

# Measuring User Influence in Financial Microblogs: Experiments Using StockTwits Data

Paulo Cortez  
ALGORITMI Centre  
Dep. Sistemas de Informação  
Universidade do Minho  
4800-058 Guimarães,  
Portugal  
pcortez@dsi.uminho.pt

Nuno Oliveira  
ALGORITMI Centre  
Dep. Sistemas de Informação  
Universidade do Minho  
4800-058 Guimarães,  
Portugal  
nunomroliveira@gmail.com

João Peixoto Ferreira  
ALGORITMI Centre  
Dep. Sistemas de Informação  
Universidade do Minho  
4800-058 Guimarães,  
Portugal  
a61572@alunos.uminho.pt

## ABSTRACT

In this paper, we study the effect of graph structure user influence measures in financial social media. In particular, we explore rich and recent data, composed of 1.2 million StockTwits messages, from June 2010 to March 2013. These data allow the creation of social network graphs by considering direct active interactions (retweets, shares or replies). Using such graphs and a realistic rolling windows evaluation, we analyzed four user influence measures (indegree, betweenness, page rank and posts) under two criteria: Percentage of Quality Users (PQU), as manually labeled by StockTwits; and the daily sentiment correlation between top lists of influential users and other users. The sentiment was based on a StockTwits labeled dataset and assessed in terms of three selections: overall sentiment (ALL) and filtered by two major technological companies (Apple – AAPL and Google – GOOG).

Promising results were obtained, with several top lists presenting PQU values higher than 80% and correlations higher than 0.6. Overall, the best results were achieved by the page rank and posts measures.

## Categories and Subject Descriptors

H.2.8 [DATABASE MANAGEMENT]: Database Applications – Data mining; E.1 [DATA STRUCTURES]: Graphs and networks; H.4.2 [DATABASE MANAGEMENT]: Types of Systems – Decision support (e.g., MIS); J.4 [SOCIAL AND BEHAVIORAL SCIENCES]: Economics

## Keywords

Sentiment Analysis, Microblogging Data, Social Networks, User Influence, Stock Markets.

## 1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WIMS '16, June 13 - 15, 2016, Nîmes, France

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4056-4/16/06...\$15.00

DOI: <http://dx.doi.org/10.1145/2912845.2912860>

Due to the expansion of the Internet and Web 2.0 phenomenon, there has been a growing interest in the automatic identification of influential users in social networks [8, 22, 2, 7, 10]. Such identification allows a better understanding of dominant trends and patterns, which can be used for decision support (e.g., designing a better product to fulfill the users' needs, making more effective marketing campaigns).

The rise of social media platforms (e.g., microblogs), has also enabled an explosion of unstructured text with opinions that increased the interest in sentiment analysis, which automatically aggregates the overall sentiment regarding a topic [16]. In particular, using sentiment analysis from social media to model and forecast stock market behavior is a very recent research topic that is presenting promising results [4, 14, 13, 20, 19, 15]. As explained in [14], microblog data is easily available at a low cost and allows a faster and less expensive creation of financial sentiment indicators when compared with traditional large-scale surveys. For instance, the popular American Association of Individual Investors (AAII) [6] survey index is only available at a weekly basis and requires conducting an explicit regular poll to the AII members. Moreover, the financial community that uses social media platforms has grown and thus is potentially more representative of all investors [18]. For instance, the amount of Amazon stock related messages on Yahoo's message board as increased from 70,000 in 1998 to 900,000 in 2006 [9]. The StockTwits data analyzed in this paper (Section 3.1.2) is another example that confirms this financial social media growth.

Despite the growing interest in measuring user influence and mining financial social media, research that combines both topics is very scarce, as explained in Section 2. Nevertheless, a good measure of user influence can potentially provide several benefits within this domain. Most forecasting approaches based on sentiment analysis simply combine financial related messages from all users (e.g., [1, 4, 14, 13, 20]), which might include a large amount of noise. Thus, as also suggested in [18, 3, 10], filtering or weighting such messages according to a user influence criterion might lead to better sentiment indexes and forecasts. Moreover, the identification of influential users can be used to create valuable and alternative social media global sentiment indicators that are more easy and faster to produce when compared with conventional survey indicators (e.g., AII). Also, the automatic detection of influential users is a valuable tool for a financial social media platforms. For instance, using such

tool, influential users could be suggested to be followed by new users, increasing their engagement with the platform.

In this paper, we address this research gap. In particular, we analyze data from StockTwits, a popular microblog service specifically dedicated to the financial domain. The studied data is very large, including a total of 1.2 million messages from 19530 users, and recent, from June 2010 to March 2013. We adapt three known social network structure measures (indegree, betweenness, page rank) and also test a new posts measure to the StockTwits data by considering any direct interactions between two users (retweets, shares or replies). These interactions allow the creation of social network graphs, built using training data and over which the user relevance measures are computed. The respective user rankings are evaluated using a robust and realistic rolling window evaluation scheme that includes five consecutive evaluation periods, each with a six month duration except for the last period, which is related with four months. The measures are analyzed in terms of two aspects: their capacity to select quality users and their daily correlation with other users, assessed in terms of global overall sentiment (ALL), and the sentiment related with two large US technological companies (Apple - AAPL and Google - GOOG).

The rest of the paper is organized as follows. Section 2 describes the related work. Next, Section 3 presents the data and methods. Then, Section 4 presents the experiments held and discusses the research results. Finally, Section 4 draws the main conclusions and also describes perspectives of future work.

## 2. RELATED WORK

Several studies have addressed the measurement of user influence in social networks. The majority of these works study the Twitter microblog. For instance, Weng et al. [22] analyzed Twitter data and proposed the TwitterRank measure, which extends the Page Rank measure by including both topical similarity and link structure, outperforming two social network structure measures (page rank and indegree). Chat et al. [8] also analyzed Twitter, under two measures: indegree and retweets (based on the number of retweets containing the user's name). Using the Spearman's rank correlation, the authors concluded that retweets presents a larger correlation value (e.g., 0.6) with the messages that mention a user when compared with indegree. Also, it was concluded that influential users hold an influence over a wide range of topics. Bakshy et al. [2] studied the diffusion of information on Twitter, concluding that large message cascades tend to be created by users that have been more influential in the past, although it is difficult to predict which particular user will generate large cascades in the future. Brown and Feng [7] also investigated Twitter under social network structure measures, in particular based on an adaptation of the k-shell decomposition level algorithm, which allows to detect the core and hierarchical structure of a network. The algorithm was evaluated in terms of a new proposed authority concept, the potential audience of a user's message excluding peers (two users that are followed by each other), and was able to define a cluster of users with a large authority value.

