Correlation between Stock Prices and polarity of companies’ performance in Tweets: a CRF-based Approach

Ekaterina Shabunina
Università degli Studi di Milano-Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione
ekaterina.shabunina@disco.unimib.it

Abstract

Information produced and shared on social networks constitute a valuable source for inferring trends and events in the real world. In this paper we show how this can be exploited concretely through quantitative analysis of social content. We present an analysis of the statistical correlation between the security prices of some IT companies and the performance polarity of the same companies as expressed in tweets. Companies’ performance polarities are obtained by applying Conditional Random Fields to the considered streams of tweets. An evaluation of both the classification model and the performed regression analysis is also presented.

I. INTRODUCTION

The analysis of the information in microblogs like Twitter is currently receiving an increasing attention from analysts and researchers. The growth in popularity and adoption of social media like Twitter offers an unmatched and unprecedented source of data for opinion and fact mining. The current research challenge is to exploit such data and relate them to real world events, to the purpose of predicting/estimating similar events or trends.

This work addresses the issue of the existence of correlation between the polarity of companies performance as expressed in tweets and the variation of these companies stock prices. To address this issue, we first have collected tweets (over a given period) containing companies ticker tags, or “cashtag” as they are referred to in Twitter. We have then classified the extracted tweets as negative, neutral, or positive, based on the information they contain about the performance of the companies. Finally we have studied the correlation between the various measures of performance polarity and the total tweets volume against the actual stock closing prices and the traded volumes of company securities.

Intuitively, one can assume that such correlation would be high, due to the fact that investors, traders and (most of) the other players in the financial markets are human beings. As such, they are indeed affected by different information, rumors and events along with their choice of actions on the stock market, and they tend to share them (once sharing was with close-by people, nowadays its on social networks). There might even be intentional usage of social networks by investors and traders, who use Twitter as a platform to communicate, share their knowledge and even try to influence the masses to perform trading actions in their favor.

Recently some research\cite{1,4,5} have addressed the problem of correlating stock prices to social network sentiments. The approach proposed in this paper is different: in fact, we do not exploit the polarity of the opinions in the tweets, but rather the polarity of the facts about the company’s performance as expressed by the Twitter users. For example, a tweet of a user declaring she is happy that Google Inc. stock price is going down, would be labeled as “negative” due to the negative performance of the company’s stock rather than the “positive” sentiment of the micro-blogger. We rely on the Conditional Random Fields probabilistic model over manually labeled datasets for multiclass classification of the companies’ performance. We achieved a classification accuracy of 93% for the 8 companies in the experiment. This is especially interesting if we consider that the task of polarity classification of tweets is very complex (even for a human being). Moreover, companies’ stock related tweets are part of a special financial domain, which employs a very specific set of jargons and slangs, with particular abbreviations and symbols. Last but not least, company-related tweets tend to be even shorter than the 140 symbols limit for a Twitter micro-blog, often stating only the action on the stock market that the investor has performed or is going to perform. An example of such an intricate and implicit semantics is the following tweet: short $GOOG. As one can see, it contains very little information for a classifier to build an accurate classification model.

Nevertheless, the performed regression analyses have proven that the total volume of tweets has a stronger statistically significant relationship with the stock measures than the performance polarity measures. This can be seen as the proof of the common saying that there’s no such a things as bad publicity.

The remainder of the paper is structured as follows: Section II introduces relevant background concepts; Section III presents our methodology; Section IV and V describe the experimental scenario and results, respectively; Section VI concludes and outlines future work directions.

II. BACKGROUND

In this Section the background notions useful to understand the study reported in this paper are pre-
sent.

I. Conditional Random Fields

For the task of classification we adopted in our work Conditional Random Fields (CRF) [3], a framework for building probabilistic models to segment and label sequence data. Due to its sequential nature, the CRF classifier performs better than the common bag-of-words approaches, especially for short texts, where bag-of-words approaches usually fail, due to the sparseness of the resulting feature vector.

CRFs have a similar structure to the Conditional Markov Model (CMM), and consequently share the same benefits of the CMM over generative models such as Hidden Markov models (HMM), but instead of using a directed graph as CMM, CRFs use an undirected graph. CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. Therefore, the label bias problem does not arise for CRFs because the weights of different states can be traded off against each other, which can lead to accuracy improvements. The results presented in [3] demonstrate that even when the models are parameterized exactly in the same way, CRFs are more robust to inaccurate modeling assumptions than MEMMs or HMMs, and resolve the label bias problem, which affects the performance of MEMMs. This is the main reason we thought that CRF could be a successful technique in the stock market field.

