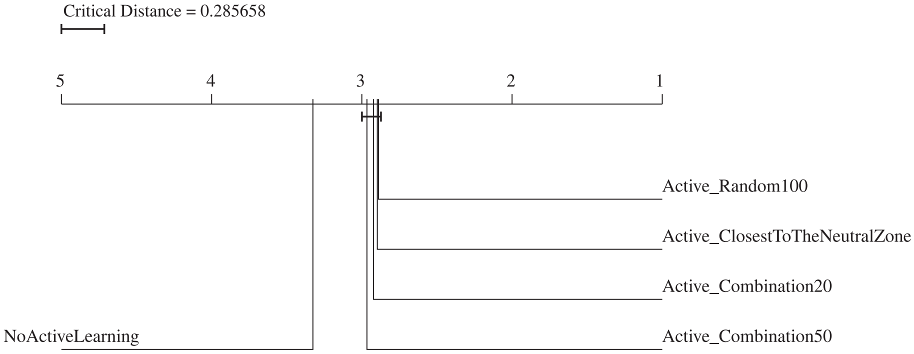


**Fig. 5.** Visualization of Nemenyi post hoc tests for the "Select 10 of 100" batch selection.

Then, as a consequence of active learning and improving the classifier with new labeled tweets, the classifier improved and started to classify new tweets as positive or negative. This improvement lasted for several batches, and then the classifier classified all new coming tweets as positive. This behavior indicates that the classifier was highly unstable since incremental learning introduced significant changes into the model with the occurrence of every new labeled tweet.

### *5.3. Stock market analysis with active learning*

Our active learning experiments showed that the best setting for learning from financial Twitter stream data is to divide tweets from the data stream into batches of 100 tweets out of which 10 tweets from each batch are selected for hand-labeling based on a querying strategy. In order to provide better randomization of the search space, we chose the strategy which combines 50% "Closest to the neutral zone" and 50% random strategy. We repeated the Granger causality analysis in order to see if this strategy improved the predictive power of financial tweets to predict the stock closing price of the Baidu company. The results can be seen in Table 5.

Since in our experiments we compute the *p*-value repetitively and make multiple comparisons of *p*-values, we applied the Bonferroni correction [1] to neutralize the problem of multiple comparisons. As in the static part of our paper (Section 4), we divide the critical *p*-value of 0.1 by 4, as this is the number of time periods (whole nine-month and three three-month periods) which we consider to be a family of tests. The *p*-values that remain significant after this correction are marked in bold in the table. As can be seen from Table 5, the best correlations were obtained for the June–August time period as in the static approach.

Additionally, we explored whether there is evidence for the reversed causality (that the price movements may influence the public sentiment). The results show that there is some causality in that direction, but after making adjustments to the critical *p*-value by applying the Bonferroni correction, no significant results were left.

### *5.4. Simulation of online experiments*

Here we present a simulation of online experiments to test whether the sentiments in tweets can predict stock prices in real time and consequently provide returns. This experimental simulation is provisional and not exhaustive, since we only experimented with a selection of basic trading strategies, but it provides an indication of the real-world value of the proposed methodology. The simulation was based on a Baidu dataset from March 14 to December 9, 2011, and the results (money + stocks' value per day) are plotted in Fig. 6. In every strategy, an investor had US$100,000 at the start. He buys 100 stocks on the first day. Buying and selling decisions for the following days depend on a selected trading strategy. The size of the neutral zone is 0.001. We experimented with three trading strategies:

**Table 5**
Statistical significance (*p*-values) of Granger causality correlation between positive sentiment probability and the closing stock price for Baidu using active learning, while changing the size of the neutral zone (i.e., the *t* value). The combined strategy for selecting 10 of 100 tweets for labeling is presented. Values which are lower than a *p*-value of 0.1, after applying the Bonferroni correction, are marked in bold.

| Size of the neutral zone (*t* value) | | 0 | 0.0001 | 0.001 | 0.01 |
|---|---|---|---|---|---|
| Select 10 of 100, combined 50% random | Lag | | | | |
| 9 Months | 1 | 0.0619 | 0.0611 | 0.0509 | 0.5883 |
| March–May | 1 | 0.8499 | 0.6551 | 0.6638 | 0.3617 |
| June–August | 1 | **0.0116** | **0.0075** | **0.0086** | 0.2732 |
| September–November | 1 | 0.9403 | 0.9572 | 0.9230 | 0.7565 |
| 9 Months | 2 | 0.0275 | 0.0326 | 0.0309 | 0.1745 |
| March–May | 2 | 0.4685 | 0.4390 | 0.5721 | 0.5757 |
| June–August | 2 | **0.0176** | **0.0163** | **0.0127** | 0.3677 |
| September–November | 2 | 0.4807 | 0.4661 | 0.4010 | 0.2200 |
| 9 Months | 3 | 0.0699 | 0.0826 | 0.0800 | 0.3177 |
| March–May | 3 | 0.3587 | 0.3886 | 0.5656 | 0.4772 |
| June–August | 3 | 0.0380 | 0.0372 | 0.0389 | 0.5708 |
| September–November | 3 | 0.3113 | 0.3292 | 0.2576 | 0.0868 |

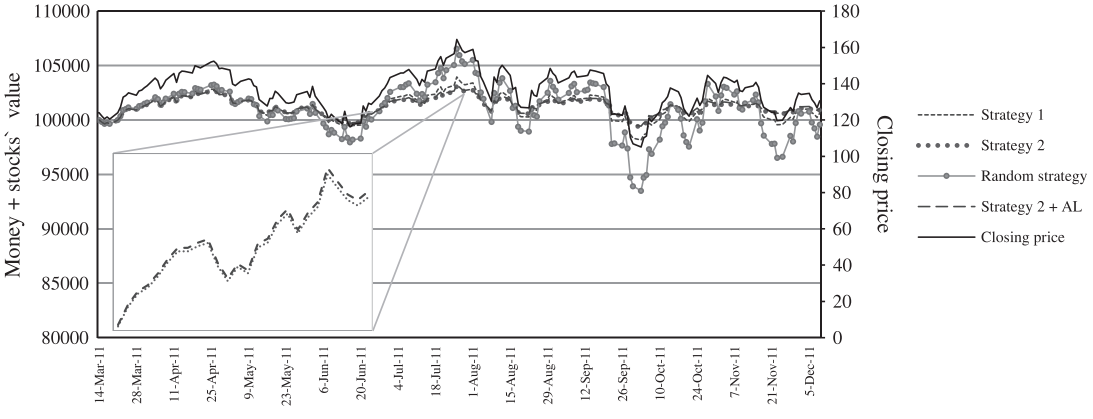


**Fig. 6.** Simulation of online experiments. The time period between June 24 and August 1 for Strategy 2 and Strategy 2 + AL is zoomed in, since in this time period, Strategy 2 + AL started to outperform Strategy 2 as a consequence of using the active learning approach.

- Strategy 1: If the daily sentiment change of the previous day was above 0, the investor buys one stock. If the daily change was below 0 and he has at least one stock, he sells it.
- Strategy 2: If the daily sentiment change of the previous day was above 0.05, the investor buys one stock. If the daily change was below 0.05 and he has at least one stock, he sells it.
- Random: Every day the investor randomly chooses whether to buy or sell one stock.