None of the presented works studied user influence on the financial domain. However, mining financial social media is a recent research trend that has presented promising re-

sults. For instance, Antweiler and Frank studied 1.5 million messages posted on Internet stock message boards (e.g., Yahoo! Finance) and found that stock messages can help in the prediction of market volatility. More recently, Bollen et al. [4] measured global (considering messages from all users) mood states using sentiment analysis when applied to Twitter, finding an accuracy of 86.7% in the prediction of the Dow Jones Industrial Average daily directions. In our previous works, we analyzed Twitter [14, 13] and StockTwits [13] messages (considering equally all users) in order to predict financial variables, such as trading volume. Sprenger and Welpé [20] studied financial Twitter texts, finding that positive tweets (i.e. bullish) are associated with abnormal stock returns. Smailović et al. [19] also analyzed financial Twitter messages, confirming that changes in the positive sentiment probability are associated with changes in stock closing prices. In their analysis, they considered equally all user's opinions.

The analysis of user influence in financial social media is a very scarce research topic. In 2011, Sabherwal et al. [18] have build a weighted sentiment index that aggregates the opinions of stock message boards taking into account each user's reputation credit score, which is attributed by other users. In the same year, Bar-Haim et al. [3] proposed a stock market prediction model based on StockTwits messages and that weights the user's opinions according to their past precision in detecting a stock rise. More closely related with this paper, Eliaçık and Erdogan [10] proposed a new aggregation method for sentiment analysis of financial messages collected from Twitter that weights the individual opinions according to two factors: the user degree of membership, measured using a social network of peers (users that mutually follow each other), and the degree of interest, defined by the number of financial terms (from a specialized Turkish lexicon with 1953 words) that a user tends to post in messages. Using the Spearman's rank correlation, the proposed sentiment analysis method presented a larger correlation (0.56) with a particular stock market (BIST 100) move when compared with a global sentiment analysis method (0.37). Despite this interesting result, it should be noted that the authors did not perform a robust experimentation. For instance, a small number of messages was considered (2408 tweets), related with just one Turkish stock (BIST 100), during a one year period but excluding 11 days related with extraordinary events.

In this paper, we analyze a specialized microblog, StockTwits, that is specifically targeted for the financial domain and thus should be less noisy than generic microblogs, such as Twitter. Moreover, this paper builds on the Twitter user influence earlier works by approaching simple social network topological measures (indegree, betweenness and page rank), which are adapted to the StockTwits data. Also, a new posts measure is explored and that is defined by the number of posts made by an individual user. Rather than using the more rigid follower relationship, as adopted in several works (e.g., [22, 8, 7]), we consider direct user interactions (e.g., retweet) for generating the social network graphs (as explained in Section 3.1.2).

Within our knowledge, this paper is the first approach that studies the effect of user influence measures on financial microblogs and in particular StockTwits. In contrast with [10], we use a much larger sample with around 1.2 million messages related with 19530 users. Also, our evaluation

is more robust. We compare the measures under two criteria: the percentage of quality users and correlation with daily sentiment from other users. These measures are tested under a realistic rolling window evaluation scheme that considers five distinct test periods, each with six or four months. Moreover, no special daily events were excluded.

### 3. MATERIALS AND METHODS

#### 3.1 StockTwits Data

##### 3.1.1 Labeled Messages and Quality Users

The data studied in this paper was kindly provided by StockTwits (stocktwits.com), which is a specialized financial microblog with more than 300,000 users that share messages about the stock market. Similarly to Twitter, messages are limited (maximum of 140 characters) and can contain ideas, Web links, charts and other data.

Recently, users were able to classify their own text messages as: “**bullish**” – optimistic opinion; or “**bearish**” – pessimistic view. In this work, we use these labeled messages, which includes a total of 341230 texts from June 2010 to March 2013, as an easy means to get the user sentiment value. StockTwits messages can be filtered by a \$TICKER tag, known as cashtag. In this work, we analyzed the sentiment of all labeled messages (selection ALL, with no filter) and filtered by two company \$TICKER tags: Apple – \$AAPL and Google – \$GOOG. The AAPL and GOOG stocks were selected since they contained a large number of messages throughout the time period that was analyzed. Figure 1 plots the total number of messages collected for the ALL, AAPL and GOOG selections, where each  $t_i$  time period has a duration of six months. The plot shows an exponential increase of messages, with much larger numbers for the last time periods ( $t_4$  and  $t_5$ ). Such increase confirms the StockTwits service popularity expansion.

We also had access to a static dataset with user information known in March 2013. Of particular interest to this work, the static dataset contains a set of 300 users ( $S_U$ ) that were labeled by StockTwits as “suggested”, meaning high quality contributors that are included in a curated list of suggested users and that is provided for normal users. These contributors were manually selected by StockTwits according to their knowledge and quality of their past behavior over time.

##### 3.1.2 Interactions and Social Network Graphs

StockTwits includes other interesting features that were used in this work to build social networks. As in Twitter, users can perform a **retweet**, where a user reposts a message or forwards it to other users. Such retweet is executed by manually adding a “RT” term to a previous message. Other features were later introduced by StockTwits, such as **shares** and **replies**. Similarly to a retweet, where users can **share** automatically a message by clicking a special button at the StockTwits platform. Users can also **reply** to a message and the thread of replies related with an original post is called a conversation.

We had access to datasets related with retweets, shares and conversations. Since some of these features were added later, the time periods differ. The retweet dataset is from June 2010 to March 2013 and includes 233722 messages. The share data is from May 2011 to March 2013 and contains

65613 texts. Finally, the dataset with conversations ranges from January 2012 to March 2013 and is made of 959417 messages. Figure 2 plots histograms for the three datasets. The plot shows an initial growing trend of retweets that is stopped when the similar, but more easy to use, share button was introduced. This effect is clearly visible at period  $t_5$ .

Most of the literature works (e.g., [22, 8, 7]) create a social network graph by considering an edge if user  $A$  follows user  $B$ . In this work, we assume a slight distinct approach, where a social network interaction is related with a retweet, a share or a reply from user  $B$  to a message posted by user  $A$ . Such distinct approach was adopted due to two main reasons. First, we did not have access to a dynamic (time evolving) dataset with the list or number of followers for all users. Second, we believe that using active actions (e.g., direct reply or clique of a share button) is a more realistic and dynamic representation of a “true” interaction between two users when compared with follower relationships. Thus, the active  $A \rightarrow B$  interactions (retweet, share or reply) are used to create directed graphs. We assume that the higher are the  $A \rightarrow B$  interactions, the stronger is the influence of  $A$  in  $B$ . The direct graph  $G_{t_i}(V_{t_i}, E_{t_i})$  is composed of:

- a vertex set  $V_{t_i}$  – with the social network users that performed retweet, share or reply actions during time period  $t_i$ ; and
- a edge set  $E_{t_i}$  – with a merge of all retweet, share or reply direct interactions made from  $V_{t_i}$  users within time period  $t_i$ .

