Unlock the Stock: User Topic Modeling for Stock Market Analysis

Patrick Siehndel and Ujwal Gadiraju
L3S Research Center, Leibniz Universität Hannover, Germany
{siehndel,gadiraju}@L3S.de

ABSTRACT

The increasing use of Twitter as a medium for sharing news related to various topics, facilitates methods for automatic news creation or event detection and prediction. However, these methods are hindered by users posting and propagating incorrect or irrelevant content. Choosing the right users is crucial in order to sample down the tweets to be analyzed, and preserve the quality of the predicted events or generated news. In this paper, we present an effective method for identifying *expert users* in defined areas related to the stock market. For each user we generate a model based on the content of their posts. The model represents the domains the user talks about, and allows a selection of users for various tasks. We show the effectiveness of the proposed approach by performing a series of experiments using large Twitter datasets related to Stock Market Companies.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

1. INTRODUCTION

Today a large amount of data is user generated content produced within social media networks like Twitter or Facebook. The availability and abundance of this kind of data has lead to various innovative real world applications. Harnessing sentiments and opinions that characterize political landscapes, using geolocation of users for analyzing earthquakes [15], or investigating social networks for predicting disease outbreaks and spread [7] are a few examples that rely on the analysis of social media data.

One of the major challenges when working with social media data, is the fact that users in the network are not homogeneous. While for some applications this might be very useful (for example, to access multiple standpoints on subjective matters from different groups of persons), in some others it is necessary to only consider the posts from certain user groups. In this paper, we present an approach towards identifying groups of users that are of particular interest in the realm of stock market analysis. Our method is based on the use of background information from Wikipedia, allowing us to generate profiles which take into account the semantic background of messages posted by the user. Additionally, our method allows to create connections between the user profiles and different areas of the stock market represented by a hierarchical model. For generating the user models we consider the textual content of the messages written by the users as well as the available metadata for each user. By applying named entity recognition on the provided posts we generate semantically enriched messages. These are aggregated into a user profile, representing the topics and fields a user writes about as well as an estimation of the user's expertise in the field. In addition, these profiles also represent how trustworthy the messages of a user in certain domains are.

The usefulness of the detection of experts in social networks related to stock markets has been analyzed by Bar-Haim et al. [3]. The authors show that training models to distinguish expert and non-expert users can improve the quality of predicted stock market changes. Our work here, is motivated further by this premise.

We test and evaluate the proposed method on a large dataset consisting of tweets related to the stock market. Our experiments show that the right composition of users can help to increase the quality and effectiveness of event prediction algorithms as well as the overall analysis of messages related to certain topics. We also show that this reduces the number of messages that need to be analyzed.

The remainder of this paper is structured as follows. In Section 2 we review the related work in the areas of topic modeling and user expertise analysis. In Section 3 we describe in detail the manifestation of complete user profiles from simple textual messages. In the penultimate Section 4, we describe a series of experiments that test the effectiveness of the proposed methods. In the last Section we draw conclusions, discuss the implications of our work and present an outlook of our imminent future work.

2. RELATED WORK

The related work in our area is split into two different directions: (i) the area of expert detection, and (ii) the domain of user and topic modeling in Twitter.

2.1 Expert Detection

In the area of expert detection the work conducted by Guy et al. [9] describes a scenario where users within an enterprise network are analyzed. The scores assigned to users representing their interest and expertise are based on search terms and documents within an index. In contrast to our work, the connection between users and

areas of expertise is made by keyword matching. In our work, the semantic relation between entities related to the topic of interest and the entities mentioned by the users is used for generating the relations. Our focus on expertise networks in Twitter also fits in the research area of automated expert finding. Here both explicit and implicit information is used for identifying experts in a particular area. Yimam-Seid and Kobsa [18] argue that for an effective use of knowledge within an organization, it is important to use hidden knowledge in various forms. The authors separate the need for information (need for people who can provide advice, help or feedback) from the need for expertise (the need for people who can perform a social or organizational role). Ghosh et al. [8] used Twitter content for seeking experts on a topic. Their results show that the use of Twitter Lists describes the expertise of a user more accurately than systems that rely on merely tweet content.