In the first three columns of Table 6, we present average daily values (money + stocks value) for Strategy 1, Strategy 2, and the Random strategy. From Fig. 6 and the average daily values in Table 6, it follows that Strategy 2 outperforms Strategy 1 and the Random strategy. Therefore, we combined the active learning algorithm with Strategy 2 to check whether the active learning approach could further improve this strategy. It turned out that, indeed, active learning improves the results as it can be observed in the time series "Strategy 2 + AL" of Fig. 6 and in the fourth column of Table 6. At the beginning of the simulation, Strategy 2 and Strategy 2 + AL had the same performance. At the end of June, Strategy 2 + AL started to outperform Strategy 2 and remained better until the end of the simulation. In the figure, the period where the first change in performance of these two strategies occurs is zoomed in. For this experiment, we used the "Combination 50% random" active learning approach and "Select 10 of 100" batch selections with neutral zone 0.001.

Therefore, the simulation of online experiments indicates that our approach is useful for online stock trading. The lowest performance was observed with the random strategy, while the best performance was obtained with the strategy which uses the active learning approach.

### *5.5. Summary of the proposed methodology for stream-based active learning for Twitter sentiment analysis in finance*

The proposed methodology for stream-based active learning for Twitter sentiment analysis in finance consists of the sequence of methodological steps presented in Fig. 7. Components which are specific to the stream-based setting and not present in the static setting (Fig. 2) are colored gray.

As can be seen from the figure, one should first provide a stream of financial tweets discussing stock relevant information concerning a company of interest. The algorithm then collects a batch of examples from the stream and conducts preprocessing on them. After classification of tweets as positive, negative, or neutral, based on a querying strategy, the algorithm selects tweets for hand-labeling. With this new labeled data, the model is updated. These steps are repeated for all batches in the data stream. For every day of a time series, the positive sentiment probability is computed by dividing the number of positive tweets per day by the total number of tweets in that day. Lastly, daily changes of the positive sentiment probability are calculated.

On the other hand, the stock closing prices of a selected company for each day should be collected. The daily returns in the stock closing price are then calculated (as presented in Section 4) in order to satisfy stationary conditions demanded by the Granger causality test.

**Table 6**
Average daily values (money + stocks value) for every strategy.

| Strategy 1 | Strategy 2 | Random strategy | Strategy 2 + AL |
|---|---|---|---|
| 101205.24 | 101214.42 | 101007.43 | 101269.81 |

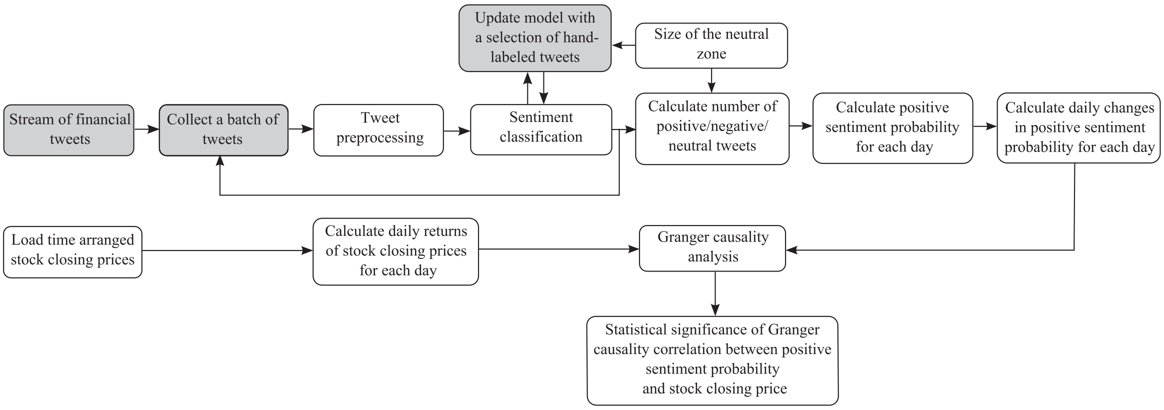

Update model with a selection of hand-labeled tweets
Size of the neutral zone
Stream of financial tweets
Collect a batch of tweets
Tweet preprocessing
Sentiment classification
Calculate number of positive/negative/ neutral tweets
Calculate positive sentiment probability for each day
Calculate daily changes in positive sentiment probability for each day
Load time arranged stock closing prices
Calculate daily returns of stock closing prices for each day
Granger causality analysis
Statistical significance of Granger causality correlation between positive sentiment probability and stock closing price

**Fig. 7.** Methodological steps for stream-based active learning for Twitter sentiment analysis in finance. Components which are specific to the stream-based setting and not present in the static setting (Fig. 2) are colored gray.

Given the daily changes of the positive sentiment probability time series and the daily returns in the stock closing price time series, Granger causality analysis is performed (considering lagged values of time series for one, two, and three days) to test whether tweet sentiment is useful for forecasting the movement of prices in the stock market and how significant the results are.

## 6. Conclusions

Predicting future values of stock prices is an interesting task, commonly connected to the analysis of public mood. Given that more and more personal opinions are made available online, various studies indicate that these kinds of analyses can be automated and can produce useful results. This paper investigates whether Twitter feeds are a suitable data source for predictive sentiment analysis. The study indicates that sentiment analysis of public mood derived from Twitter feeds can be used to eventually forecast movements of individual stock prices. Additionaly, the SVM neutral zone gave us the ability to classify tweets into the neutral category and proved to be useful for improving the predictive power by strengthening the correlation between the opinionated tweets and the stock closing price.

Furthermore, the methodology was adapted to a stream-based setting using the incremental active learning approach, which provides the algorithm with the ability to choose new training data from a data stream for hand-labeling. A series of experiments was conducted to find the best querying strategy for financial Twitter data. The experiments indicate that by using the active learning approach, the prediction power of the sentiment classifier in the stock market application is further improved. With our study, we introduced stream-based active learning for sentiment analysis of microblogging messages in the financial domain, which contributes both to sentiment analysis and the active learning research area, since this issue is still insufficiently explored.

Our approach seems to be useful also for online stock trading. We presented initial experimental results of a simulation where we tested whether the sentiments in tweets can predict values of future stock prices in real time and, consequently, provide returns. Initial results indicate that by augmenting a trading strategy with consideration of the changes in the values of positive sentiment probability one could improve the returns.

In further work, we plan to expand the number of companies to further test our stream-based methodology for sentiment analysis of microblogging messages in the financial domain. Moreover, our work would benefit from using other (preferably hand-labeled) datasets in order to obtain more realistic performance estimates in tests of preprocessing settings and an even better initial sentiment classifier. Since in this study we used a rather simple neutral zone, we plan to improve its sophistication and adaptability with time. Finally, we plan to improve our methodology by developing techniques of specific tweet concept drift detection and by tackling the detection of irony and sarcasm.

## Acknowledgments

The work presented in this paper was partially funded by the European Commission in the context of the FP7 projects FOC and FIRST (Grant Agreement Nos. 255987 and 257928, respectively), and the Ad Futura Programme of the Slovenian Human Resources and Scholarship Fund. We are grateful to Ulli Spankowski and Sebastian Schroff for their kind cooperation as financial experts in the stock analytics application presented in this paper. We are also grateful to Dragi Kocev and Vladimir Kuzmanovski from Jožef Stefan Institute, Ljubljana, Slovenia, for their help in the statistical evaluation of the results, and to Igor Mozetič for useful comments and suggestions. Finally, we are grateful to Martin Saveski for his help with the implementation of the active learning algorithms.