The exception is related with self interactions ( $A \rightarrow A$ ), which were removed since we consider that self interactions are not relevant for the purpose of this work (user influence measurement). The topological user influence measures (indegree, betweenness and page rank) are computed over the obtained direct graph, as explained in the Section 3.2.

For demonstration purposes, Figure 3 shows two social network direct graphs, related with the time periods  $t_0$  (left) and  $t_5$  (right). To simplify the visualization of the graphs, users (nodes) and direct interactions (edges) are plotted only if there is a minimum of 10  $A \rightarrow B$  interactions. The comparison between the left and right graphs of Figure 3 shows a large increase in the amount of users and interactions between the two periods, which again attests the StockTwits platform expansion during this period.

#### 3.2 User Influence Measures and List of Top Users

In this work, we adopted three topological user influence measures (indegree, betweenness and page rank) that computed over the social network direct graph of interactions. As explained in Section 3.1.2, one unique graph is built at given time period and that merges all retweet, share and reply interactions. Also, the three influence measures were computed using unweighted graphs, i.e., they considered only if there was an edge between two users but not the number of interactions associated with such edge. We also test one simple measure, called posts. The four measures are:

- **Indegree** – a popular user influence measure (e.g., used in [22, 8]) based on the number of users (type  $B$ ) that interact another user (type  $A$ );

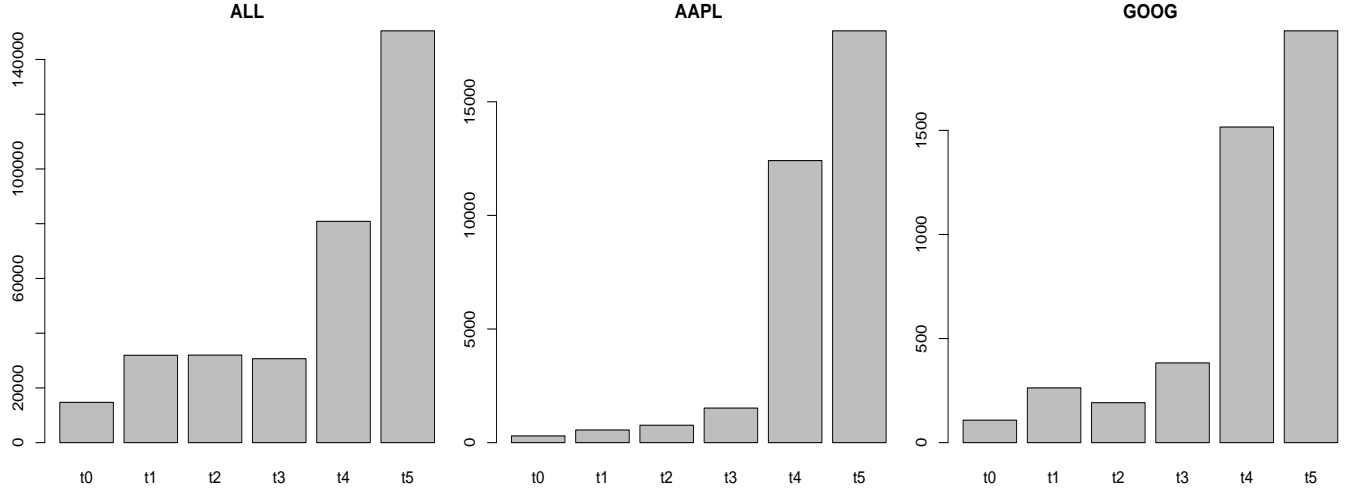


Figure 1: Histogram of StockTwits messages over the time periods analyzed and for the ALL, AAPL and GOOG selections ( $x$ -axis denotes the time period  $t_i$ ;  $y$ -axis the frequency of the messages)

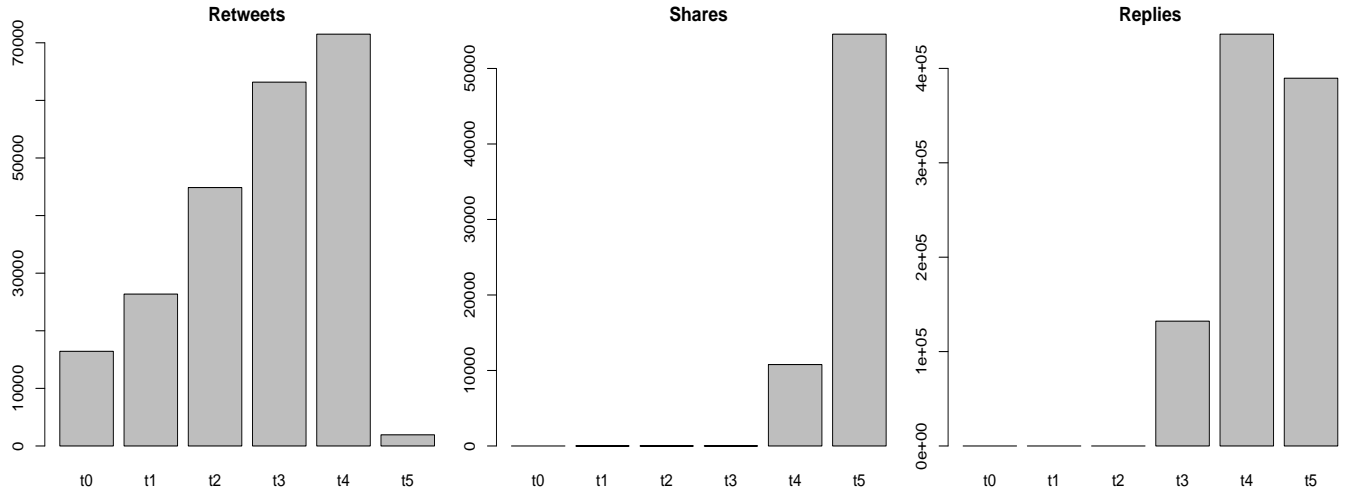


Figure 2: Histogram of StockTwits messages over the time periods analyzed and for retweets, shares and replies ( $x$ -axis denotes the time period  $t_i$ ;  $y$ -axis the frequency of the messages)

- **Betweenness** – an indicator of the centrality of user in a network, it is equal to the number of shortest paths of interactions that go through a user [11];
- **Page Rank** – based on the famous page rank algorithm used by Google [5] and that takes into account the number and quality of the edges to estimate the importance of a node; and
- **Posts** – a simple measure explored in this paper and that counts the number of posts made by user  $A$  that received a direct interaction (retweet, share or reply) from other users.