II. Stock Market

Despite the variety of analyses that are performed by the players in the financial markets in order to determine whether to buy or sell a given security (and at which prices to do so), they can be generally divided into two main categories: technical and fundamentalist. While the first one is basically an attempt of applying mathematical models to understand the behaviour of a given security and try to forecast its future movements, the second one is based on the study of intrinsic value of the company whose shares are under consideration. The value of a company is based on its capacity of generating cash in the future. Buying a company share, in this sense, is the same as buying an expected future cash flow.

The relationship between the general ascertained performance polarity in tweets related to a given company, in this sense, may be a good proxy of its future profitability. Therefore, the quantitative performance polarity analysis of these tweets may be a good indicator of the future profitability of a company, in a way that it could be correlated to its stock performance. This relationship has been put to test in this work.

III. Methodology

Figure 1 shows the architectural overview of the proposed approach. The whole process is divided into five main parts: Data Pre-processing, Data processing, Training, Twitter data labeling, and Regression analysis of tweets against stock market variables.

I. Data Pre-processing

First of all, we have crawled the tweets to be analyzed. This has been obtained through the Twitter Search API by the cashtags of the companies considered for the analysis. A cashtag consists of the company’s ticker code with a “$” sign before it, for example, $msft for Microsoft Corp.

For the Twitter raw data filtering process numerous text processing steps have been applied: multiple lines breaks elimination; e.g., filtering out of all the non-English data, using an off-the-shelf language detection technique [2] etc. Then, the original tweet timestamps were normalized to the American Eastern Time zone to be comparable to the NASDAQ stock exchange time and grouped in reference to the trading hours of the Nasdaq Stock Exchange. This is based on the assumption that the tweets posted after the closing time of the stock market (16:00) have to be correlated to the next days stock market performance, on Friday after 16:00 as well as within the weekends chatter are assumed to have a reflection on the following Monday.

II. Data Processing

The tweets to be used for the training of the classifier (first 20% of the collected data per each company, reported in Table 1) were manually labeled at tweet level in the sense of the company’s performance, as either positive, negative or neutral, where neutral stands for the tweets with no specific indicator of the company’s performance or those with a mixed one.

<table>
<thead>
<tr>
<th>Company</th>
<th>Total</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
<th>Manually labelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>2,589</td>
<td>366</td>
<td>21</td>
<td>2,202</td>
<td>250</td>
</tr>
<tr>
<td>ORCL</td>
<td>2,642</td>
<td>611</td>
<td>101</td>
<td>1,930</td>
<td>405</td>
</tr>
<tr>
<td>GOOG</td>
<td>31,529</td>
<td>7,184</td>
<td>3,216</td>
<td>21,128</td>
<td>9,177</td>
</tr>
<tr>
<td>MSFT</td>
<td>16,646</td>
<td>4,433</td>
<td>350</td>
<td>11,663</td>
<td>3,516</td>
</tr>
<tr>
<td>ADBE</td>
<td>1,399</td>
<td>307</td>
<td>102</td>
<td>990</td>
<td>265</td>
</tr>
<tr>
<td>INFY</td>
<td>383</td>
<td>100</td>
<td>3</td>
<td>280</td>
<td>71</td>
</tr>
<tr>
<td>LOGI</td>
<td>261</td>
<td>147</td>
<td>1</td>
<td>113</td>
<td>63</td>
</tr>
<tr>
<td>YHOO</td>
<td>9,606</td>
<td>2,821</td>
<td>886</td>
<td>5,898</td>
<td>2,191</td>
</tr>
</tbody>
</table>

Table 1: Companies Twitter dataset statistics

Each tweet has been tokenized at word level and Part-of-Speech tagged with the java CMU ARK Twitter POS Tagger.

The tweets have then been represented as an array of strings (both the training and the testing data) as required by the CRF.

III. Model Training, Data Labeling, and Regression Analysis

In our work we have chosen to use a general-purpose tool CRF++ a simple, customizable, and open source implementation of Conditional Random Fields for segmenting and labeling sequential data.

CRF++ has been trained with the manually labeled corpus of tweets in which each instance (tweet) is a sequence of words, labeled according to the class of the instance to which they belong.

To obtain an as much unbiased estimate of the performance of the system as possible, the classification model has been trained using 10-fold cross-validation. We have performed numerous experiments with bigram templates, varying different combinations of word and POS tag features in the search for the best classifiers performance.

