Master Thesis
Computer Science
Thesis no: 1MSC:2013-01
September 2013

Hashtags and followers

An experimental study of the online social network Twitter

Eva García Martín

School of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona
Sweden

This thesis is submitted to the School of Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 10 weeks of full time studies.

Contact Information:
Author(s):
Eva García Martín
E-mail: eva.g.596@gmail.com

External advisor(s):
Mina Doroud, Ph.D.
Data Scientist, Twitter Inc.
1355 Market Street Suite 900 San Francisco, CA 94103
E-mail: mdoroud@twitter.com

University advisor(s):
Niklas Lavesson, Ph.D.
Associate professor of Computer Science
School of Computing/Blekinge Institute of Technology

School of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona
Sweden

Internet : www.bth.se/com
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

ABSTRACT

Context. Social media marketing is constantly gaining interest as a powerful tool, for advertisement campaigns, in order to maximize their audience to reach potential new customers. To efficiently target customers, the knowledge of social network structure and user behavior is of crucial importance. Among these online social networks, Twitter's popularity is rapidly increasing. Its key feature is to link different topics and posts by using the hashtag symbol. This particular characteristic is one of the principal causes that direct users to specific topics, and lead them to expand their network.
Objectives. In this study we investigate a correlation between hashtags and increase of followers motivated by a specific research question. The question is whether the addition of hashtags to tweets produces new followers.
Methods. We designed a controlled experiment in which we gather tweets from two types of users: users tweeting with hashtags and users tweeting without hashtags. Users tweeting with hashtags will belong to the experimental group and users tweeting without hashtags will form part of the control group.
Their statistical behavior is analyzed by conducting the non-parametrical Mann-Whitney U-test.
Results. The results of the Mann-Whitney U-Test show that the null hypothesis is rejected at confidence level 0.05.
Based on that, a correlation is shown between hashtags and followers, therefore tweets that contain hashtags are more likely to lead to a higher increase in the number of followers than tweets without hashtags.
Conclusions. This thesis contributes to describe the functionality of hahstags in the online social network Twitter. It provides an original correlational study on the use of hashtags and increase of followers. We discover that users tweeting with hashtags are more likely to increase their number of followers than users that tweet without hashtags.
This discovery opens a new research direction regarding hashtags and followers, specifically to discover which hashtags increase the number of followers and which do not.

Keywords: Experiment, Twitter, hashtags, followers, correlation.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor Dr. Niklas Lavesson. I will always be grateful for his constant support, encouragement, help, patience and advice. I am completely honored for having the possibility to present this work under his supervision, without him this thesis would have not been possible.

Second, I would like to thank Mina Doroud, my co supervisor in this thesis, for her help throughout the whole thesis.
I would also like to thank Ana B. Rodríguez, without her support I would not be writing this acknowledgments section and I would not have pursued this master. Thank you for believing in me when I would not.

I would like to thank my family for their constant support and patience: Felisa, Ángel, Mati, Matilde, Maribel, Sylvie, Elena.

Last but not least, I would like to give special thanks to one exceptional person that magically and unexpectedly appeared in my life and made me smile all my way through it, J. Eslava.

Finally this thesis is completed thanks to the support and help from my brother Jacopo, and from my closest friends that had always been there no matter what: Iker, Amaia (Areta), Nicolò, David (Zall), Ahmed, Despoina, Vanesa, Wenzo, Bob, Nader, Javi (Tronxo), Amaia (Rentería), Layla, Barby, Jenny, Manu, Cloi, Steffi, Loreto, Antonio.
Gracias.

"The biggest risk is not taking any risk…" – Mark Zuckerberg

Theoretical Background

In this chapter we are going to briefly introduce the thesis. In detail, we are going to portray the research methodology that we have applied. First of all, we explain the statistical analyses that have been conducted and the reasons for choosing specifically those analyses. Secondly, we clarify the background of our quantitative research, why we consider our experiment as quantitative research and the variables involved in the experiment. Finally we describe the purpose of our study in the last subsection.

Statistical Background

In this study we have performed several statistical analysis in order to test our data. First of all we performed a normality test. Based on the results of this test, we will choose which type of test will be performed, specifically: If the data is positively tested for normality, we will perform a parametric test. On the other hand, if the data is negatively tested for normality, we will perform a non-parametric test.