For all these measures, the higher the value, the more relevant is the user. Thus, for a particular social network direct graph, we compute the four measures and rank all users according to a decreasing ordering of the measure values. For a particular user ranking, related with one measure, we de-

fine a list with the top (most influential)  $r$  users and that is denoted as  $Tr$ . For instance,  $T12$  represents a list with the 12th most relevant users.

For demonstration purposes, Table 1 presents examples of user influence measures computed for five users ( $a$ ,  $b$ ,  $c$ ,  $d$  and  $e$ ). Considering Table 1, the top list  $T3$  for Indegree includes users  $\{e, b, d\}$  (with the highest measure values). Similarly, the other  $T3$  lists are:  $\{c, e, a\}$  – for Betweenness;  $\{e, d, c\}$  – for Page Rank; and  $\{e, a, d\}$  – for Posts.

### 3.3 Evaluation

In order to achieve a realistic and robust evaluation of the user influence measures, we adopt a rolling window method [21, 12], where the data is split into several train and test time periods, ordered in time and from June 2010 until March 2013. The rolling window consists of several iterations. In the first iteration, the first period ( $t_0$ ) is used to compute the social network graph ( $G_{t_0}$ ) and user relevance

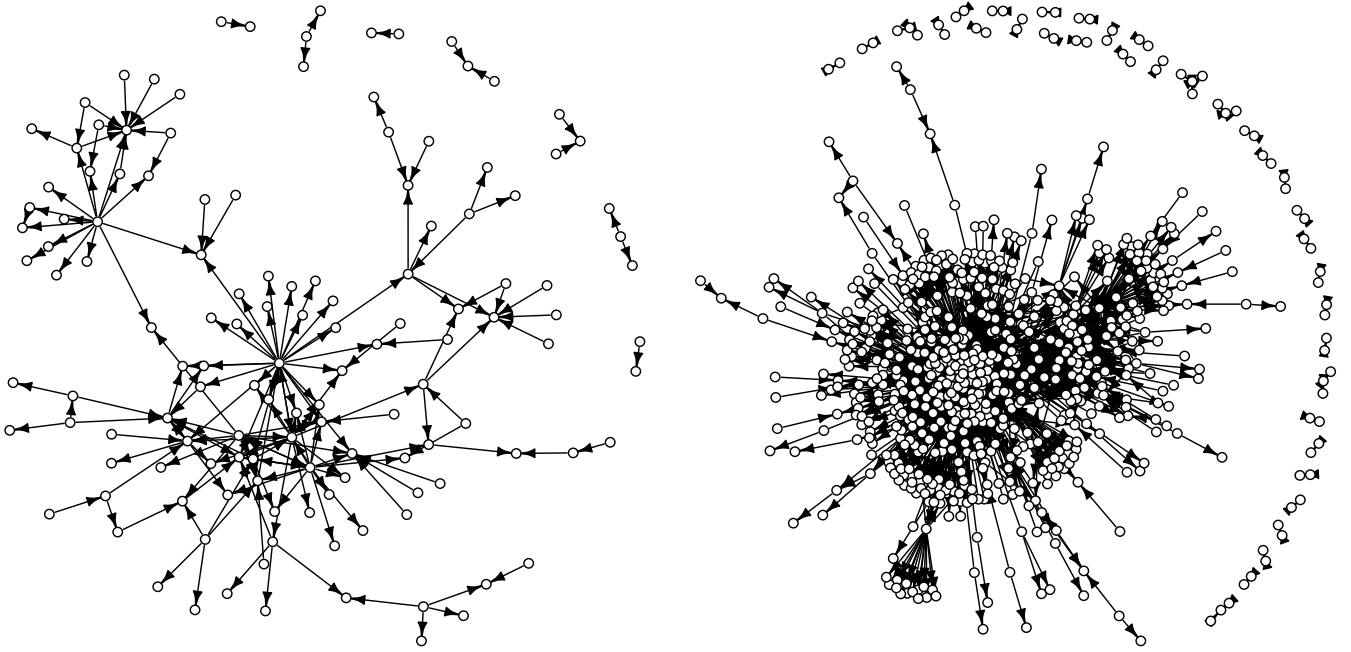


Figure 3: StockTwits social network graph (with minimum of 10 interactions)  $G_0$  (left) and  $G_4$  (right)

Table 1: Example of user influence measure values

	User Indegree	Betweenness	Page Rank	Posts
<i>a</i>	2	167.5	0.011	67
<i>b</i>	3	0.0	0.012	20
<i>c</i>	1	175.0	0.013	13
<i>d</i>	3	61.0	0.017	41
<i>e</i>	5	171.7	0.020	109

measures, while the next sequential period ( $t_1$ ) is used to evaluate the quality of the measure. Thus,  $t_0$  works like a “training set”, while  $t_1$  is a “test set”. In the next iteration, the “training” and “test” sets are slided, thus allowing to build the graph  $G_{t_1}$  and assess the quality of the user influence measures on test set  $t_2$ , and so on. The size of each period was set to six months. The exception is for the last period, which includes only four months but that has a larger amount of messages and interactions, as shown in Table 2. The six month period was selected as a reasonable compromise between the amount of rolling window iterations (five) and test period length (six months), in order to get meaningful correlation values (with enough number of days  $N$ ). In the table, column **Time** is the time period,  $N_G$  is the total number of users for graph  $G_{t_i}$ , **Messages** is the total number of StockTwits labelled messages (shown in Figure 1), **Begin** and **End** are the beginning and ending dates of the period. For time period  $t_5$ , the  $N_D$  is empty because it corresponds to the last test set, thus no graph was built for this period.

In this work, the user influence measures and their respective lists of top users are evaluated using two distinct aspects: the capacity to select quality users and sentiment analysis correlation.