2.2 User and Topic Modeling in Twitter

Abel et al. [1] compared different approaches for extracting professional interests from social media profiles. They showed that tag based profiles and self-created user profiles are most suitable for this task. Recent work [4] has shown how user trust models can be used for increasing the performance of event detection methods on Twitter by using textual features and meta-information about the user. Based on the user profile a classifier decides whether to take messages from these users into account or discard them. This application is relatively close to the one analyzed in this paper. The main differences lie in the use of word vectors (wherein we use detected entities in contrast) and the focus on the trustworthiness of a user, while we focus on areas of interest and keep users based on the amount of related information they posted in the past.

Besides the analysis of textual patterns and user information several works also consider the structural information inherent in social networks like Twitter. The task of finding content of high quality has also been analyzed in domains like question answering communities [2] or forums [19] which are comparable in several dimensions as they are also social networks. The use of standard authority estimation approaches within the network has also been used to gather experts for specific topics [17]. In [5], Chelaru et al. analyzed how different user groups are connected to each other and showed that within professional networks groups centered around different skills exist. For our analysis these groups are interesting since the user in these groups can be considered to be experts for the specific domain.

3. METHODOLOGY

In this Section we describe how the creation of user sets for defined topics is carried out. We select special groups of users with the main goal of reducing the content which is not of interest for our domain; the main topics with relevance for Systemic Risk and Stock market analysis are only a small fraction of the messages posted by most of the users. Additionally many users have different purposes in mind when using social media [12]. Nevertheless Twitter contains a large portion of news related content [11] which is of interest for generating topic specific user models within the analyzed domain.

Our user model describes the topics a user writes about based on the Thomas Reuters Business Classification System (TRBC)[1]. This System contains a hierarchy for different business sectors, allowing us to model the interests and expertise of the monitored users in a way, which directly describes and partitions the different sectors of the stock market. In our profile we model the Economic sectors, Business sectors and Industries to corresponding Wikipedia categories. These different categories contain various aspects of the stock market in different granularities. For instance, the Economic Sector *Basic Material* is split into Business Sectors like *Chemicals* and *Mineral Resources*. These are further grouped into industries like *Agricultural Chemicals* or *Steel*. Our aggregated user profiles together with the different levels of granularity allows us to find and monitor users for various domains.

In our scenario we are looking for expert users within specific domains. An *expert user* can be described as a person who has a deeper knowledge regarding a certain domain than the average user. In particular we choose the users related to a domain by first annotating the messages posted by the user, and then calculating their relevance to the defined domain of interest. The models and computed connections are comparable to the methods presented in [16] and [10], albeit without the need for performing graph walking or traversal algorithms. Our approach for the creation of a user profile consists of the following 4 main steps.

- Topic selection
- Message enrichment
- Entity linking
- User finalization

3.1 Creating User Profiles

Topic selection The Thomson Reuters Business Classification System builds the base for the topics of the generated profiles, it contains a hierarchy for different business sectors, allowing us to model the interests and expertise of the monitored users in a way that indicates their relatedness to companies of the different stock market sectors. We linked the Economic sectors, Business sectors and Industries of the TRBC to corresponding Wikipedia categories. Based on these Wikipedia categories the users interests are described. As a result, our profile only indicates the relatedness of the user to different sectors of the stock market. These user profiles together with the different levels of granularity allows us to find and monitor users for various scenarios.

Message enrichment In order to relate the messages written by the analyzed users we annotate all tweets of the user using the Wikipedia Miner Toolkit[2] [13]. We use named entity recognition, which provides us relations between entities or concepts mentioned in the tweets of a user to the corresponding Wikipedia article. The links discovered by Wikipedia Miner have a similar style to the links which can be found inside a Wikipedia article. Not all words which have a related article in Wikipedia are used as links, but only words which are relevant for the whole topic are used as links. Figure 1 shows an example tweet with the related Wikipedia articles which are used to build the user profile.