Non-parametric test means that it is used over a population of samples whose statistical behavior is not influenced by an a-priori defined parameter. Hence, if the data is not following a normal distribution (which is defined by 2 parameters) it has the characteristics of a non parametric population (Hettmansperger et al., 1998). Non-parametric test involve testing a hypothesis where there is no statement about the distribution's parameters. Many researchers prefer to avoid using non-parametric tests since their results are less reliable. In many non-parametric tests the data is ranked in order. When you transform a complete data set into ranks, a lot of information is lost. In the end, non-parametric tests have less power, and maybe a larger sample size might be needed to draw the same conclusions than with a parametric test.

The normality test that was performed was the Kolmogorov-Smirnov Goodness-of-Fit test for a single sample. Goodness-of-fit tests are used to determine whether the distribution of scores in a sample conforms to the distribution of scores in a specific theoretical or empirical population distribution (Sheskin ,2003). They are used to demonstrate if a sample is derived from a specific distribution. In our case we use the Kolmogorov-Smirnov test to determine if our samples are derived from a normal distribution.

The Kolmogorov-Smirnov test is designed to be used with a continuous variable, that is our case. It is a test of ordinal data that requires a cumulative frequency distribution to be constructed.

We are going to perform a Mann-Whitney U Test. This test is deeply described in Section IV-C. Our main hypothesis is whether two independent samples represent two populations with different median values (Sheskin ,2003). Since we are collecting users via simple random sampling, we can

say that our samples are independent. With this test we are able to see if there is a significant difference between two sets.

If the test for normality would have been positive, we would have performed the unpaired t-test.

Quantitative Research

Quantitative research refers to the systematic empirical investigation of a specific phenomena via statistical, mathematical or computational techniques (Given, 2008). The objective of quantitative research is to develop hypotheses and then support or reject them. We refer to quantitative research when the results of the studies are provided as mathematical measurements and statistical analysis.

In quantitative methods, we need to understand which factors or variables influence a specific outcome. For example, in our study, we need to understand the factors or variables that influence the increase of followers.

A variable refers to a characteristic or attribute of an individual that can be measured. We have four types of variables: Independent, dependent, control and confounding. Independent variables are the ones that are probably causing the outcome that is going to be studied. These are the main ones that we are going to analyze.

We are going to design the experiment so that we can correctly measure the effect of the independent variable on the dependent variable.

Dependent variables are the outcome that wants to be determined.

Control variables are a type of independent variables that we need to measure because they can potentially influence and affect the dependent variable. We need to be aware of their existence in order to correctly design the experiment, so that we are measuring the effect of the independent variable and not of the control variable.

Confounding variables are not measured or observed. We are aware of their existence, and that they can influence somehow the dependent variable. However, we can not control for them.

In the next subsection we describe these variables as part of our experiment.

Purpose of this thesis

The purpose statement indicates the reasons for the study and the accomplishment we want to obtain. It sets the objectives, the intention and the idea of the research that is going to be conducted.

The purpose of our experiment is to test the relationship between the use of hashtags and the increase of followers controlling for age, sex, gender, retweets and nationality.

We want to measure the increase of followers. This will be the desired outcome, therefore our dependent variable. Since we want to study the effect of hashtags on the increase of followers, hashtags will be our independent variable. Age, sex, nationality, retweets, location and user popularity are our control variables. Finally, we consider Twitter suggesting new followers and a user posting their account publicly on the Internet as confounding variables.

We are going to detail the purpose statement in Section III.

CONTENTS

Hashtags and followers: An experimental study of the online social network Twitter

Eva García Martín

***Abstract*—In this thesis we have conducted a statistical analysis of the online social network Twitter. Specifically, we focused on investigating a correlation between hashtags and increase of followers to determine whether the addition of hashtags to tweets produces new followers. We have designed a controlled experiment in which we create two groups of users: one tweeting with hashtags and the other tweeting without hashtags. We analyzed their statistical behavior by conducting a non-parametrical Mann-Whitney U-test. The results showed that there is a correlation between hashtags and followers, therefore tweets that contain hashtags produce a higher increase in the number of followers than tweets without hashtags.**

***Index Terms*—Twitter, followers, analysis, experiment, hashtags.**

I. Introduction

Twitter is a social network that publishes around 1 billion tweets every two days (Terdiman, 2012). Currently people use Twitter for massively spreading world breaking news and for publicly displaying their opinions (Wang et al., 2011).

Since 2006, when Twitter was founded, many researchers have taken advantage of the public Twitter API to build different models and applications. In the literature, we have presented several works related to Twitter. Among those, the most common research topics are sentiment analysis, event prediction and retweet prediction, as explained in section II-B.