Table 2: Details of the time periods

Time	$N_G$	Messages	Begin	End
$t_0$	1336	14729	June 2, 2010	December 1, 2010
$t_1$	1755	31924	December 2, 2010	June 1, 2011
$t_2$	2880	31983	June 2, 2011	December 1, 2011
$t_3$	3851	30667	December 2, 2011	June 1, 2012
$t_4$	12414	80865	June 2, 2012	December 1, 2012
$t_5$	–	150463	December 2, 2012	March 30, 2013

The first aspect is based on the Percentage of Quality Users (PQU), defined as:

$$\begin{aligned} S_T &= T_j \cap S_U \\ PQU &= \#(S_T) / \#(T_j) \end{aligned} \quad (1)$$

where  $T_j$  denotes a particular list of top users using a measure,  $S_U$  is the static set of StockTwits suggested users (i.e., the gold standard),  $S_T$  is the set of suggested users included on list  $T_j$  and  $\#$  denotes the cardinality operator.

The second aspect of evaluation is based on popular Spearman’s rank coefficient (used also in [8, 10]), often used as a measure of association strength between two variables:

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{N^3 - N} \quad (2)$$

where  $N$  denotes the number of observations. The Spearman’s rank correlation calculation is a nonparametric statistic for measuring the strength of a monotonic function and it can range from -1 (strong negative correlation between  $x_i$  and  $y_i$  values) to 1 (strong positive correlation), where a 0 value means no correlation. In this paper, we assume a daily analysis and the labeled dataset of messages. For a particular day, each user might issue several positive (bullish) or

negative (bearish) sentiment opinions. Thus, we aggregate all user sentiment opinions under an overall bullishness index  $B_d(U)$  [13], also termed as positive sentiment probability in [19]:

$$B_d(U) = \frac{\#(Bull_d)}{\#(Bull_d) + \#(Bear_d)} \quad (3)$$

where  $Bull_d$  and  $Bear_d$  are the set of bullish and bearish messages for day  $d$  and considering all  $U$  users. The  $B_d(U)$  values range from 0 (full negative opinion) to 1 (full positive opinion).

We assume that a particular top list of users ( $T_j$ ) is relevant if their associated bullishness index  $B_d(T_j)$  correlates with the bullishness index of other users  $B_d(U')$ ,  $U' = U_d - T_j$ , where  $U_d$  is the set of users that post a StockTwits message at day  $d$ . In order to achieve a fair comparison, the correlation ( $\rho$ ) is computed using the same set of days for all top lists ( $T_j, j \in \{1, 2, \dots, M\}$ ). For some message selections (e.g., GOOG), periods (e.g.,  $t_2$ ) and top lists (e.g.,  $T_j = T12$  for betweenness), the number of days with labeled messages related with  $T_j$  is too low. Thus, in order to get more robust correlation value, we define a minimum amount of days  $N_{\min}$  in order to consider the list  $T_j$  in the correlation comparison. The final correlation ( $\rho$ ) is computed using  $x_i = B_i(T_j)$ ,  $y_i = B_i(U')$  and  $i \in \mathcal{D}$  for a particular time period, message selection and all  $M$  top lists, such that:

$$\mathcal{D} = \bigcap_{j \in J} D_j, J = \{1, \dots, M\} : \#(D_j) \geq N_{\min} \wedge \#(\mathcal{D}) \geq N_{\min} \quad (4)$$

where  $D_j$  denotes the set of days with labeled messages for top list  $T_j$  and the number of correlation observations is  $N = \#(\mathcal{D})$ .

## 4. EXPERIMENTS AND RESULTS

All experiments here reported were conducted using the open source **R** tool [17]. For each of the four measures (in-degree, betweenness, page rank and posts), we defined five lists of top users, according to the growing scale: T12, T25, T50, T100 and T200. Thus, the number of analyzed top lists is  $M = 4 \times 5 = 20$ .

### 4.1 Capacity to Select Quality Users

The obtained PQU results for the distinct social network graphs ( $G_{t_0}, G_{t_1}, \dots, G_{t_4}$ ) and list of top influential users are presented in Table 3. In the table, the PQU values are shown in percentage and the best value of each row is underlined. The row **Avg.**, contains the average value over all test periods ( $t_1, t_2, \dots, t_5$ ) and for the same top list ( $T_j$ ), while the last row **O.A.** shows the overall average value for a user measure (for all top lists and test periods).

An analysis to the table reveals that there is a majority of cases where PQU values are higher or equal to 50% (highlighted in **bold**). In some cases, high PQU values (e.g., higher or equal to 80%) we obtained, such as: T25 for In-degree and  $t_1$ , T12 for betweenness and  $t_1$  and  $t_2$ ; T25 for page rank and  $t_1$ ; T50 for posts and  $t_1$  and  $t_2$ . The best vertical average results are obtained by T12 for betweenness and T25 for page rank. The best overall PQU measure value is achieved by posts, although with slight differences when compared with page rank and in-degree.

Overall, the table shows two major patterns that all four measures follow. First, in general, higher PQU values are

obtained for the initial test periods when compared with the most recent ones (vertical pattern). Second, the smaller lists, such as T12 and T25, tend to present higher PQU values when compared with the larger tops, such as T100 and T200 (horizontal pattern). Regarding the vertical pattern, much lower results were obtained for the last two test periods ( $t_4$  and  $t_5$ ), which corresponds to a stronger use of the StockTwits platform, with much more users and interactions.

We note that the static list of suggested users  $S_U$  is rather small, with just 300 users. While StockTwits does not explicitly describe how often it updates their curated list of suggested users, the subjective mentioning of contributors that are “constructive over time”<sup>1</sup> indicates that StockTwits tends to favor older contributors, which might explain the results from Table 3, with much higher PQU values for the first three periods ( $t_1, t_2$  and  $t_3$ ) and lower PQU values for the most recent periods ( $t_4$  and  $t_5$ ). In effect, the obtained results suggest that the StockTwits curated list  $S_U$  is too conservative and thus it should be updated by StockTwits in order to include more users. Given the large amount of StockTwits users (with currently more than 300,000 users), such update could be facilitated by considering automatic measures of user influence similar to the ones explored in this paper.