## Appendix A

In this appendix, we present some empirical support for considering smiley-labeled tweets as a reasonable approximation for manually-annotated positive/negative sentiments of tweets.

Table 7 presents the most relatively positive and negative sentiment-bearing terms from the smiley-labeled Stanford dataset after replacing URLs and negation words with common tokens *URL* and *NEGATION*. Since some terms were presented in both positive and negative tweets, we took the 1000 terms with the highest document frequency in positive tweets, and the 1000 terms with the highest document frequency in negative tweets and calculated the difference between document frequencies of individual terms. From the table, it follows that positive/negative smiley-labeled tweets contain numerous common positive/negative sentiment-bearing words, such as ‘‘thanks,’’ ‘‘love,’’ ‘‘good,’’ and ‘‘great’’ for positive sentiments, and ‘‘miss,’’ ‘‘sad,’’ and ‘‘bad’’ for negative sentiments.

Interestingly, from the table it follows that mentioning other users (by writing the ‘@’ sign and a username) and writing web URLs is associated with positive emoticons, and probably positive feelings. On the other hand, it seems that writing negation words (e.g., ‘‘not,’’ ‘‘isn't,’’ ‘‘aren't,’’ ‘‘wasn't,’’ etc.) is associated with negative emoticons, and probably negative feelings.

To address this concern even further, we conducted an additional experiment: we obtained a manually-labeled collection (subset) of smiley-labeled tweets and computed how accurate emoticons are as labels. Their accuracy on 1500 positive and

**Table 7**
The most relatively positive and negative sentiment-bearing terms from the smiley-labeled Stanford dataset.

| Positive terms | Document frequency difference | Negative terms | Document frequency difference |
|---|---|---|---|
| ‘@’ | 134,856 | ‘go’ | −74,192 |
| ‘!’ | 77,611 | ‘NEGATION’ | −69,678 |
| ‘going’ | 55,777 | ‘miss’ | −37,075 |
| ‘thanks’ | 41,668 | ‘work’ | −28,721 |
| ‘love’ | 36,643 | ‘.’ | −28,414 |
| ‘good’ | 31,178 | ‘sad’ | −28,144 |
| ‘URL’ | 22,168 | ‘t’ | −27,682 |
| ‘look’ | 21,239 | “ t’ | −27,062 |
| ‘,’ | 21,027 | ‘want’ | −22,440 |
| ‘great’ | 16,474 | ‘. . .’ | −21,947 |
| ‘–’ | 16,432 | ‘feel’ | −21,684 |
| ‘&’ | 15,483 | ‘wish’ | −18,981 |
| ‘happy’ | 13,679 | ‘looking’ | −17,395 |
| ‘;’ | 13,040 | ‘can ” | −17,048 |
| ‘lol’ | 12,581 | ‘bad’ | −16,580 |

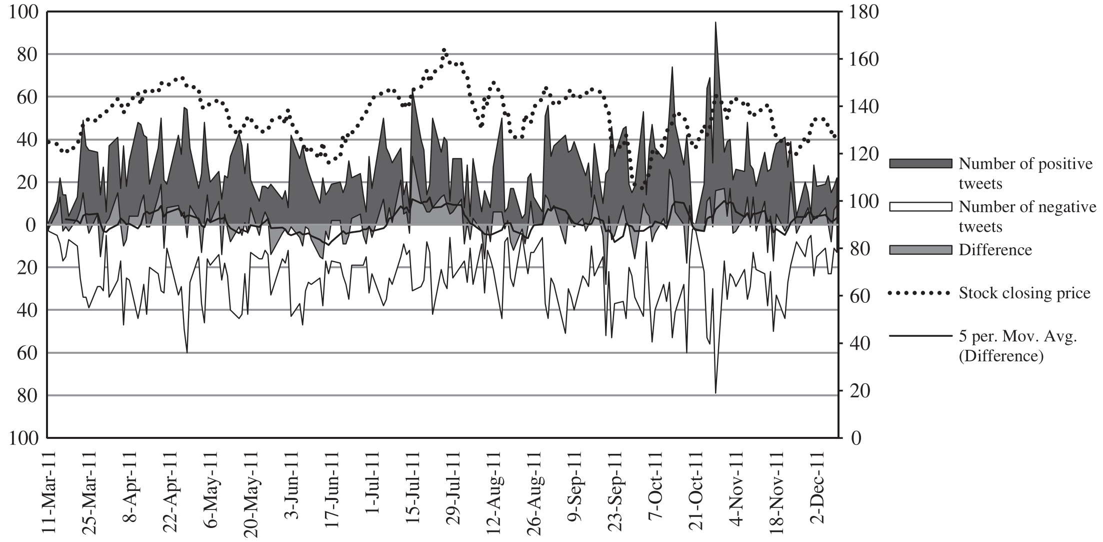


**Fig. 8.** Number of tweet posts classified as positive or negative, their difference, the moving average of the difference (averaged over 5 days), and the stock closing price per day for Baidu.

negative manually-labeled tweets was 86.40%. This result provides an error estimate and illustrates that smiley-labeled tweets are a reasonable approximation for manually-annotated positive/negative sentiment of tweets.

Furthermore, using the best preprocessing setting, as explained in Section 3.4, we trained a classifier and classified the Baidu tweets with the learned SVM classifier into one of two categories (positive or negative), counted the number of positive and negative tweets for each day of the time series, and plotted them together with their difference, the moving average of the difference (averaged over five days), and the stock closing price per day. The visual presentation of the sentiment time series for Baidu can be seen in Fig. 8. The peaks show the days when people intensively tweeted about the stocks. In the experiments for the visualization, we classified only the tweets whose dates corresponded to the dates when the stock market was open. The rest of the tweets (e.g., tweets which were written on weekends or non-working days of the stock market) were discarded and not analyzed.

In order to inspect how accurate the smiley-based sentiment classifier is, we calculated the *F*-measure on all of the hand-labeled tweets from the Baidu dataset using the sentiment classifier learned on the Stanford smiley-labeled dataset. In this experiment we achieved an *F*-measure of 0.517 (baseline is 0.33). We have also calculated the *F*-measure on the positive and negative hand-labeled Baidu tweets and achieved an *F*-measure of 0.671. Additionally, we employed ten-fold cross-validation experiments, using SVM on the all (positive, negative and neutral) hand-labeled tweets from the Baidu dataset and achieved average *F*-measure of 0.645.