The novelty of the platform offers a wide horizon for further research, since there has not been a lot of research conducted in it yet.

In order to provide an original contribution in this research field we conduct an experiment in Twitter. The aim of this experiment is to find out if there exists a relationship between the use of hashtags and the increase of followers. To the best of our knowledge and based on scientific literature, this experiment has not been performed beforehand. If we show a correlation between hashtags and followers, we would make an original discovery that was previously unknown.

The results could be implemented in social marketing campaigns and streams, i.e. customer targeting, advertising; to better understand what helps decrease or increase followers. We can discover how users interact with each other, who they are influenced by and how they make decisions (Pentland, 2011; Pentland et al., 2012; Nikolov, 2012).

Based on this contribution, our results could benefit the research community by providing further insights on Twitter's structure and behavior. It opens a new research direction regarding hashtags and followers. More specifically, and related to the computer science field, the discovery that hashtags are correlated with the increase of followers encourages and motivates researchers to continue investigating in order to find out specifically which hashtags do attract followers. They can apply Natural Language Processing techniques to see which type of hashtag produces this increase of followers. Moreover, they can apply machine learning techniques to group hashtags into different types and see which groups are related to the increase of followers. Finally, it also opens a new research idea, the possibility of creating a predictor model in order to predict the increase of followers.

Our results could also benefit marketing companies since they will know one of the reasons that makes users gain followers. If a user has more followers his or her audience is bigger. Lastly, the contribution could benefit users, since they will have more knowledge on how to spread their information more widely and make that information reach more people.

The document is structured as follows: In Section II we present the background with the Twitter-related terminology and its related work. In Section III we detail the purpose statement. In Section IV we deeply describe the research methodology, the experimental design, data collection and statistical analysis. In Section V we present the raw and analyzed results. Last but not least in Section VI we summarize the conclusions. Lastly, in Section VII we provide further perspectives of future work.

II. Background

A. Terminology

Tweets are short 140 character messages. To tweet is the action of posting a tweet in Twitter. Twitter users tweet to show to the public their thoughts about a specific matter, to post breaking news or information about topics they are interested in (Mathioudakis and Koudas, 2010). Mentions are used for connecting users. If user U_a wants to mention user U_b, U_a posts the character @ followed by U_b username, e.g. *@username*. Retweet is a particular case of mentioning. A retweet is the action of a user tweeting the tweet of another user. We have two types of retweets: retweeting the complete tweet, and retweeting and adding something more to the tweet. The second type of retweet includes the text indicator *RT*.

When users want to categorize their messages into specific topics, they add the hash symbol to the topic. For example *#computerscience*. The hash symbol plus the topic or word is called a hashtag (Wang et al., 2011).

Users connect with each other by following each other. If U_a follows U_b, then U_a receives in his or her timeline all the tweets from U_b. In this case, U_a is a follower of U_b and U_b is a followee of U_a.

Eva García Martín is with the Department of Computer Science, Blekinge Institute of Technology, Karlskrona, 371 79 Sweden e-mail: evgb12@student.tbh.se .

For gathering data from Twitter, there is an open API available for developers[1]. This API makes it easy for the developer to send requests to Twitter to ask specific information (Makice, 2009). We are interested in two main requests:

- GET statuses/sample
- GET users/lookup

The first request belongs to the Streaming API[2]. The streaming API is a collection of APIs that give the developer access to the global Twitter stream. The main difference with a normal API request, is that in the streaming API the HTTP connection is persistent. A normal API request would send a request and get a response every time we would want some data. In this case the connection is always open and the end user is constantly receiving data. The benefits of using the streaming API are the small restrictions and limits we have in order to gather data. With the first request shown above we obtain a random sample of the global Twitter stream.

On the other hand, we have the REST v1.1 API[3]. This follows a normal API request. The majority of requests are under this API. For instance, if we want to obtain some specific data of certain users or search for specific keywords; the requests will be REST requests. The second request belongs to the REST v1.1 API. We use this request to gather all the information from 100 users per request. The downside of this API is that there is a very strict limitation in the number of requests we can send. In order to allow Twitter to monitor the number of requests we make, we need to follow an authentication protocol, *Oauth* [4]. This way we obtain information in the name of the user and the registered twitter application.