### 4.2 Sentiment Analysis Correlation

The obtained sentiment analysis correlation values for the ALL, AAPL and GOOG message selections are presented in Table 4. In the experiments, the minimum amount of days was set to a reasonable value of  $N_{\min} = 20$  that allows to obtain more top list values for the AAPL and GOOG selections. In practice, using such  $N_{\min}$  threshold led to a higher minimum number of observations, which is  $N = 26$  for GOOG and  $t_2$ . For the ALL selection, which includes a larger amount of messages, the number of observations are higher, reaching a maximum of  $N = 184$  for  $t_1$  and  $t_2$ . In the table, column  $N_M$  denotes the total number of users that posted messages within that test period:

$$N_M = \#(\bigcup_{d \in \mathcal{D}} U_d) \quad (5)$$

The  $N_M$  values attest the increase of user’s adoption of the StockTwits platform. Similarly to Table 3, the best value within a test period is underlined and there are rows with the vertical average over all test periods (**Avg.**) and overall user measure (**O.A.**) Also, missing correlations are signaled with the “-” symbol (e.g., AAPL, T12,  $t_1$  and betweenness). Such missing correlations correspond to a  $T_j$  with a very small amount of daily messages, i.e.,  $\#(D_j) < N_{\min}$  or with a constant bullishness value (non meaningful correlation).

A large portion of Table 4 values are related with moderate correlations (higher than 0.40, highlighted in **bold**), in particular for selection ALL (from  $t_1$  to  $t_4$ ), AAPL ( $t_2, t_3$  and  $t_4$ ) and GOOG ( $t_1$  and  $t_5$ ). In some cases, there are stronger correlations, such as: 0.64 – ALL,  $t_2$ , T50 and T100 for page rank; 0.63 – ALL,  $t_3$ , T12 for page rank; 0.58 – ALL,  $t_4$ , T200 for betweenness. A visual demonstration of these and other example correlations is provided in Figure 4, which plots the  $x_i$  bullishness index values for the top list

<sup>1</sup>See: <http://help.stocktwits.com/customer/en/portal/articles/1877-how-do-i-find-people-to-follow->.

Table 3: Capacity to select quality users (PQU values, in %, best value per period is underlined, values  $\geq 50\%$  are in bold)

Time	Indegree					Betweenness					Page Rank					Posts				
	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200
$t_1$	<b>75</b>	<u>84</u>	<b>76</b>	<b>64</b>	<b>52</b>	<b>83</b>	<b>72</b>	44	46	30	<b>75</b>	<u>84</u>	<b>76</b>	<b>63</b>	<b>50</b>	<b>67</b>	<b>76</b>	<b>74</b>	<b>66</b>	<b>53</b>
$t_2$	<b>67</b>	<b>80</b>	<b>78</b>	<b>72</b>	<b>56</b>	<u>83</u>	<b>76</b>	<b>54</b>	46	39	<b>75</b>	<b>80</b>	<b>80</b>	<b>69</b>	<b>55</b>	<b>75</b>	<b>80</b>	<b>80</b>	<b>72</b>	<b>55</b>
$t_3$	<b>58</b>	<b>76</b>	<b>78</b>	<b>69</b>	<b>54</b>	<b>58</b>	<b>56</b>	<b>56</b>	43	34	<b>67</b>	<u>80</u>	<b>78</b>	<b>67</b>	<b>54</b>	<b>58</b>	<b>76</b>	<u>80</u>	<b>71</b>	<b>56</b>
$t_4$	25	16	16	16	15	33	<u>36</u>	26	27	18	17	20	16	14	18	33	24	18	21	22
$t_5$	25	20	18	12	10	25	<u>28</u>	20	14	12	25	20	22	13	12	25	20	20	13	12
Avg.	<b>50</b>	<b>55</b>	<b>53</b>	47	38	<u>57</u>	<b>54</b>	40	35	27	<b>52</b>	<u>57</u>	<b>54</b>	45	38	<b>52</b>	<b>55</b>	<b>54</b>	49	40
O.A.	48					42					49					<u>50</u>				

Table 4: Correlation sentiment analysis values (best value per period is underlined; values  $\geq 0.40$  are in bold)