The Twitter data labeling phase has consisted of automatically labeling all the data that has not been manually labeled (80% of the collected data per each company), with the created classification model.

For the statistical analysis of the regression, we have employed the software Minitab 17.

IV. EXPERIMENTAL SETTING

For the experiments we selected 8 companies in the IT sector from the Nasdaq-100, in particular those that showcased a diverse Twitter activity distribution. They are Google Inc. (GOOG), Microsoft Corp. (MSFT), Yahoo! Inc. (YHOO), Electronic Arts Inc. (EA), Adobe Systems Inc. (ADBE), Infosys Technologies (INFY), Logitech International S.A. (LOGI), and Oracle Corp. (ORCL).

For each company two datasets have been created. The first one has consisted of tweets about the company’s stock market performance and has been constructed by crawling Twitter only based on the company’s cashtag. The second dataset contained the company’s stock closing prices and negotiated volumes for the same time period as of the Twitter dataset, and has been created using Bloomberg.

The experiment datasets enclose tweets over a 5 weeks time span, from 17th of April, 2014 until 24th of May, 2014. The first week (17th of April until 24th of April, 2014) has been manually labeled and used as the training data for the classifier. Table 1 shows the numerical statistics related to these datasets. One important note: as clearly observable from Table 1, the number of negative tweets is remarkably lower than of positive tweets for all 8 companies studied.

V. EXPERIMENTAL RESULTS

In this Section we present the evaluation of the classification model and the results of the statistical correlation analysis.

I. Classifier Performance

We have conducted several experiments with different templates, by varying the involved features. The best result has been achieved with a template which included: two previous words, current and two next words, their POS tags and these words combinations: word before previous word / previous word, previous/current, current/next, and next/next after next words.

Table 2 presents the best achieved performance measures of the classification models for each of the 8 companies studied, selected out of the 10-folds.

The average accuracy over all companies is 93% and for some companies it has even exceeded 94% (Logitech Systems Inc.: 94.4%, Google Inc.: 94.1%). While evaluating these results, it is necessary to consider, that since the experiments were conducted on real Twitter data: a) the training sets for all companies were non-symmetric; b) the training sets for some companies were rather numerically limited; c) the sets have been manually labeled with inherent subjectivity; d) the study has been performed over

---

3http://www.ark.cs.cmu.edu/TweetNLP/
5http://www.minitab.com/enus/products/minitab/
6http://www.bloomberg.com/
the rather complex financial markets domain; e) the datasets consisted of short by nature Twitter messages and; f) as features the analysis has only used word combinations and Part-of-Speech tags.

It is relevant to note that, as mentioned in the Section IV, people tend to tweet mostly positive comments about the companies stock performance instead of negative ones, therefore the difference between the number of positive and negative tweets is very large. Due to this finding, the performance measures for the negative tweets are not very coherent to those achieved for the positive and neutral classes. For example, referring to Table 2, Logitech (logi) had only 1 negative tweet out of 258 for the whole 5 weeks period considered and, similarly, Infosys (infy) had 3 negative tweets out of 383. For these two companies, for the negative class the accuracy is ideal, while the recall and precision are consequently 0.

II. Visual correlation

Before starting to discuss the results, it is important to stress that correlation does not imply causality. This means that, for instance, observing a high number of positive tweets may not be (and is very likely not) the reason why the stock performance for the studied day was positive. On the contrary, it may simply mean that the stock variation caused a positive reaction on the Twitter users, which on their turns posted positive content about the company.

The aim of this preliminary study is to visually identify whether a similar trend is present in the number of classified and total amount of tweets with the movements of the stocks closing prices and negotiated volumes.

In this Section, we firstly would like to understand if the daily variation on the stocks’ closing prices presented itself as a better variable than the absolute closing price (in USD) to be related to the results obtained in terms of the difference in between the number of positively and negatively classified tweets per each trading day. Secondly, we present the comparison of the total number of tweets, disregarding the classification, as a correlation variable to two measures absolute closing price change (the module of the % change in a given day) and the total negotiated volume. The second analysis was to infer if the closing prices movement or the total traded volume presented better adherence measures to the total chatter on Twitter about the stocks of companies. An additional overall goal pursued by this study was the comparison of the polarity identified in the tweets and the total volume of tweets as the correlation measures to the stocks’ movements.

As pointed out in the Section VII, people tend to tweet rather less when the company has a negative situation. Due to this finding, it became impossible to use a logical measure for the analysis of the stock price accumulated inertia, which was infinitely increasing because of a recurrent higher number of positive tweets in the datasets.