B. Related Work

Online social networks are becoming more and more popular nowadays. There have been several models built upon them (Kumar et al., 2010; Mislove, 2009). Many marketing companies are increasing their interest in social networks as a way of advertising their products to the end user. The reason behind this is the increase of active users using different social networks, therefore targeting customers can be performed in a much accurate and efficient way (Lerman and Galstyan, 2008; Granovetter, 1973; Richardson and Domingos, 2002; Domingos and Richardson, 2001).

The results of an initial search indicate that there are three main areas of research in relation to online social networks: Sentiment analysis, Prediction and Graph models.

There is a great interest in building a model that shows the growth of a specific social network. The most common model is the preferential attachment, created by Barabási and Albert (1999) and tested with positive results by Newman (2001) and Jeong et al. (2003). Preferential attachment states that new links tend to form towards already popular links. Popular links are users that have a higher number of followers compared to the average Twitter user. Not only celebrities are popular users.

On the other hand, Lang and Wu (2011) built a growth model of the social network *Buzznet*[5] looking for evidence of preferential attachment. They unexpectedly discovered that *Buzznet* follows an anti-preferential attachment model; therefore high-degree nodes create edges to low-degree nodes. This means that users that we expect to have a higher number of followers than followees (high-degree nodes), such as celebrities, end up having a lower number of followers than followees(low-degree nodes).

Mislove et al. (2008) discovered the reasons behind link formation in Flickr[6]. They discovered that links are usually created by users who already have many links. We could match this with the previous discovery from Lang and Wu (2011). If we take a user that has many links, i.e. a celebrity, the results of the study from Mislove et al. (2008) demonstrate that it will tend to create new links with more users; therefore the difference between followers and followees would be higher. This agrees with the claim from Lang and Wu (2011).

In addition to graph models, several studies such as Sakaki et al. (2010); Ritterman et al. (2009); Qiu et al. (2011) have been made on event prediction using Twitter as a source. Based on hashtags or trends they try to predict the appearance of future events.

Finally, we observe that Twitter has been used by many researchers to detect the sentiment from different users to certain products. This is called sentiment analysis. It is useful for marketing companies so that they can manage to correctly advertise their products (Bifet and Frank, 2010; Pak and Paroubek, 2010; Wang et al., 2011; Thelwall et al., 2011; Kouloumpis et al., 2011; Go et al., 2009; Diakopoulos and Shamma, 2010).

Equally important, there has been a lot of specific research conducted in the Twitter area; interesting and important investigations, such as the ones held by Suh et al. (2010); Naveed et al. (2011); Macskassy and Michelson (2011); Ye and Wu (2010); Tsur and Rappoport (2012). These investigations focus on the reasons behind information spread in Twitter. For instance, Mendoza et al. (2010); Boyd et al. (2010); Macskassy and Michelson (2011); Petrovic et al. (2011) have thoroughly investigated how to predict retweets.

Yang and Counts (2010) have conducted a study where they predict whether a post will get mentioned or not by building a very extensive diffusion network. They also measured the speed of a certain post, by analyzing how fast a tweet is retweeted. Related to this, Huberman et al. (2008) and Java et al. (2007) study the follower and followee network.

While several studies focus on retweet prediction, there is a study by Boyd et al. (2010) that focuses on the motivations behind a retweet, rather than predicting a new one, by interviewing Twitter users.

Cha et al. (2010); Romero et al. (2011) show that users with active followers are more likely to be retweeted, however the fact of having more followers does not necessarily lead to the user being more popular. While several studies by Yang and Counts (2010); Kwak et al. (2010); Cha et al. (2010) focus

[1] https://dev.twitter.com/
[2] https://dev.twitter.com/docs/streaming-apis
[3] https://dev.twitter.com/docs/api/1.1
[4] https://dev.twitter.com/docs/auth/oauth
[5] www.buzznet.com
[6] www.flickr.com

on the context of the tweet and the user who has posted that tweet; Naveed et al. (2011) center their research on the precise content of the tweet applying Natural Language Processing. In the work done by Suh et al. (2010), an interesting discovery was the relationship between hashtags and retweets. Tweets containing hashtags and URLs are more likely to be retweeted. To the best of our knowledge this work is the most closely related to the topic of hashtag usage and prediction.

If a tweet is retweeted several times then the information in that tweet is diffused between more people. If the source of the tweet has more followers then the information is not only more probable to be retweeted (Suh et al., 2010) but is also spread to a larger number of people. The reason behind this argument has already been explained. If U_a has more followers than U_b, then the audience of U_a is bigger than the audience of U_b (Lerman, 2007). Based on this, it is important to know why certain tweets get more retweets than others and why specific users gain more followers than others.