	Indegree					Betweenness					Page Rank					Posts						
Time	$N_M$	$N$	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200	T12	T25	T50	T100	T200
ALL																						
$t_1$	861	177	<b>0.42</b>	0.34	0.38	<b>0.47</b>	<b>0.48</b>	0.27	0.33	0.35	<b>0.43</b>	<b>0.45</b>	0.39	<b>0.40</b>	0.38	<b>0.44</b>	<b>0.49</b>	<b>0.43</b>	0.36	<b>0.45</b>	<b>0.48</b>	<b>0.56</b>
$t_2$	1207	184	0.33	<b>0.62</b>	<b>0.53</b>	<b>0.56</b>	<b>0.60</b>	<b>0.54</b>	<b>0.54</b>	<b>0.58</b>	<b>0.60</b>	<b>0.57</b>	<b>0.61</b>	<b>0.61</b>	<u>0.64</u>	<u>0.64</u>	<b>0.59</b>	<b>0.57</b>	<b>0.63</b>	<b>0.55</b>	<b>0.56</b>	<b>0.60</b>
$t_3$	1407	184	<b>0.58</b>	<b>0.61</b>	<b>0.61</b>	<b>0.58</b>	<b>0.60</b>	<b>0.47</b>	<b>0.49</b>	<b>0.59</b>	<b>0.59</b>	<b>0.58</b>	<u>0.63</u>	<b>0.60</b>	<b>0.62</b>	<u>0.63</u>	<b>0.60</b>	<b>0.58</b>	<b>0.59</b>	<b>0.61</b>	<b>0.58</b>	<b>0.61</b>
$t_4$	4560	171	0.36	<b>0.40</b>	<b>0.40</b>	<b>0.45</b>	<b>0.46</b>	<b>0.40</b>	<b>0.45</b>	<b>0.44</b>	<b>0.46</b>	<b>0.58</b>	<u>0.41</u>	0.39	<b>0.43</b>	<u>0.41</u>	<b>0.45</b>	<b>0.43</b>	<b>0.40</b>	<b>0.42</b>	<b>0.47</b>	<b>0.44</b>
$t_5$	6057	119	0.08	0.14	0.14	0.07	-0.02	0.03	0.06	0.02	0.21	<u>0.30</u>	0.02	0.11	0.22	0.12	0.23	0.06	0.13	0.14	0.14	0.00
Avg.			0.35	<b>0.42</b>	<b>0.41</b>	<b>0.43</b>	<b>0.43</b>	0.34	0.37	0.39	<b>0.46</b>	<u>0.50</u>	<b>0.41</b>	<b>0.42</b>	<b>0.46</b>	<b>0.45</b>	<b>0.47</b>	<b>0.41</b>	<b>0.42</b>	<b>0.43</b>	<b>0.45</b>	<b>0.44</b>
O.A.			<b>0.41</b>					<b>0.41</b>					<u>0.44</u>					<b>0.43</b>				
AAPL																						
$t_1$	89	76	0.22	0.22	0.21	0.19	0.14	-	<u>0.34</u>	0.15	0.16	0.14	0.22	0.22	0.21	0.15	0.14	0.22	0.22	0.19	0.19	0.06
$t_2$	187	92	<b>0.58</b>	<b>0.55</b>	<b>0.48</b>	<b>0.51</b>	<b>0.48</b>	<b>0.42</b>	0.38	0.35	<b>0.43</b>	<b>0.42</b>	<b>0.54</b>	<b>0.54</b>	<b>0.51</b>	<b>0.47</b>	<b>0.44</b>	<b>0.54</b>	<b>0.52</b>	<b>0.48</b>	<b>0.51</b>	<b>0.48</b>
$t_3$	262	121	<u>0.52</u>	<u>0.52</u>	<b>0.51</b>	<b>0.51</b>	<b>0.51</b>	<b>0.43</b>	<b>0.41</b>	<b>0.51</b>	<b>0.44</b>	<b>0.46</b>	<u>0.52</u>	<u>0.52</u>	<b>0.51</b>	<b>0.51</b>	<b>0.50</b>	<u>0.52</u>	<u>0.52</u>	<b>0.51</b>	<b>0.51</b>	<b>0.44</b>
$t_4$	1233	114	0.24	0.38	<b>0.43</b>	<b>0.42</b>	<b>0.42</b>	0.31	0.31	0.26	0.19	0.36	0.37	0.39	<b>0.43</b>	<b>0.46</b>	<b>0.41</b>	0.37	0.38	<b>0.45</b>	<b>0.43</b>	<b>0.41</b>
$t_5$	1587	99	0.23	0.28	0.33	0.25	0.30	0.20	0.27	0.31	0.39	0.34	0.26	0.36	0.35	0.18	0.30	0.23	0.25	<u>0.40</u>	0.21	0.31
Avg.			0.36	0.39	<b>0.40</b>	0.38	0.37	0.34	0.34	0.32	0.32	0.34	0.38	<u>0.41</u>	<b>0.40</b>	0.35	0.36	0.38	0.38	<u>0.41</u>	0.37	0.34
O.A.			<u>0.38</u>					0.33					<u>0.38</u>					0.37				
GOOG																						
$t_1$	54	32	<b>0.58</b>	<b>0.49</b>	<b>0.49</b>	<b>0.49</b>	<b>0.49</b>	-	-	-	-	-	<b>0.58</b>	<b>0.58</b>	<b>0.49</b>	<b>0.49</b>	<b>0.50</b>	<b>0.58</b>	<b>0.49</b>	<b>0.49</b>	<b>0.49</b>	<b>0.50</b>
$t_2$	70	26	<u>0.22</u>	0.22	0.22	0.22	0.22	-	-	-	-	-	0.22	0.22	0.22	0.21	0.21	0.22	0.22	0.22	0.22	0.22
$t_3$	110	52	0.31	0.31	0.31	0.31	0.35	-	-	0.31	0.36	0.34	0.32	0.31	0.31	0.33	<u>0.37</u>	0.31	0.31	0.31	0.31	0.33
$t_4$	423	55	-0.03	-0.05	-0.03	-0.02	0.09	0.02	0.01	-0.04	0.02	0.03	-0.03	-0.07	-0.04	-0.06	0.08	0.04	-0.04	-0.04	-0.02	<u>0.10</u>
$t_5$	486	40	0.18	0.39	<b>0.43</b>	<b>0.41</b>	<b>0.44</b>	0.09	0.34	0.31	<b>0.41</b>	<b>0.49</b>	0.18	0.18	<b>0.45</b>	0.39	<b>0.45</b>	0.18	0.39	<b>0.43</b>	<b>0.41</b>	<b>0.45</b>
Avg.			0.25	0.28	0.28	0.28	<u>0.32</u>	0.05	0.18	0.19	0.26	0.29	0.26	0.25	0.29	0.27	<u>0.32</u>	0.27	0.28	0.28	0.28	<u>0.32</u>
O.A.			<u>0.28</u>					0.21					<u>0.28</u>					<u>0.28</u>				