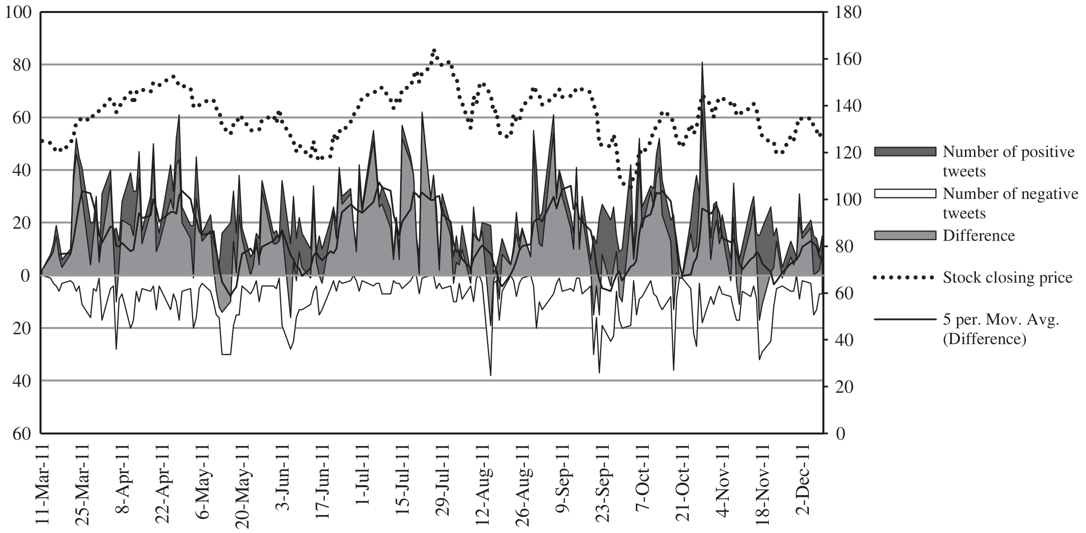


**Fig. 9.** Number of hand-labeled positive and negative tweet posts, their difference, the moving average of the difference (averaged over 5 days), and the stock closing price per day for Baidu.

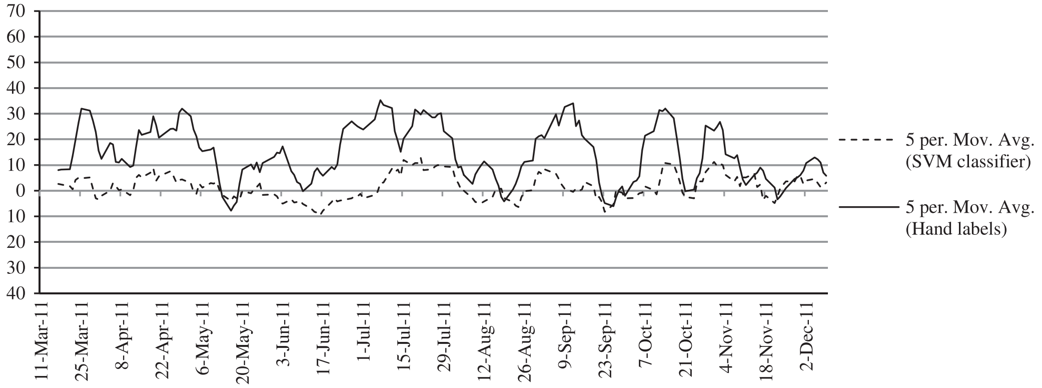


**Fig. 10.** The moving average of the difference (averaged over 5 days) for hand-labeled positive and negative tweets and the ones classified as positive or negative by the SVM sentiment classifier.

Given the manually labeled 11389 Baidu tweets, we were able to plot a figure also for this data (see Fig. 9) based on true positive and negative labels. Additionally, we plotted a graph (see Fig. 10) which presents the moving average of the sentiment difference (averaged over five days) for hand-labeled positive and negative tweets and the ones classified as positive or negative by our SVM sentiment classifier. As it can be seen from the figure, the biggest peaks of differences and general trend remain basically the same which leads to conclusion that classification obtained with our classifier learned on smiley-labeled tweets is comparable with manual labels.

## Appendix B

This appendix reports experimental results of Granger causality correlation between positive sentiment probability and closing stock price for 8 companies (Apple, Amazon, Baidu, Cisco, Google, Microsoft, Netflix and Research In Motion). Results are shown in Table 8.

**Table 8**
Statistical significance ($p$-values) of Granger causality correlation between positive sentiment probability and closing stock price for 8 companies, while changing the size of the neutral zone, i.e., the $t$ value from 0 to 1. Values which are lower than $p$-value of 0.1 are marked with bold.