Since all the research in this area has been made on prediction of retweets, knowing the reasons behind the increase of followers could lead to useful information about how information is spread trough Twitter.

III. Purpose Statement

The purpose of this experimental study is to investigate the potential correlation between hashtag usage and the increase or decrease of followers; controlling for retweets, user characteristics and user popularity. Our independent variable is hashtags. The dependent variable is the change in the number of followers. In this experiment we control for retweets, user popularity, user characteristics, *the million follower fallacy* and new mentions. We take into account the effect of retweets since one of the key reasons for new followers is that U_a starts to follow U_b because they had a friend in common that retweeted a tweet from U_b.

Due to Avnit (2009) and Cha et al. (2010), *The million follower fallacy* is a term used when U_a starts following U_b just for etiquette or for being polite to follow someone that already follows you. As for user characteristics, we sample random users so that we have a set of users that is representative of the whole population, with random ages, genders, nationalities and languages. We also have to randomly sample famous users, since usually these users get their number of followers increased at a much faster pace than the rest of users. Famous users are those ones that are widely known to the public, such as celebrities, famous sport players and so on. They have a lot of links towards them from other users. Finally, we also consider the fact that U_a mentions U_b in his or her tweet. The username of U_b is publicly being seen by all the followers of U_a, therefore it increases the chance of U_b of getting new followers.

Finally, as for confounding variables, we consider these two:

- A user posting his or her user account on a public place in the Internet.
- Twitter sometimes suggests a user to follow new users.

At this point, we have disclosed which control and confounding variables we are going to take into consideration. The way these variables are going to affect our experiment design and how we are going to control them is explained in section IV.

IV. Research Methodology

In order to investigate the possible correlation between the use of hashtags and the increase in followers, we propose a specific research question. Our research question is whether the addition of hashtags to tweets produces new followers.

Our motivation for this research question is hashtags. In Twitter, if a hashtag becomes popular, you can see that hashtag on a specific section of Twitter's timeline, available to all users. If you click on any of these trendy hashtags, you can see all tweets with such hashtag. Therefore, users tweeting with trendy hashtags are more likely to be visible on Twitter.

Moreover, you can search for a hashtag even if it is not trendy, e.g. a hashtag related to an event. Hence, we want to discover if users that search tweets based on a specific hashtag, actually start to follow the authors of those tweets. For that reason, we hypothesize that there is an increase in the number of followers for users tweeting with hashtags.

A. Experimental design

The aim is to perform a controlled experiment in order to answer the stated question. The main characteristics that we want to achieve with this design are randomization, non biased choices and a clear distinction between users tweeting with hashtags and without hashtags.

We created two independent groups of users: a control group and an experimental group. The control group is formed by users that have tweeted in the moment of the gathering without a hashtag. On the other hand, the experimental group consists of users that have tweeted with a hashtag in the moment of the gathering. Every user is picked following the same random procedure, that will be explained in section IV-B. The only difference between both groups is the hashtag usage.

The reason for this specific design is the fact that we need to control several variables; such as age, gender, nationality and retweets. The way we have controlled these variables is collecting users in the same way, therefore we ensure the existence of all type of random users in both groups, all of them presenting very similar characteristics of age, sex, etc. For instance, if the increase of followers is due to retweets, this effect is going to be presented in the same way in both groups.

We achieve a non-biased choice of users by picking random users, 24 hours a day during a whole week. This way the study is not biased by the location, language or age.

After obtaining the information from every user, we analyze the increase of followers between both groups. We have performed several statistical analyses for this procedure, that will be explained further on in section V.

If the statistical analyses show that there is a significant difference in the number of followers, this change in the number of followers is probably caused by hashtags. With this experiment we are able to answer our research question.

B. Data collection

Tweets and users are obtained from Twitter using the Twitter API. From the Twitter API we used two requests:

- GET statuses/sample
- GET users/lookup

The first one returns a small sample of random tweets from Twitter's public timeline. It does not depend on a location or language. We use this request to gather the different users for both the control and the experimental group.

The second request returns information from a hundred users at the most. Information such as number of followers, number of tweets etc. We use this last request to update the information from each user.

There are some special considerations we need to take into account. The Twitter API has a limit in the number of requests the developer can make. Every update could be of a maximum of 18,000 users every 15 minutes. We gathered the number of users based on this limit.

We developed a Python script to make the different requests to the Twitter API. We used an open source package called Twython, from developer Ryan McGrath [7]. With Twython we were able to make all the requests to Twitter in an efficient way.