Entity linking: The third stage, entity linking, relates the entities that have been mentioned in the users tweets to the categories chosen beforehand representing the different sectors of the stock market. For each of the entities we calculate the relatedness to every article belonging to the chosen categories. As relatedness measure we use the one proposed by Milne et al. [14], this measure is modeled based on the normalized Google distance measure [6] and

[1]http://thomsonreuters.com/en/products-services/financial/market-indices/business-classification.html

[2]http://wikipedia-miner.cms.waikato.ac.nz/

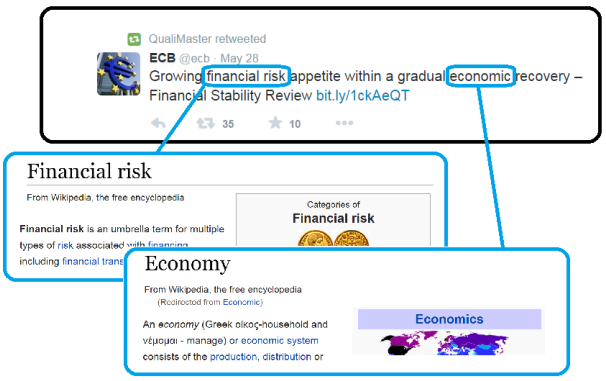

Figure 1: Example of an annotated Tweet with the related entites within Wikipedia.

calculates the ratio between the inlinks two articles have in common and the overall articles linking to them. In order to reduce the influence of articles being mentioned several times we used the log of the number of articles as a weight for the profiles. The formula for the relatedness of two articles() is defined as shown below.

$$R(a,b) = \frac{\log(max(|A|,|B|)) - \log(|AB|)}{\log(|W|) - \log(min(|A|,|B|))}$$

User finalization: In the final stage, we perform an aggregation over all of the tweets and related entities of a user in order to generate the final user profile. The generated user profile displays the topics a user talks about, based on the amount of followers or how focused a user is in a certain topic. We also get an estimation about the expertise of the user in this domain.

Algorithm 1 depicts the steps in the creation of a user profile. In line 10 we perform the main update of the profile using the related category and the calculated relatedness.

The several loops in the algorithm require some time for calculation, to speed up the process our implementation contains a buffer storing the weights for all categories for every entity. This reduces the three inner loops to a single lookup in a hashmap, making the method applicable for very large sets of users and tweets, making our approach presented here scalable.

4. EXPERIMENTS FOR USER MODELING

In this section we describe the experimental evaluation performed for analyzing our generated user models. The experimental evaluation is a two-fold approach. In a first set of experiments we evaluate the generated user models as they are, without correlating them with additional information. This set of experiments gives us an overview of how the analyzed users behave in general and how their topics of interest influence their position within the network. We also evaluate the temporal stability of the generated profiles in order to give recommendations for future use of the proposed methods. In the second set of experiments, we evaluate the usefulness of the generated profiles in a series of experiments, for applications related to event detection. We show how the generated profiles influence the results.

4.1 Evaluating the Users

Algorithm 1: Update of User Profiles

Input: T: Set of Tweets of a User C: Set of Categories for the Profile

1 **foreach** $c_k \in C$ **do**
2 $\quad ILC \leftarrow (c_k, getInlinks(c_k))$
3 **foreach** $T_i \in T$ **do**
4 $\quad A_i \leftarrow getAnnotations(T_i)$
5 $\quad$ **foreach** $a_j \in A_i$ **do**
6 $\quad\quad ILA_i \leftarrow getInlinks(a_j)$
7 $\quad\quad$ **foreach** $ilak \in ILA_i$ **do**
8 $\quad\quad\quad$ **foreach** $ilc_k \in ILC$ **do**
9 $\quad\quad\quad\quad R \leftarrow calulateRelatedness(ilc_k, a_j)$
10 $\quad\quad\quad\quad updateProfile(ilc_k, R)$

Table 1: Statistics about Users in Stock related dataset.

Unique Users	Avg Tweets per User	Avg Friends per User	Avg Follower per User	Avg Listed per User
333,704	7389.98	918.57	2138.47	29.59

The first set of experiments focuses on the evaluation of methods described for modeling user interest and expertise, and the resulting user profiles. In this section we aim to answer 2 main research questions.