| Size of neutral zone ($t$ value) | | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time period | Lag | | | | | | | | | | | |
| *Apple* | | | | | | | | | | | | |
| 9 Months | 1 | 0.7446 | 0.9211 | 0.6220 | 0.4407 | 0.3576 | 0.4793 | 0.4235 | 0.3455 | 0.4696 | 0.4415 | 0.5168 |
| March–May | 1 | 0.4480 | 0.5921 | 0.3978 | 0.6687 | 0.6496 | 0.7556 | 0.3841 | 0.3485 | 0.2970 | 0.1969 | 0.2378 |
| June–August | 1 | 0.7307 | 0.5956 | 0.9755 | 0.6513 | 0.7153 | 0.9270 | 0.8409 | 0.9704 | 0.8173 | 0.8370 | 0.8726 |
| September–November | 1 | 0.5228 | 0.7405 | 0.7477 | 0.5463 | 0.3463 | 0.2288 | 0.2625 | 0.2773 | 0.3874 | 0.5021 | 0.6250 |
| 9 Months | 2 | 0.8440 | 0.9694 | 0.9154 | 0.5005 | 0.5766 | 0.6619 | 0.3964 | 0.3187 | 0.4171 | 0.5550 | 0.7341 |
| March–May | 2 | **0.0085** | 0.2395 | 0.2613 | 0.1452 | 0.2879 | 0.2202 | **0.0707** | 0.1157 | **0.0927** | 0.2053 | 0.2636 |
| June–August | 2 | 0.5560 | 0.6098 | 0.8244 | 0.9440 | 0.8974 | 0.8375 | 0.9658 | 0.9921 | 0.9389 | 0.9234 | 0.9437 |
| September–November | 2 | 0.8773 | 0.9781 | 0.9881 | 0.6626 | 0.4476 | 0.4147 | 0.3597 | 0.2557 | 0.4301 | 0.5279 | 0.6131 |
| 9 Months | 3 | 0.9307 | 0.9421 | 0.9291 | 0.8461 | 0.8695 | 0.9567 | 0.8628 | 0.7414 | 0.7986 | 0.7271 | 0.7277 |
| March–May | 3 | **0.0917** | 0.3378 | 0.4054 | 0.3662 | 0.5031 | 0.5769 | 0.3614 | 0.5064 | 0.4660 | 0.3180 | 0.4197 |
| June–August | 3 | 0.6805 | 0.7147 | 0.6528 | 0.8010 | 0.7286 | 0.6720 | 0.5144 | 0.4786 | 0.5496 | 0.4397 | 0.3039 |
| September–November | 3 | 0.9054 | 0.9874 | 0.9887 | 0.7135 | 0.5025 | 0.4639 | 0.3932 | 0.2521 | 0.4951 | 0.6118 | 0.7319 |
| *Amazon* | | | | | | | | | | | | |
| 9 Months | 1 | 0.6345 | 0.9798 | 0.9552 | 0.9362 | 0.6590 | 0.7368 | 0.3633 | 0.4433 | 0.7776 | 0.8148 | 0.6641 |
| March–May | 1 | 0.9912 | 0.7029 | 0.5652 | 0.5190 | 0.4136 | 0.4510 | 0.8119 | 0.7580 | 0.5641 | 0.6111 | 0.5377 |
| June–August | 1 | 0.3837 | 0.6021 | 0.8388 | 0.5821 | 0.9085 | 0.9921 | 0.5518 | 0.7976 | 0.8120 | 0.6791 | 0.9950 |
| September–November | 1 | 0.9036 | 0.7460 | 0.4902 | 0.8200 | 0.8425 | 0.9537 | 0.5568 | 0.4676 | 0.7286 | 0.5466 | 0.6946 |
| 9 Months | 2 | 0.5635 | 0.9153 | 0.8843 | 0.7442 | 0.2989 | 0.3351 | 0.1557 | 0.2282 | 0.2816 | 0.2323 | 0.3285 |
| March–May | 2 | 0.4927 | 0.2821 | 0.3812 | 0.4562 | 0.4542 | 0.4483 | 0.5508 | 0.9571 | 0.8094 | 0.8909 | 0.8664 |
| June–August | 2 | 0.5519 | 0.7992 | 0.9382 | 0.8123 | 0.6380 | 0.8333 | 0.8178 | 0.9929 | 0.8619 | 0.8835 | 0.8987 |
| September–November | 2 | 0.9756 | 0.8713 | 0.6644 | 0.1261 | **0.0334** | **0.0231** | **0.0042** | **0.0143** | **0.0458** | **0.0166** | **0.0232** |
| 9 Months | 3 | 0.7602 | 0.9759 | 0.9614 | 0.8624 | 0.5145 | 0.5477 | 0.3060 | 0.3685 | 0.4925 | 0.4106 | 0.4868 |
| March–May | 3 | 0.4056 | 0.2370 | 0.1970 | 0.1338 | 0.1336 | 0.1831 | 0.4617 | 0.8733 | 0.8205 | 0.6032 | 0.9265 |
| June–August | 3 | 0.8305 | 0.9054 | 0.9227 | 0.9161 | 0.7835 | 0.9103 | 0.8925 | 0.9998 | 0.7835 | 0.7259 | 0.4328 |
| September–November | 3 | 0.9838 | 0.8480 | 0.6074 | 0.1075 | **0.0344** | **0.0253** | **0.0070** | **0.0176** | **0.0475** | **0.0171** | **0.0291** |
| *Baidu* | | | | | | | | | | | | |
| 9 Months | 1 | **0.0661** | 0.3726 | 0.4165 | 0.2280 | 0.3184 | 0.5041 | 0.2134 | 0.2171 | 0.5494 | 0.5845 | 0.9052 |
| March–May | 1 | 0.4035 | 0.3062 | 0.4633 | 0.3588 | 0.4391 | 0.3665 | 0.5424 | 0.8645 | 0.9701 | 0.8957 | 0.6142 |
| June–August | 1 | **0.0686** | 0.1243 | 0.2025 | **0.0697** | 0.1639 | 0.3923 | 0.1574 | **0.0681** | 0.2403 | 0.3761 | 0.3399 |
| September–November | 1 | **0.0607** | 0.3992 | 0.4195 | 0.3766 | 0.3720 | 0.2428 | 0.1931 | 0.6731 | 0.8376 | 0.7920 | 0.9692 |
| 9 Months | 2 | **0.0470** | 0.3006 | 0.3374 | 0.2620 | 0.2491 | 0.3826 | 0.2988 | 0.3877 | 0.5397 | 0.5178 | 0.8296 |
| March–May | 2 | 0.4704 | 0.4031 | 0.5336 | 0.4136 | 0.3805 | 0.1335 | **0.0332** | **0.0409** | 0.3572 | 0.1630 | 0.1073 |
| June–August | 2 | **0.0501** | **0.0378** | **0.0327** | **0.0393** | **0.0200** | **0.0116** | **0.0043** | **0.0101** | 0.1218 | 0.1401 | 0.3111 |
| September–November | 2 | **0.0693** | 0.5913 | 0.6412 | 0.5719 | 0.5935 | 0.4454 | 0.3829 | 0.8635 | 0.7595 | 0.6386 | 0.7422 |
| 9 Months | 3 | **0.0415** | 0.2945 | 0.3858 | 0.2988 | 0.3226 | 0.4245 | 0.3593 | 0.3673 | 0.5434 | 0.5150 | 0.8300 |
| March–May | 3 | 0.6636 | 0.6133 | 0.7557 | 0.6609 | 0.5678 | 0.2033 | **0.0503** | 0.1038 | 0.5105 | 0.3403 | 0.1969 |
| June–August | 3 | **0.0984** | **0.0759** | **0.0590** | **0.0995** | **0.0499** | **0.0208** | **0.0112** | **0.0291** | 0.1995 | 0.1708 | 0.2932 |
| September–November | 3 | **0.0283** | 0.2774 | 0.3405 | 0.2644 | 0.3434 | 0.4368 | 0.4709 | 0.7899 | 0.8771 | 0.8048 | 0.8981 |
| *Cisco* | | | | | | | | | | | | |
| 9 Months | 1 | 0.3118 | 0.3725 | 0.3629 | 0.7742 | 0.9601 | 0.9658 | 0.9345 | 0.7237 | 0.3475 | 0.4044 | 0.4781 |
| March–May | 1 | 0.9664 | 0.3588 | 0.4149 | 0.3774 | 0.2866 | 0.1402 | 0.1928 | 0.1103 | 0.4691 | 0.6100 | 0.8670 |
| June–August | 1 | 0.6395 | 0.8037 | 0.5996 | 0.9649 | 0.5772 | 0.5328 | 0.5097 | 0.3414 | 0.1090 | 0.2052 | 0.2465 |
| September–November | 1 | 0.4107 | 0.5913 | 0.8078 | 0.9287 | 0.9568 | 0.9370 | 0.9931 | 0.9015 | 0.8982 | 0.8119 | 0.9346 |
| 9 Months | 2 | 0.2970 | 0.7836 | 0.6619 | 0.7989 | 0.7312 | 0.6285 | 0.7809 | 0.8541 | 0.6064 | 0.5774 | 0.6835 |
| March–May | 2 | 0.9149 | 0.3060 | 0.5873 | 0.6915 | 0.5558 | 0.3647 | 0.4760 | 0.3713 | 0.7580 | 0.9703 | 0.9714 |
| June–August | 2 | 0.7471 | 0.9557 | 0.7590 | 0.6253 | 0.4594 | 0.4041 | 0.5016 | 0.5058 | 0.2081 | 0.3886 | 0.4378 |
| September–November | 2 | 0.3710 | 0.6240 | 0.8306 | 0.8808 | 0.9401 | 0.9636 | 0.9390 | 0.9507 | 0.9135 | 0.8185 | 0.8616 |
| 9 Months | 3 | 0.4299 | 0.9543 | 0.9225 | 0.8689 | 0.8052 | 0.6434 | 0.6323 | 0.8036 | 0.4910 | 0.3918 | 0.4664 |
| March–May | 3 | 0.9813 | 0.6299 | 0.7885 | 0.7101 | 0.5140 | 0.3389 | 0.4745 | 0.2374 | 0.3343 | 0.1625 | 0.3438 |
| June–August | 3 | 0.7944 | 0.9506 | 0.7932 | 0.4264 | 0.2766 | 0.1812 | 0.2594 | 0.4166 | 0.1520 | 0.1781 | 0.1865 |
| September–November | 3 | 0.5070 | 0.7070 | 0.9287 | 0.9313 | 0.9693 | 0.9239 | 0.7458 | 0.8369 | 0.8635 | 0.7775 | 0.8375 |
| *Google* | | | | | | | | | | | | |
| 9 Months | 1 | 0.8868 | 0.5420 | 0.9140 | 0.9694 | 0.9395 | 0.9307 | 0.9332 | 0.9452 | 0.7282 | 0.9186 | 0.9919 |
| March–May | 1 | **0.0411** | 0.1051 | **0.0139** | **0.0193** | **0.0142** | **0.0290** | **0.0447** | **0.0497** | **0.0641** | **0.0498** | **0.0753** |
| June–August | 1 | 0.7607 | 0.5689 | 0.6991 | 0.8971 | 0.8554 | 0.9389 | 0.9629 | 0.7535 | 0.5878 | 0.9496 | 0.7805 |
| September–November | 1 | **0.0141** | **0.0116** | **0.0067** | **0.0014** | **0.0209** | **0.0314** | **0.0616** | **0.0833** | 0.1302 | 0.2168 | 0.4405 |
| 9 Months | 2 | 0.9518 | 0.8147 | 0.7848 | 0.9980 | 0.9790 | 0.9849 | 0.9708 | 0.9173 | 0.6978 | 0.7798 | 0.6635 |