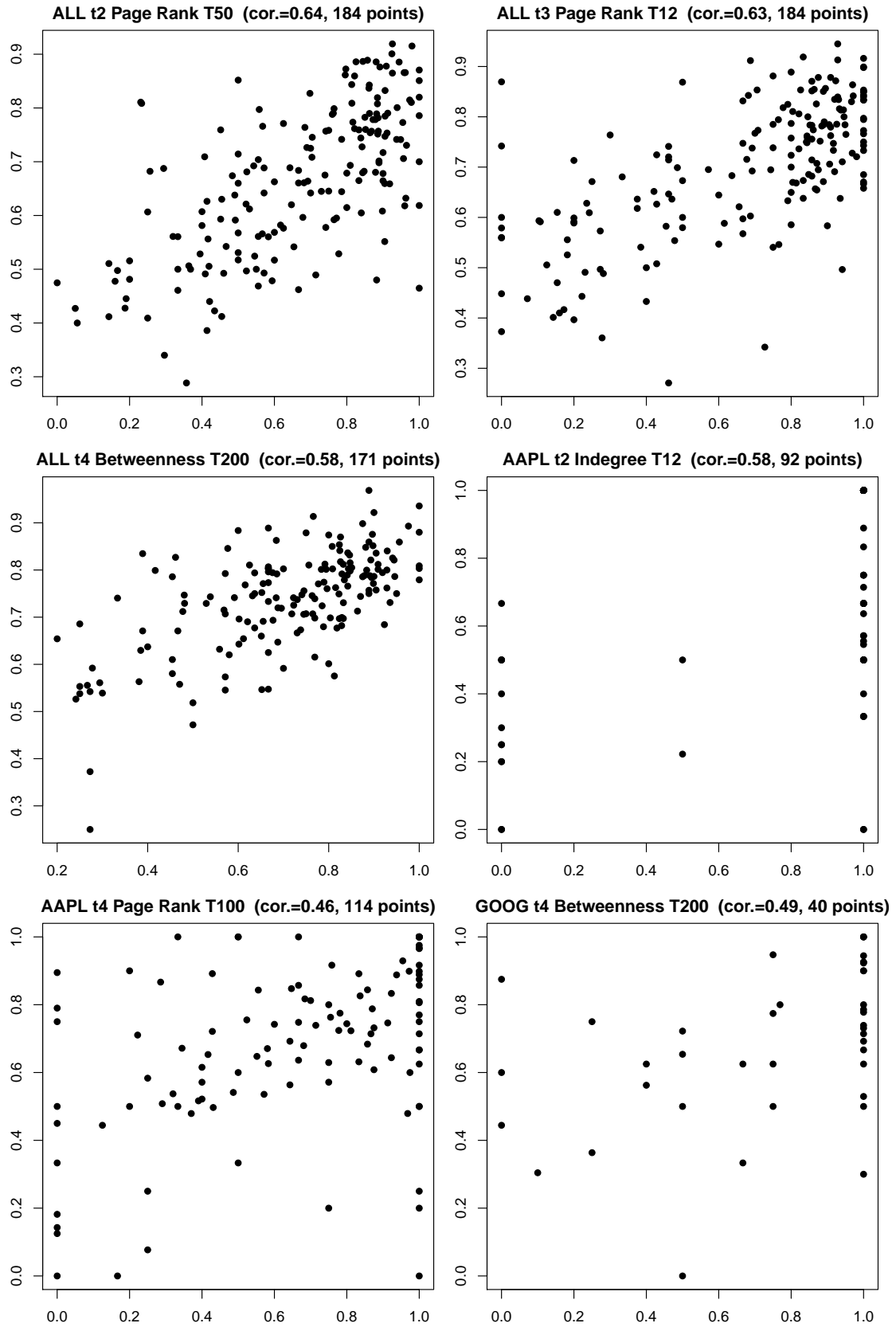


Figure 4: Examples of top list measure sentiment correlations ( $x$ -axis – bullishness index values for the top list;  $y$ -axis – bullishness index values for other users)



$T_j$  ( $x$ -axis) and  $y_i$  bullishness values for other users ( $U'$ ,  $y$ -axis). In particular, Figure 4 shows interesting correlations for the ALL selection.

In general, the results are better for the general sentiment (ALL selection), followed by the Apple stock (AAPL). Also, some interesting vertical average values were obtained, such as: ALL selection – T200 for betweenness; and AAPL – T50 for page rank. Moreover, the best overall average (O.A.) results are provided by page rank, followed by posts and indegree.

The correlation differences between the two examined companies (AAPL and GOOG) might be explained by the distinct numbers of StockTwits messages (much higher for AAPL when compared with GOOG, as shown in Figure 1) or by differences in AAPL and GOOG communities of investors. This latter possibility will be addressed in future work.

## 5. CONCLUSIONS

The main goal of this work was to study user influence measures computed using the social network structure of financial microblogs. In particular, we focused on recent StockTwits data, which included 1.2 million messages from 19530 users from June 2010 to March 2013. Using direct interactions between two users (retweets, shares or replies), we have generated five social network graphs, each built using a distinct and sequential six month training time period. Then, four measures (indegree, betweenness, page rank and posts) were computed over these graphs in order to create several top lists of users (T12, T50, T100 and T200). Next, the top lists were evaluated, under a realistic rolling window evaluation scheme that considers unseen test data related with the next six/four months. We adopted two evaluation criteria: Percentage of Quality Users (PQU), as classified by StockTwits, and the daily sentiment correlation between the top list and other users, using a user labeled dataset and under three message selections: ALL – general sentiment; and two technological stocks, AAPL - Apple and GOOG - Google.

Interesting results were achieved, with some top lists resulting in high quality values (e.g., PQU higher than 80%, Spearman correlation higher than 0.6). Overall, the best selection of quality users (PQU) was obtained by the posts measure (50%), followed by page rank (49%). Regarding the correlations, the page rank measure achieved the best overall result (0.44 for ALL, 0.38 for AAPL, 0.28 for GOOG), followed by the posts measure. Some specific top lists obtained interesting average (over all five test periods) results, such as: betweenness T200 for ALL (0.50), page rank T25 for AAPL (0.41) and posts T50 for AAPL (0.41).

Within our knowledge, this is the first work that studies user influence measures on financial microblogs. The obtained results are promising, showing that direct interactions from financial microblogs can be used to rank the quality and influence of users. This opens several future research directions. For instance, we intend to explore other features, such as number of mentions (retrieved using text analysis) or number of likes (which were more recently implemented, after 2013). Also, we plan to study more complex effects on sentiment influence, such as by considering other sentiment aggregation periods (e.g., hourly) or the diffusion of sentiment message cascades. Also, we aim to analyze other financial microblog messages, such as Twitter (by using cashtag filters). Since Twitter does not have labeled messages, clas-

sified by their own users, specialized financial lexicons, such as publicly provided in [15], can be used to automatically get the overall sentiment (“bullish” or “bearish”) of a message. Also, we intend to explore influence measures and user rankings to weight the sentiment opinions when designing models to forecast stock market variables (e.g., returns, volatility).

## 6. ACKNOWLEDGMENTS

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. We also thank StockTwits for the provision of their data.

## 7. REFERENCES

- [1] W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.
- [2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the 3rd International Conference on Web Search and Data Mining (WSDM)*, pages 65–74, Hong Kong, 2011. ACM.
- [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] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
- [5] S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.
- [6] G. W. Brown and M. T. Cliff. Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1):1–27, 2004.
- [7] P. Brown and J. Feng. Measuring user influence on twitter using modified k-shell decomposition. In *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM)*, pages 18–23, Barcelona, Spain, 2011. AAAI.
- [8] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM)*, pages 10–17, Washington, DC, USA, 2010. AAAI.
- [9] S. R. Das and M. Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.
- [10] A. B. Eliaçik and N. Erdogan. User-weighted sentiment analysis for financial community on twitter. In *Proceedings of the 11th International Conference on Innovations in Information Technology (IIT)*, pages 46–51, Dubai, United Arab Emirates, 2015. IEEE.
- [11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.
- [12] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, June 2014.

- [13] N. Oliveira, P. Cortez, and N. Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer Berlin Heidelberg, Angra do Heroísmo, Açores, Portugal, 2013.
- [14] N. Oliveira, P. Cortez, and N. Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In D. Camacho, R. Akerkar, and M. D. Rodríguez-Moreno, editors, *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS)*, page 31, Madrid, Spain, 2013. ACM.
- [15] N. Oliveira, P. Cortez, and N. Areal. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73, 2016.
- [16] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.
- [17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [18] S. Sabherwal, S. K. Sarkar, and Y. Zhang. Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting*, 38(9-10):1209–1237, 2011.
- [19] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203, 2014.
- [20] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welp. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.
- [21] L. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Forecasting Journal*, 16(4):437–450, 2000.
- [22] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the 3rd International Conference on Web Search and Data Mining (WSDM)*, pages 261–270, New York City, USA, 2010. ACM.