- **RQ#1. How broad are the topics that users talk about and how are these topics interconnected?** We analyze if there are major differences between users regarding the variance of the topics they tweet about, or more precisely how focused the users are with respect to a certain topic. We also evaluate how the different topics within the evaluated realm are connected.
- **RQ#2. How stable are the generated user profiles over time?** With the availability of large amounts of users, it becomes crucial to reduce the time for recalculating or updating the user profiles with every new post. With this question we want to answer how stable user profiles are over time, allowing us to estimate the best time spans for periodically updating the profiles.

4.2 Dataset

The dataset gathered for our experiments consists of around 5.3 million tweets related to more than 3000 Stocks mainly listed at the NYSE. These tweets were collected using the Twitter Streaming API together with a set of filters. The filters were selected based on the stock symbols of the different companies (for instance, $AAPL for Apple). The idea behind this approach is to collect a series of tweets which contain a direct relation to a stock market company and are thereby of high interest to our domain. The collected tweets were posted by 333,704 different users who posted more than 2.4 billion tweets overall (see Table 1).

Out of this large amount of users we randomly selected a set of 10,000 active users (i.e., users with >100 followers and >100 tweets) for further analysis. For each of the users we downloaded up to 2,400 of the most recent tweets using the Twitter API resulting in a dataset of 12,308,376 tweets. The tweets were annotated using the described method resulting in a set of more than 50 million annotations. The user profiles were generated based on these annotations.

Table 2: Focus of expert users based on the sum of weights of their *top-x* topics

Experts in	Top 1	Top 5	Top 10
Energy	11.6	27.26	61.01
Natural resources	11.82	28.15	62.89
Materials	11.11	27.53	63.04
Transport	9.33	24.6	60.19
Automobiles	10.53	26.25	61.71
Goods	9.54	25.22	59.77
Manufactured goods	9.54	25.22	59.77
Foods	11.39	27.89	63.66
Banking	12.14	31.15	64.64
Insurance	11.84	31.17	64.77
Real estate	9.72	26.35	60.79
Financial services	12.15	31.28	64.92
Health care	17.85	35.62	66.06
Pharmacology	21.9	38.87	68.53
Technology	11.31	28.98	63.6
Software	13.11	31.73	63.94
Information technology management	13.19	33.06	64.04
Telecommunications	13.14	32.47	65.93
Public utilities	9.51	24.59	59.05
Average Expert	12.14	29.34	63.07
All Users	7.89	21.33	54.99

4.3 Topics and Expertise

In Table 2 we show how much the $top-50$ experts of the different domains are focused on their topics. The Table shows the percentage of the weights of the top categories within the profile. We can see that domain experts in the area of *Pharmacology* are very focused on this topic. As expected the domain experts show a strong focus towards their area of expertise when compared to an average user.

Additionally, we analyzed how the topics the users talk about are connected to each other, as shown in Figure 2. The diagram shows the Pearson correlation between the different industries based on the user profiles. We can see some strong correlations within the different sectors. This is expected since the topics of industries like *Computer Hardware* and *Computer Software* are very related. It is interesting to see that there are also other relations between some of the sectors. For instance, the industries around healthcare are also connected to the industries around food and chemistry, or the industries around computer hardware show some connections to entertainment. These connections can help finding users who possess some domain knowledge in a certain area even though their posting behavior does not indicate this.

4.4 Temporal Aspects and Stability of Profiles

Finally, we focus on temporal aspects of the profiles and analyze how stable the profiles are over time. This allows us to estimate the required update interval for different user groups. In order to understand and analyze the evolution of user profiles, we consider the *top-100* users for every domain. Then, we compared the generated profiles based on the first and the second half of the collected tweets from these users. We calculated the Pearson correlation between these profiles to investigate how similar they are. The results are presented in Table 4. We clearly see that the profiles for expert users are more stable than those for average users. This indicates that users who write only about a certain topic, continue to do so. The average time between the profiles was measured by taking the first tweet of both profiles and calculating the time difference. For the tweets we crawled, this time varies between 100 and 200 days. Over this timespan, the user profiles are very stable, so an update of the users is not required very frequently.