**Table 8** (*continued*)

| Size of neutral zone ($t$ value) | | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time period | Lag | | | | | | | | | | | |
| March–May | 2 | **0.0760** | 0.2664 | **0.0477** | **0.0653** | **0.0457** | **0.0653** | **0.0766** | 0.1059 | 0.1125 | **0.0867** | 0.1347 |
| June–August | 2 | 0.6283 | 0.5343 | 0.5107 | 0.6456 | 0.5366 | 0.6455 | 0.8501 | 0.6965 | 0.4376 | 0.7773 | 0.8606 |
| September–November | 2 | **0.0138** | **0.0099** | **0.0246** | **0.0021** | **0.0277** | **0.0324** | 0.1072 | **0.0881** | 0.1812 | 0.3162 | 0.4338 |
| 9 Months | 3 | 0.6646 | 0.5766 | 0.6087 | 0.5101 | 0.4084 | 0.4739 | 0.7731 | 0.8193 | 0.5283 | 0.7775 | 0.7141 |
| March–May | 3 | **0.0990** | 0.1981 | **0.0534** | 0.1313 | 0.1428 | 0.1851 | 0.1940 | 0.1606 | 0.1845 | 0.1165 | 0.1159 |
| June–August | 3 | 0.4358 | 0.3669 | 0.3011 | 0.2063 | 0.1550 | 0.1852 | 0.4540 | 0.5222 | 0.2428 | 0.4229 | 0.5547 |
| September–November | 3 | **0.0284** | **0.0187** | **0.0286** | **0.0025** | **0.0517** | **0.0568** | 0.1980 | 0.1729 | 0.2672 | 0.3963 | 0.4534 |
| *Microsoft* | | | | | | | | | | | | |
| 9 Months | 1 | 0.6679 | 0.3568 | 0.9189 | 0.9904 | 0.6432 | 0.5673 | 0.4413 | 0.1396 | 0.1362 | 0.1637 | 0.1617 |
| March–May | 1 | 0.6332 | 0.2487 | 0.4302 | 0.8703 | 0.8729 | 0.9392 | 0.5836 | 0.3472 | 0.4615 | 0.5502 | 0.5186 |
| June–August | 1 | 0.7526 | 0.7787 | 0.7475 | 0.8899 | 0.5411 | 0.6378 | 0.7815 | 0.4765 | 0.2588 | 0.3150 | 0.3336 |
| September–November | 1 | 0.4650 | 0.7614 | 0.7444 | 0.6252 | 0.5487 | 0.7196 | 0.8068 | 0.9712 | 0.9020 | 0.9952 | 0.9776 |
| 9 Months | 2 | 0.9182 | 0.4222 | 0.7743 | 0.9600 | 0.6652 | 0.7169 | 0.7694 | 0.3678 | 0.2956 | 0.3127 | 0.3655 |
| March–May | 2 | 0.7817 | **0.0796** | 0.1162 | 0.2965 | 0.2726 | 0.6301 | 0.9859 | 0.7704 | 0.8678 | 0.8381 | 0.8103 |
| June–August | 2 | 0.8801 | 0.9219 | 0.7157 | 0.5411 | 0.4959 | 0.7137 | 0.4260 | 0.4561 | 0.2599 | 0.2092 | 0.3626 |
| September–November | 2 | 0.6129 | 0.4038 | 0.1949 | 0.3665 | 0.3150 | 0.3455 | 0.6157 | 0.7471 | 0.9247 | 0.8871 | 0.8367 |
| 9 Months | 3 | 0.1963 | 0.2644 | 0.8766 | 0.7262 | 0.4828 | 0.5401 | 0.7305 | 0.4060 | 0.2596 | 0.4274 | 0.4639 |
| March–May | 3 | 0.6349 | 0.2929 | 0.3842 | 0.6915 | 0.7104 | 0.8493 | 0.7085 | 0.7042 | 0.9156 | 0.8840 | 0.7872 |
| June–August | 3 | 0.7867 | 0.8895 | 0.9172 | 0.7417 | 0.6345 | 0.9046 | 0.7276 | 0.7819 | 0.5362 | 0.5037 | 0.7192 |
| September–November | 3 | **0.0769** | 0.1053 | 0.1184 | 0.2333 | 0.2847 | 0.3899 | 0.6335 | 0.5726 | 0.5449 | 0.6526 | 0.6384 |
| *Netflix* | | | | | | | | | | | | |
| 9 Months | 1 | 0.9418 | 0.5072 | 0.5229 | 0.6627 | 0.7087 | 0.6704 | 0.5634 | 0.3848 | 0.2457 | 0.2784 | 0.1886 |
| March–May | 1 | 0.3323 | 0.2970 | 0.9020 | 0.9091 | 0.6241 | 0.1399 | **0.0695** | 0.1279 | 0.5991 | 0.9096 | 0.7988 |
| June–August | 1 | 0.2645 | 0.1996 | 0.1929 | 0.3070 | **0.0654** | 0.1364 | 0.2063 | 0.4115 | 0.9797 | 0.8229 | 0.7951 |
| September–November | 1 | 0.8447 | 0.8630 | 0.9653 | 0.9798 | 0.7571 | 0.5296 | 0.5541 | 0.4377 | 0.3088 | 0.3520 | 0.2598 |
| 9 Months | 2 | **0.0025** | **0.0089** | **0.0162** | **0.0322** | **0.0573** | **0.0588** | 0.1097 | 0.2117 | 0.1163 | **0.0920** | 0.1188 |
| March–May | 2 | 0.6181 | 0.5854 | 0.9448 | 0.9558 | 0.8478 | 0.8944 | 0.7219 | 0.8333 | 0.9406 | 0.8737 | 0.7988 |
| June–August | 2 | 0.7579 | 0.6329 | 0.6350 | 0.7644 | 0.1848 | 0.2453 | 0.3260 | 0.3848 | 0.5909 | 0.5712 | 0.3979 |
| September–November | 2 | **0.0005** | **0.0026** | **0.0067** | **0.0064** | **0.0054** | **0.0048** | **0.0134** | **0.0250** | **0.0199** | **0.0175** | **0.0174** |
| 9 Months | 3 | **0.0026** | **0.0099** | **0.0145** | **0.0200** | **0.0292** | **0.0338** | **0.0791** | 0.1612 | **0.0969** | **0.0400** | **0.0543** |
| March–May | 3 | 0.5532 | 0.8481 | 0.9888 | 0.9874 | 0.8801 | 0.9556 | 0.8333 | 0.9016 | 0.8808 | 0.8962 | 0.7288 |
| June–August | 3 | 0.9717 | 0.8966 | 0.9062 | 0.9196 | 0.5645 | 0.6326 | 0.6748 | 0.6902 | 0.7521 | 0.5838 | 0.3437 |
| September–November | 3 | **0.0007** | **0.0039** | **0.0080** | **0.0065** | **0.0042** | **0.0039** | **0.0155** | **0.0300** | **0.0270** | **0.0137** | **0.0168** |
| *Research in motion* | | | | | | | | | | | | |
| 9 Months | 1 | **0.0467** | 0.1974 | 0.2492 | 0.4061 | 0.5319 | 0.4297 | 0.4333 | 0.3741 | 0.3235 | **0.0686** | **0.0222** |
| March–May | 1 | 0.1361 | **0.0965** | **0.0355** | **0.0386** | **0.0415** | 0.1453 | 0.1634 | 0.2533 | 0.3042 | 0.3317 | 0.3578 |
| June–August | 1 | 0.4033 | 0.3245 | 0.5356 | 0.9549 | 0.5168 | 0.8072 | 0.9564 | 0.7196 | 0.8101 | 0.2457 | 0.1698 |
| September–November | 1 | 0.2127 | 0.8165 | 0.8799 | 0.9744 | 0.5442 | 0.5476 | 0.9324 | 0.9327 | 0.7226 | 0.4989 | 0.1960 |
| 9 Months | 2 | 0.1076 | 0.2965 | 0.2422 | 0.4390 | 0.7683 | 0.5019 | 0.4867 | 0.3297 | 0.1858 | 0.1050 | **0.0257** |
| March–May | 2 | 0.3524 | 0.1979 | 0.1240 | 0.1237 | 0.1176 | 0.1639 | 0.2027 | 0.2781 | 0.3745 | 0.3021 | 0.2665 |
| June–August | 2 | 0.5624 | 0.4978 | 0.7891 | 0.9998 | 0.7448 | 0.7990 | 0.8972 | 0.9059 | 0.9779 | 0.2865 | 0.3614 |
| September–November | 2 | **0.0859** | 0.1703 | 0.1436 | 0.3715 | 0.4917 | 0.1120 | 0.1502 | **0.0578** | **0.0368** | **0.0268** | **0.0221** |
| 9 Months | 3 | **0.0609** | 0.1407 | **0.0512** | 0.2095 | 0.7124 | 0.4059 | 0.4169 | 0.3591 | 0.1648 | **0.0459** | **0.0138** |
| March–May | 3 | 0.5580 | 0.3390 | 0.1449 | 0.1117 | **0.0853** | **0.0572** | **0.0572** | 0.1329 | 0.4074 | 0.2432 | 0.4005 |
| June–August | 3 | 0.6641 | 0.6635 | 0.7495 | 0.9885 | 0.9445 | 0.9682 | 0.9852 | 0.9679 | 0.9522 | 0.3449 | 0.3866 |
| September–November | 3 | **0.0884** | 0.1365 | **0.0899** | 0.1989 | 0.4663 | 0.1926 | 0.2526 | 0.1356 | **0.0763** | **0.0419** | **0.0344** |