In this Section, although the charts have been plotted for all companies studied, Oracle Inc. is used as the demonstrative example.

First we looked at the two charts per each company: Pos-Neg (number of positive tweets minus number of negative tweets per day) and the closing prices (Figure 2a), against Pos-Neg and the change in the closing prices (Figure 2b). As becomes clearly observable, only sporadically occurs to take place a slight correlation in the first figure, while on the second figure it is explicitly visible that the plots follow similar patterns. This finding proves the necessity of the price change measure in assessing the adherences between stocks closing prices and the social indicators of the companies performances in tweets.

The second visual comparison of the measures is between the total number of tweets (independently from classification) in a given day and the absolute (module) price change (Figure 2c), against the total number of tweets and the total negotiated volume for the given security (Figure 2d). Even if this comparison does not imply anything about the quality of the classification procedure, it is important for the understanding of the relationships between social networking activities and real world phenomena. As observable, in both figures the visual correlation is strongly present.

<table>
<thead>
<tr>
<th>Company</th>
<th>Total Acc</th>
<th>Total Prec</th>
<th>Total Recall</th>
<th>F</th>
<th>Positive Acc</th>
<th>Positive Prec</th>
<th>Positive Recall</th>
<th>F</th>
<th>Negative Acc</th>
<th>Negative Prec</th>
<th>Negative Recall</th>
<th>F</th>
<th>Neutral Acc</th>
<th>Neutral Prec</th>
<th>Neutral Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>0.93</td>
<td>0.95</td>
<td>0.67</td>
<td>0.79</td>
<td>0.94</td>
<td>0.97</td>
<td>0.85</td>
<td>0.90</td>
<td>0.95</td>
<td>0.17</td>
<td>0.29</td>
<td>0.91</td>
<td>0.87</td>
<td>1</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>ORCL</td>
<td>0.93</td>
<td>0.94</td>
<td>0.78</td>
<td>0.85</td>
<td>0.94</td>
<td>0.97</td>
<td>0.84</td>
<td>0.90</td>
<td>0.95</td>
<td>0.50</td>
<td>0.67</td>
<td>0.91</td>
<td>0.86</td>
<td>1</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>GOOG</td>
<td>0.94</td>
<td>0.90</td>
<td>0.88</td>
<td>0.95</td>
<td>0.88</td>
<td>0.84</td>
<td>0.84</td>
<td>0.86</td>
<td>0.95</td>
<td>0.84</td>
<td>0.86</td>
<td>0.92</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>MSFT</td>
<td>0.92</td>
<td>0.88</td>
<td>0.81</td>
<td>0.84</td>
<td>0.90</td>
<td>0.88</td>
<td>0.85</td>
<td>0.86</td>
<td>0.96</td>
<td>0.88</td>
<td>0.63</td>
<td>0.74</td>
<td>0.90</td>
<td>0.89</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>ADBE</td>
<td>0.92</td>
<td>0.92</td>
<td>0.76</td>
<td>0.83</td>
<td>0.94</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.93</td>
<td>0.40</td>
<td>0.57</td>
<td>0.90</td>
<td>0.86</td>
<td>0.97</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>INFY</td>
<td>0.89</td>
<td>0.57</td>
<td>0.48</td>
<td>0.52</td>
<td>0.86</td>
<td>0.88</td>
<td>0.47</td>
<td>0.61</td>
<td>0.98</td>
<td>0.00</td>
<td>0.00</td>
<td>0.84</td>
<td>0.84</td>
<td>0.98</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>LOGI</td>
<td>0.94</td>
<td>0.68</td>
<td>0.63</td>
<td>0.61</td>
<td>0.92</td>
<td>1</td>
<td>0.89</td>
<td>0.94</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.92</td>
<td>0.75</td>
<td>1</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>YHOO</td>
<td>0.94</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.91</td>
<td>0.88</td>
<td>0.79</td>
<td>0.83</td>
<td>0.98</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.92</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Classifiers Performance for all companies at study. Acc: Accuracy, Precis: Precision, Recall: Recall, F: F measure.
III. Regression Analysis

The aim of this section is to go beyond the simple visual analysis of the plotted data, with the purpose of providing statistical certainty on the inferred conclusions. To formally validate the adherence of the classification results with the real world observations, a set of statistical regression analyses have been performed, using the same pairs of variables reported in Section VII.