Table 3: Relation between Influence and Topics

Top 100 Users	Avg Follower	Avg Statuses	Avg Friends	Avg Listed
Energy	2559.94	17243.63	1590.79	65.37
Natural resources	4685.93	17365.61	3040.12	83.96
Materials	5259.67	23385.05	3023.05	66.88
Transport	4583.44	22488.75	1845.28	70.97
Automobiles	2700.87	27310.71	1961.97	28.91
Goods	2800.68	10056.88	1134.21	80.32
Manufactured goods	4300.66	12220.56	2690.9	28.45
Foods	5153.03	24691.78	2047.7	53.06
Banking	4601.98	17592.06	1327.5	125.3
Insurance	4780.14	14988.76	1053.38	136.6
Real estate	2460.12	14024.47	1154.1	51.82
Financial services	2870.38	7825	1126.87	63.59
Health care	3096.98	8971.55	1110.2	110.4
Pharmacology	3173.98	6302.31	597.75	113.59
Technology	4343.17	23063.76	1625.44	129.33
Software	4109.08	19703.43	1393.48	118.25
IT management	3575.85	15880.88	1462.06	124.63
Telecommunications	3894.77	18417.04	1281.52	100.71
Public utilities	1467.21	12218.47	696.52	27.72
Average	3706.2	16513.19	1587.52	83.15

Table 4: Temporal Stability of Profiles

Experts in	Correlation	Avg. Time between Profiles (Days)
Energy	0.879	161
Natural resources	0.881	125
Materials	0.895	103
Transport	0.813	133
Automobiles	0.924	112
Goods	0.903	178
Manufactured goods	0.801	170
Foods	0.866	108
Banking	0.944	167
Insurance	0.925	158
Real estate	0.882	134
Financial services	0.912	167
Health care	0.961	215
Pharmacology	0.98	197
Technology	0.965	161
Software	0.929	183
IT Management	0.951	223
Telecommunications	0.927	189
Public utilities	0.87	102
Average Expert	0.906	157
All Users	0.782	131

4.5 Discussion

In this section, we evaluated different properties of the generated user profiles. We showed that the topics that users talk about are relatively broad on average. However, our models allow the separation of domain experts who mainly focus on one topic or a set of connected topics. Our last series of experiments showed that