## References

[1] H. Abdi, Bonferroni and Šidák corrections for multiple comparisons, in: N.J. Salkind (Ed.), Encyclopedia of Measurement and Statistics, 2007.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment analysis of twitter data, in: Proceedings of the Workshop on Languages in Social Media, 2011, pp. 30–38.

[3] S. Asur, B.A. Huberman, Predicting the future with social media, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp. 492–499.

[4] L. Bachelier, Théorie de la Spéculation, Gauthier-Villars, 1900.

[5] A. Bifet, R. Kirkby, Data Stream Mining: A Practical Approach, 2009.

[6] A. Bifet, E. Frank, Sentiment Knowledge Discovery in Twitter Streaming Data, Discovery Science, Springer, Berlin Heidelberg, 2010. pp. 1–15.

[7] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[8] A. Broder, S. Glassman, M. Manasse, G. Zweig, Syntactic clustering of the web, in: Proceedings of the 6th International World Wide Web Conference, 1997, pp. 393–404.

[9] K.C. Butler, S.J. Malaikah, Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia, J. Bank. Finan. 16 (1) (1992) 197–210.

[10] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.

[11] R. Chen, M. Lazer, Sentiment analysis of twitter feeds for the prediction of stock market movement, in: CS 229 Machine Learning: Final Project, 2011.

[12] C. Cortes, V.N. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[13] A.R. Damasio, Descartes Error: Emotion, Reason, and the Human Brain, Harper Perennial, 1995.

[14] S. Das, M. Chen, Yahoo! for Amazon: extracting market sentiment from stock message boards, in: Proceedings of the 8th the Asia Pacific Finance Association Annual Conference (APFA), 2001.

[15] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[16] L. Dugan, Twitter to surpass 500 million registered users on Wednesday, 2012 <http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842> (accessed 01.02.13).

[17] E. Fama, Random walks in stock market prices, Finan. Anal. J. 21 (5) (1965) 55–59.

[18] R. Feldman, J. Sanger, The Text Mining Handbook – Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007.

[19] Y. Freund, H. Seung, E. Tishby, Selective sampling using the query by committee algorithm, Mach. Learn. 28 (2-3) (1997) 133–168.

[20] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Statist. (1940) 86–92.

[21] A. Go, R. Bhayani, L. Huang, Twitter Sentiment Classification Using Distant Supervision, CS224N Project Report, Stanford, 2009.

[22] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (1969) 424–438.

[23] D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomkins, The predictive power of online chatter, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005, pp. 78–87.

[24] E. Haddi, X. Liu, Y. Shi, The role of text pre-processing in sentiment analysis, Proc. Comp. Sci. 17 (2013) 26–32.

[25] E. Ikonomovska, J. Gama, S. DŽeroski, Learning model trees from evolving data streams, Data Min. Knowl. Disc. 23 (1) (2011) 128–168.

[26] E. Ikonomovska, Algorithms for Learning Regression Trees and Ensembles on Evolving Data Streams, Doctoral Dissertation, 2012.

[27] R.L. Iman, J.M. Davenport, Approximations of the critical region of the Friedman statistic, Commun. Statist. – Theory Meth. 9 (6) (1980) 571-–595.

[28] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent Twitter sentiment classification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 151—160.

[29] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: Proceedings of the European Conference on Machine Learning, 1998, pp. 137–142.

[30] T. Joachims, A support vector method for multivariate performance measures, in: Proceedings of the 22nd International Conference on Machine Learning (ICML), 2005.

[31] T. Joachims, Training linear SVMs in linear time, in: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006, pp. 217–226.

[32] T. Joachims, C.-N.J. Yu, Sparse kernel SVMs via cutting-plane training, Mach. Learn. 76 (2–3) (2009) 179–193.

[33] M. Kavussanos, E. Dockery, A multivariate test for stock market efficiency: the case of ASE, Appl. Finan. Econ. 11 (5) (2001) 573–579.

[34] E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: the good the bad and the OMG!, in: ICWSM, 2011.

[35] D. Lewis, W. Gale, A sequential algorithm for training text classifiers, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 3–12.

[36] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, 2012.