In order to verify the statistical significance of the relationship between two variables the ANOVA (ANalysis Of VAriance) methodology is applied to the linear and quadratic regression. The test examines the statistical significance (p-value) of the relationship between the two variables at a given threshold of 5% ($p<0.05$). It presents the explanatory capacity of the regression model (R-sq) and the correlation (r) between two variables.

For the comparison of the first two pairs, similarly to what was observable in the figures 2, the closing price change seems to be a better correlation measure to the difference between the number of positive and negative tweets about the companies’ performance. We achieved only one case of $p=0.049$, on the edge of the given statistical significance threshold ($p<0.05$), for the daily performance polarity difference and closing prices. For the performance polarity difference and the price change 3 out of 8 companies had a statistically significant relationship: two of which with an ideal $p<0.001$, and one even with a strong regression models explanatory capacity of over 90% and 0.85 correlation between these two variables.

Using the total number of tweets as the independent variable and as the dependent variables both the absolute (modular) price change and the total traded volume, similar results were obtained as already noted in the Section VII. 5 out of 8 relationships were statistically significant for both cases. For these statistically significant cases the average explanatory capacity was: 36% for the absolute price change case and 39% for the total traded volume case. It is worth noticing that for Electronic Arts Inc. has been observed an almost perfect R-sq of 95.43% in the quadratic regression model for the absolute price change case; and an 82.28% respectively for the traded volume case.

IV. Granger Causality

The Granger Causality test [2] is widely used in determining if one time-series $X(t)$ is useful in forecasting another time-series $Y(t)$. Since the regression analysis shows only the presence of correlation between the variables and not the causality of the relationship, the Granger tests were applied in the two possible directions for each variable at study of all the 8 companies considered for the experiment.

The obtained results for this test are rather random and did not enable us to make coherent conclusions. The cases with a causal relationship were mostly not in accordance with the regression analysis results. Though in most cases in which there
is the presence of Granger causality the stocks are the cause of the tweets generation and not the other way around, as predicted. In general, only half of the companies had at least one positive Granger Causality relationship. Only for Microsoft Inc. all cases but one (total number of tweets V.S. the traded volume) have a causality relationship. But always the stock markets variable appears as the forecaster for the tweets measure.

VI. CONCLUSIONS AND FUTURE WORK

The polarity classification method presented in this paper achieved an average accuracy of 93%, using as features only word combinations and Part-of-Speech tags.

The observation referring to the strongly lower amount of negative tweets for all 8 companies compared to positive or negative tweets brings to a curious conclusion: Twitter users in general tend either to post more actively tweets about the positive companies’ performance or not to post much when the company has a negative event or price movement.

The explicitly visible correlation was achieved for all companies in three cases: the polarity difference and the closing prices change, the total number of tweets and the absolute by module price change, and the total number of tweets and the total negotiated volume for the given security. The regression analysis results presented 5 out of 8 statistically significant relationships for both pairs with the total tweets volume as independent variable; and 3 out of 8 for the polarity difference and price change. The polarity difference and the stock’s closing price neither visually nor quantitatively present valuable relationship. Therefore, it is possible to make a strong conclusion that the performance polarity in tweets has less information then the general volume of tweets posted, making possibly unnecessary to undergo the time consuming and demanding task of tweets classification.

The Granger causality test showed that the Twitter chatter about the stocks of companies at study did not have a forecasting capacity of the stocks prices or volumes.

Although the experimental results do not allow to extract general conclusions for the whole stock market, due to the limitation in time and number of companies analyzed, the achieved observations referring to the accuracy of the classification models and the found adherences of the polarity tweets with the stock market values show that, nevertheless, the proposed approach is promising and has a strong potential of adoption, especially if properly combined with other stock market analysis tools.

There are many areas in which this work could be expanded with the objective of further refinements in the methodology and of being able to grasp more generic and embracing conclusions.

First, to increase classifiers performance measures, one initial direction would be to perform the research study for a longer period of time and expand the manually labeled data sample. This is crucial, considering the complexity of classifying small information nuggets like tweets. To reduce the consumption of time of experts and also the subjectivity of the data labels, the crowdsourcing approach could be adopted. As for the model refinement, we expect to obtain relevant classification model improvements by relying on the introduction of more complex features, rather than using solely combinations of words and POS tags.

Finally, it is also planned to test the presented approach on a wider range of companies and perhaps on different knowledge or even language domains. For instances a similar analysis could be applied to the general stock market evolution instead of single companies.

VII. ACKNOWLEDGEMENTS

This research was partially supported by the ERC grant Search Computing from Politecnico di Milano.

REFERENCES