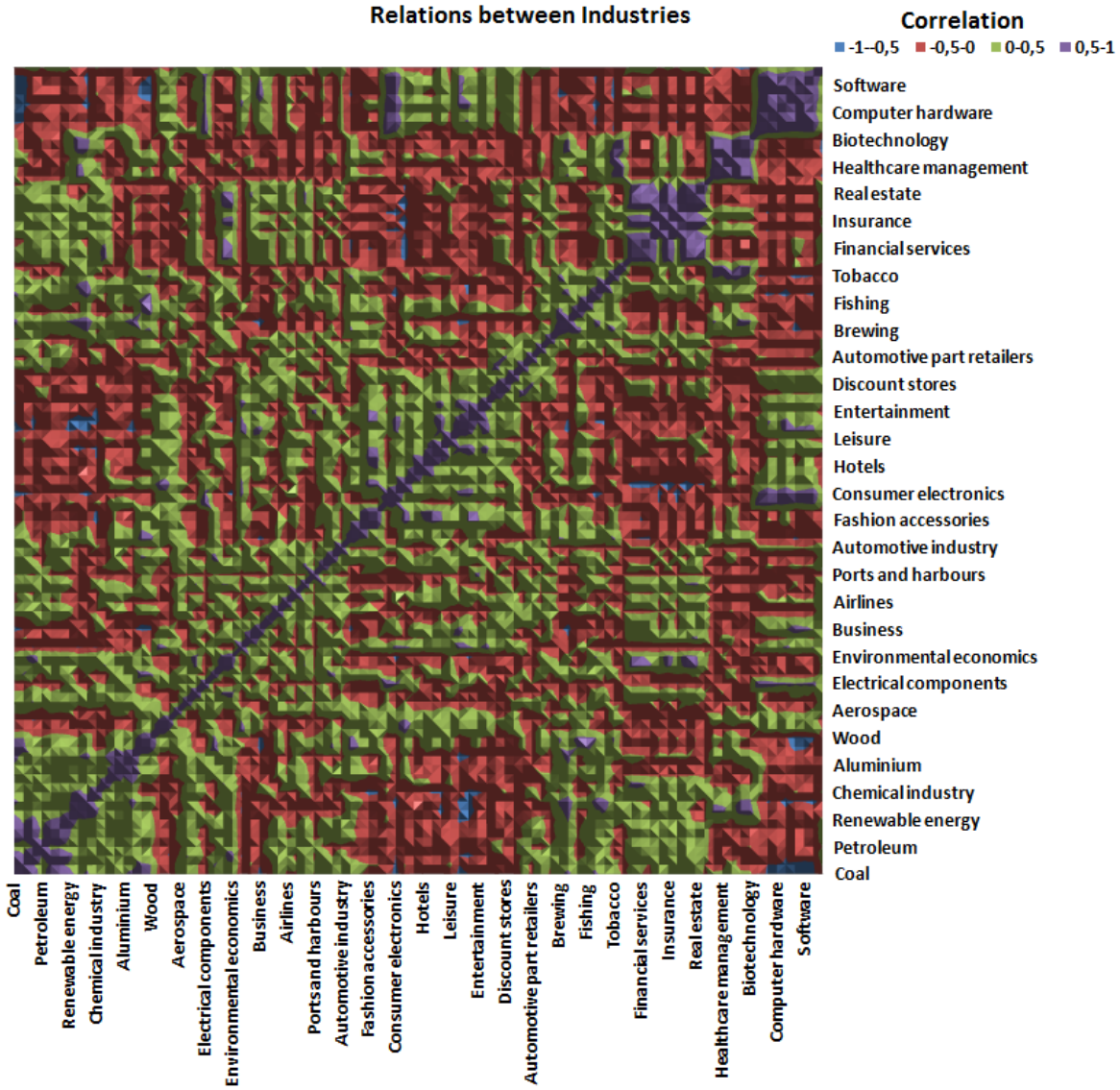

Figure 2: Correlation of different Industries based on User Messages

the topics users write about are relatively stable, indicating that frequent updates or recrawling of user data is not required.

5. EXPERIMENTS FOR TOPIC MODELING AND EVENT DETECTION

In this section, we will describe a series of experiments showing how the expert detection methods can improve the effectiveness of event detection and trend modeling in the area of stock market analysis. The main research questions addressed by these experiments are:

- **RQ#1. Which users talk about which companies?** We can distinguish users based on the calculated interests and posting behavior in certain domains and industry sectors. Based on this, we analyze how effective a selection of certain users for a specified company is. We also analyze how many of the related messages for this company we can retrieve based on the selection of a few expert users.
- **RQ#2. Are small sets of expert users suitable for event detection and prediction?** Our dataset contains tweets related to stocks, these stocks are related to events. By analyzing the tweets from different user groups we can arrive at different predictions. We analyze the timeframes in which expert users talk about a certain stock and in which timeframe the "normal" users mention this stock.

5.1 Dataset

For this series of experiments we use the same dataset as described in Section 4. Additionally, we generated profiles for a set of 10 different companies corresponding to different sectors. These profiles were generated based on the Wikipedia pages of the corresponding companies and the out links of these pages. Two example profiles for the companies "Ford" and "Merck" are shown in Figure 3. We can see a strong focus in the areas of Automobiles and Pharmacology, which are the main areas of interest for these two companies.