[37] J. Martineau, T. Finin, Delta TFIDF: an improved feature space for sentiment analysis, in: Proceedings of the Third AAAI International Conference on Weblogs and Social Media, 2009.

[38] P. Melville, R.J. Mooney, Diverse ensembles for active learning, in: Proceedings of the 21th International Conference on Machine learning, 2004, pp. 74.

[39] A. Mittal, A. Goel, Stock Prediction Using Twitter Sentiment Analysis, 2012, Stanford.edu <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf> (accessed 12.08.13).

[40] G. Mishne, N. Glance, Predicting movie sales from blogger sentiment, in: AAAI Symposium on Computational Approaches to Analysing Weblogs AAAI-CAAW, 2006, pp. 155–158.

[41] R. Morris, Branding OHIO Through Social Media, 2012 <http://www.ohio.edu/compass/stories/11-12/8/branding-ohio-social-media.cfm> (accessed 01.02.13).

[42] S. Nann, J. Krauss, D. Schoder, Predictive analytics on public data – the case of stock markets, in: Proceeding of 21st European Conference on Information Systems, 2013 (Paper 116).

[43] P.B. Nemenyi, Distribution-Free Multiple Comparisons, Doctoral dissertation, Princeton University, 1963.

[44] J.R. Nofsinger, Social mood and financial economics, J. Behav. Finan. 6 (3) (2005) 144–160.

[45] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2010), 2010, pp. 122–129.

[46] C. Oh, O.R.L. Sheng, Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement, in: Proceedings of the International Conference on Information Systems, ICIS, 2011.

[47] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, 2002, pp. 79–86.

[48] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Ret. 2 (1–2) (2008) 1–135.

[49] Y. Rao, Q. Li, X. Mao, L. Wenyin, Sentiment topic models for social emotion mining, Inform. Sci. (2014).

[50] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, in: Proceedings of the ACL Student Research Workshop, 2005, pp. 43–48.

[51] J. Regnault, Calcul des Chances et philosophie de la Bourse, Mallet-Bachelier and Castel, Paris, 1863.

[52] J.A. Rice, Mathematical Statistics and Data Analysis, 3rd ed., Duxbury Advanced, 2006.

[53] E.J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, A. Jaimes, Correlating financial time series with micro-blogging activity, In: Proceedings of the fifth ACM International Conference on Web Search and Data Mining, 2012, pp. 513–522.

[54] M. Saveski, M. Grčar, Web services for stream mining: a stream-based active learning use case, in: Proceedings of the PlanSoKD Workshop at ECML PKDD, 2011.

[55] M.J. Schervish, P values: what they are and what they are not, Am. Statist. 50 (3) (1996) 203–206.

[56] D. Sculley, Online active learning methods for fast label-efficient spam filtering, in: Fourth Conference on Email and AntiSpam (CEAS), 2007.

[57] D. Sculley, Combined regression and ranking, in: Proceedings of the 16th Annual SIGKDD Conference on Knowledge Discover and Data Mining, 2010, pp. 979–988.

[58] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. (CSUR) 34 (1) (2002) 1–47.

[59] V. Sehgal, S. Charles, SOPS: stock prediction using web sentiment, in: Proceedings of the International Conference on Data Mining Workshops, IEEE ICDMW, IEEE, 2007, pp. 21–26.

[60] A. Seth, Granger Causality, Scholarpedia 2(7) (2007) 1667 <http://www.scholarpedia.org/article/Granger_causality> (accessed 27.08.13).

[61] B. Settles, M. Craven, An analysis of active learning strategies for sequence labeling tasks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 1070–1079.

[62] B. Settles, M. Craven, L. Friedland, Active learning with real annotation costs, in: Proceedings of the NIPS Workshop on Cost-Sensitive Learning, 2008, pp. 1–10.

[63] B. Settles, Active Learning Literature Survey, University of Wisconsin, Madison, 2010.

[64] B. Settles, Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1467–1478.

[65] B. Settles, From theories to queries: active learning in practice, in: Workshop on Active Learning and Experimental Design, 2011, pp. 1–18.
[66] H. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 287-294.
[67] S. Shalev-Shwartz, Y. Singer, N. Srebro, Pegasos: primal estimated sub-gradient solver for SVM, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 807–814.
[68] J. Smailović, M. Grčar, M. Žnidaršič, Sentiment analysis on tweets in a financial domain, in: 4th Jožef Stefan International Postgraduate School Students Conference, 2012, pp. 169–175.
[69] J. Smailović, M. Grčar, M. Žnidaršič, N. Lavraš, Predictive sentiment analysis of tweets: a stock market application, in: Human–Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Lecture Notes in Computer Science, vol. 7947, 2013, pp. 77–88.
[70] T.O. Sprenger, A. Tumasjan, P.G. Sandner, I.M. Welpe, Tweets and trades: the information content of stock microblogs, Euro. Finan. Manage. (2013).
[71] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment in Twitter events, J. Am. Soc. Inform. Sci. Technol. 62 (2) (2011) 406–418.
[72] R.M. Tong, An operational system for detecting and tracking opinions in on-line discussion, in: Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification (OTC), 2001, p. 6.
[73] S. Tong, E. Chang, Support vector machine active learning for image retrieval, in: Proceedings of the ACM International Conference on Multimedia, 2001, pp. 107–118.
[74] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, in: Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010), 2010, pp. 178–185.
[75] P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the Association for Computational Linguistics, 2002, pp. 417–424.
[76] C.J. van Rijsbergen, Foundation of evaluation, J. Document. 30 (4) (1974) 365–373.
[77] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[78] V. Vapnik, Statistical Learning Theory, John Wiley and Sons Inc., New York, 1998.
[79] P. Wang, P. Zhang, L. Guo, Mining multi-label data streams using ensemble-based active learning, in: Proceedings of SDM, 2012, pp. 1131–1140.
[80] P. Wessa, Bivariate Granger Causality (v1.0.0) in Free Statistics Software (v1.1.23-r7), 2008, Office for Research Development and Education <http://www.wessa.net/rwasp_grangercausality.wasp/> (accessed 01.02.13).
[81] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[82] Y. Yang, X. Liu, A re-examination of text categorization methods, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 42–49.
[83] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Dec. Supp. Syst. 55.4 (2013) 919–926.
[84] P. Zhang, X. Zhu, L. Guo, Mining data streams with labeled and unlabeled training examples, in: Proceedings of the International Conference on Data Mining ICDM'09, 2009, pp. 627–636.
[85] X. Zhang, H. Fuehres, P.A. Gloorm, Predicting stock market indicators through twitter I hope it is not as bad as I fear, Proc.-Soc. Behav. Sci. 26 (2011) 55–62.
[86] X. Zhu, P. Zhang, X. Lin, Y. Shi, Active learning from data streams, in: Proceedings of the International Conference on Data Mining ICDM'07, 2007, pp. 757–762.
[87] X. Zhu, P. Zhang, X. Lin, Y. Shi, Active learning from stream data using optimal weight classifier ensemble, IEEE Trans. Syst, Man, Cybernet.: Part B 40 (6) (2010) 1607–1621.
[88] I. Žliobaitė, A. Bifet, B. Pfahringer, G. Holmes, Active learning with evolving streaming data, in: Machine Learning and Knowledge Discovery in Databases, Springer, Berlin Heidelberg, 2011, pp. 597–612.