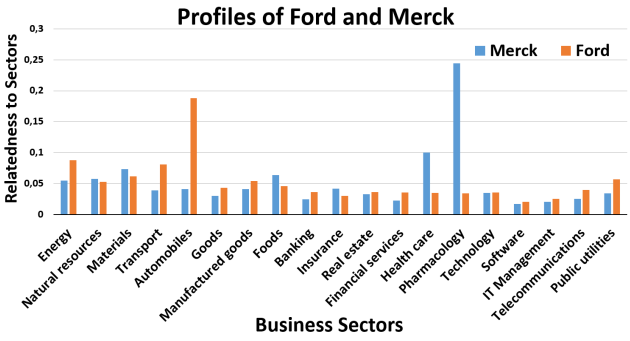

Figure 3: Examples of 2 different companies and their corresponding profiles.

Table 5: Percentage of User Groups Mentioning Stock Names

Company Name	All Users	Active Users 10k	Top 1000 Experts	Top 100 Experts
Apple Inc	38.42	63.69	62.1	**81.0**
Amazon	11.23	**39.39**	22.6	17.0
Alibaba	4.44	22.5	22.6	**35.0**
IBM	2.25	14.97	13.8	**19.0**
Yelp	0.69	10.07	6.1	**11.0**
Dow Chemical	0.31	**4.95**	2.2	0.0
General Electric	1.35	9.19	**14.8**	13.0
Ford Motor Company	1.64	9.44	6.7	**19.0**
Merck & Co	0.72	10.24	33.7	**73**
Pepsico	0.87	8.77	2.6	**9.0**
Average	6.192	19.32	18,72	**27.7**

For further evaluation we selected 10,000 users out of the 333,000 users in our dataset. These users where randomly selected with the constrains that the users had more than 100 tweets and more than 100 followers, since we intend to focus on active users to have a comparable baseline. For each of the chosen companies, we collected the sets of 100 and 1000 most similar users based on the Pearson correlation between the generated profiles of users and companies. To address our first research question, we analyzed the percentages of our user groups which had mentioned one of the analyzed stock names within the monitored period. The results are shown in Table 5. When comparing the top users or the very active users with all users, it is evident that most of the content is posted by the top users. Only highly popular companies like Amazon or Apple got mentioned by a large fraction of average users. In two of the analyzed cases the selected user groups did not match the users talking about the stock well, so the percentage of users within the top expert users was smaller than the percentage within the set of active users. For the other companies the sets of top users contained a higher percentage of users who talked about the company.

In order to assess how useful a selection of specific users can be in the area of event detection, we looked up events for all companies from Yahoo Finance[3]. For each company we choose the 3 events with the strongest impact on the number of posted tweets per day. All these were events which may have direct influence on the stock market, such as the announcement of the quarterly reports. Table 6 shows the companies we used and the number of tweets found within our original dataset.

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

Table 6: Statistics about Stock related dataset.

Company Name	Stock - Symbol	Tweets
Apple Inc	$AAPL	625,950
Amazon	$AMZN	152,952
Alibaba	$BABA	82,733
IBM	$IBM	27,875
Yelp	$YELP	10,777
Dow Chemical	$DOW	4,198
General Electric	$GE	10,981
Ford Motor Company	$F	15,308
Merck & Co	$MRK	12,784
Pepsico	$PEP	9,321

Table 7: Ratio between Tweets on Average Days and Event Days

Company Name	Active - Users	Top 1000	Top 100
Apple Inc	3.908	**4.322**	2.845
Amazon	4.862	7.324	**12.987**
Alibaba	2.852	3.365	**4.625**
IBM	3.272	4.688	**14.223**
Yelp	2.615	11.065	**20.708**
Dow Chemical	2.998	**5.865**	-
General Electric	4.837	4.423	**9.016**
Ford Motor Company	1.942	**6.628**	6.491
Merck & Co	5.106	**9.042**	7.339
Pepsico	1.785	5.031	**7.607**

Each of the chosen events shows a clear spike corresponding to the number of tweets containing the stock symbol of the company. Figure 4 shows one of our example events. We can see that all 3 groups of users show the same pattern. This indicates that all 3 groups of users could be used for detecting the related events.

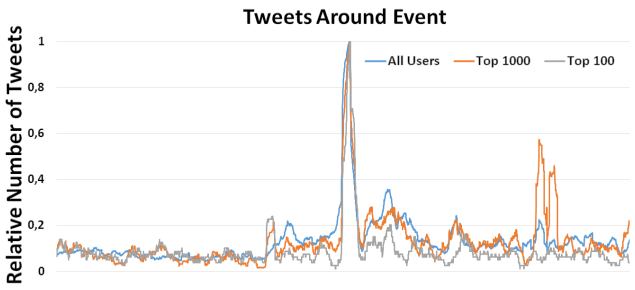

Figure 4: Number of tweets posted around the timespan of an event from different user groups.

Table 7 shows the overall results of our event detection experiments. Since the main focus of this paper is not the evaluation of the event detection algorithm, we choose to use a simple metric for evaluating how good event detection algorithms could work on the different times series. We measured the ratio between the number of tweets at the events and the average number of posts per day. A ratio of 7 for example indicates that on the days of the events, 7 times more tweets were posted as compared to a normal day without a special event. We can see that based on this metric the user groups chosen by expertise and topic outperform the active users by far. This indicates that by using these groups of users event

detection on the generated time series becomes considerably easier, since the events generate spikes which are higher above the average and therefore easier to distinguish from the surrounding noise.

5.2 Discussion

We evaluated how useful the generated sets of users are in terms of the amount of content we can collect per company and in terms of event detection for the different companies. The experiments showed that the selection of enough users is not trivial when small companies are monitored. In these scenarios the selection of more general users might be required. For most of the analyzed companies we found sets of users which were large enough and allowed us to analyze the tweets posted by these users. We evaluated how useful these tweets are for an event detection scenario. For the event detection, special groups of expert users showed the tendency to post content more focused and related to events, which makes detection of these events easier when only these users are monitored.

6. CONCLUSION & FUTURE WORK

In this paper, we presented and evaluated a method for generating user profiles for users from social networks in the domain of stock market analysis. We evaluated the performance of the generated models by analyzing how users who are selected based on their user models, perform in different tasks compared to normal users and very active users. Our findings clearly indicate that the use of expert users based on our proposed approach entails the following benefits.

- This approach allows us to attain high quality content related to the domain or company of interest.
- We can attain a high quality of content while optimizing the set of users.
- We showed that in the area of event detection, the use of a relatively small set of expert users facilitates the detection of relevant events, and can improve the event detection quality.

In our imminent future work we will delve further into the content of the posts from different user groups. Especially in the area of sentiment analysis, abridged with the correlation between the social network content and the stock market data, further evaluations which distinguish expert users from normal users are interesting.

7. REFERENCES

[1] F. Abel, E. Herder, and D. Krause. Extraction of professional interests from social web profiles. *Proc. UMAP*, 34, 2011.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

[3] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein. Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319. Association for Computational Linguistics, 2011.

[4] T. Bodnar, C. Tucker, K. Hopkinson, and S. G. Bilen. Increasing the veracity of event detection on social media networks through user trust modeling. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 636–643. IEEE, 2014.

[5] S. Chelaru, E. Herder, K. D. Naini, and P. Siehndel. Recognizing skill networks and their specific communication and connection practices. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 13–23. ACM, 2014.

[6] R. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.

[7] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd. In *ICWSM*, 2012.

[8] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[9] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen. Mining expertise and interests from social media. In *Proceedings of the 22nd international conference on World Wide Web*, pages 515–526. International World Wide Web Conferences Steering Committee, 2013.

[10] R. Kawase, P. Siehndel, B. P. Nunes, E. Herder, and W. Nejdl. Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources. In *HT*, 2014.

[11] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[12] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. Social media & mobile internet use among teens and young adults. millennials. *Pew Internet & American Life Project*, 2010.

[13] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[14] D. N. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239, 2013.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[16] P. Siehndel and R. Kawase. Twikime! - user profiles that make sense. In *International Semantic Web Conference (Posters & Demos)*, 2012.

[17] R. Yeniterzi and J. Callan. Constructing effective and efficient topic-specific authority networks for expert finding in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 45–50. ACM, 2014.

[18] D. Yimam-Seid and A. Kobsa. Expert-finding systems for organizations: Problem and domain analysis and the demoir approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1–24, 2003.

[19] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.