Our main target is to measure the increase of followers for the different users from the control and experimental group. For collecting the users, we analyze each tweet, from the random sample of tweets that we extracted. For every tweet that does not contain a hashtag we save its author (user) as a member of our control group. On the other hand, for every tweet that does contain a hashtag we save the author as a member of our experimental group. In summary, all users are picked using the same random request, the only difference is the presence of a hashtag in the user's tweet.

For each tweet we are going to save the following data: Username, number of tweets, number of followers, number of followees, date and time of gathering, the tweet, number of times that tweet has been favored and number of times the tweet has been retweeted. We store all this information in a database. We have two different tables inside the database, one for the control group and a second one for the experimental group.

The next step is to discover if the users gain, maintain or loose followers. Every six minutes during half an hour we update all the users' information. After these six minutes, we gather again new users for both control and experimental groups. We then update them every six minutes for half an hour. And so on and so forth during a period of one week.

In the end, we acquired a total of 1,546,742 users, summing both groups. 1,193,569 users tweeting without hashtags and 353,173 tweeting with hashtags. As we can see, the number of users from the control group is much higher since they were always more users tweeting without hashtags than tweeting with hashtags.

We have made our calculations based on a data set of approximatively 1.5 million users. For each one of these users there are 5 updates of information.

[7] https://github.com/ryanmcgrath/twython

The data collection procedure is not completely accurate. The reason is because on rare occasions when updating the users, Twitter's service was unavailable due to overcapacity. When this happened 100 users were not updated. However, this is not too relevant since we are making a total of 5 updates, so the last update is always available in order to calculate the increase or not on the number of followers. Furthermore, we can notice that among a total data set of 1.1 million users, only 20 exhibited an anomalous behavior in the increase on the number of followers, always coming from the request statues/sample. However a user having a fast increase of followers is possible in Twitter, the code has been thoroughly checked, and Twitter has not reported that the request status/sample gives erroneous information. As a consequence we did not remove those instances from our data set, even if they clearly do not represent a visible portion of the examined population and hence they do not influence its overall behavior.

C. Data analysis

We provide a statistical analysis on our data to discover if there is a difference between users tweeting with hashtags and users tweeting without hashtags.

First of all, we tested both our data sets for normality, with the Kolmogorov-Smirnov test (Kolmogorov, 1933). The null hypothesis is that the distribution of samples that represent the change in the number of followers follows a normal distribution.

The hypothesis is rejected, therefore the normality test shows that with a confidence level of 0.05 the samples are not drawn from a normal distribution.

Since our data is tested negative for normality, we are going to perform a non-parametric test.

The chosen test is Mann-Whitney U-test. The hypothesis is whether two independent samples represent two populations with different median values. Developed by Mann and Whitney (1947) and thoroughly explained in the book by Sheskin (2003). We are going to base all our computations on the latest one.

Our way of collecting data, explained in section IV-A, is known as simple random sampling with replacement. Each user is chosen from a large data set, in this case Twitter's public timeline, randomly and by chance, therefore each user has the same probability of begin chosen. This procedure ensures to gather independent sample values (Moore and McCabe, 2011; Cochran, 1977, 2007; Starnes et al., 2010; Given, 2008).

In this case, group 1 is the control group, and group 2 the experimental group.

Null hypothesis:

$$H_0 : \theta_1 = \theta_2$$

The median of the populations of Group 1 and Group 2 are equal. This can be also expressed as : The means of the ranks of the two groups are equal ($\overline{R_1} = \overline{R_2}$).

Alternative hypothesis :

$$H_1 : \theta_1 < \theta_2$$

The mean of the ranks of the experimental group is bigger than the mean of the ranks of the control group. It represents that the median of the population of group 1 is less than the median of the population of group 2. This is a directional alternative hypothesis, that has to be evaluated with a one-tailed test.

The other possible alternative hypothesis would be the following one:

$$H_1 : \theta_1 > \theta_2$$

However, we have chosen the first one, since, as Sheskin (2003) states, we need to choose the alternative hypothesis based on which mean of the rank is higher. In this case, $\overline{R_1} < \overline{R_2}$, therefore the chosen hypothesis is that $\theta_1 < \theta_2$.

This test is conducted by comparing $U = min\{U_1, U_2\}$, to the critical value for the confidence interval, which is tabulated in literature. The parameters U_1 and U_2 are derived according to eq. 12.1 & 12.2 of reference (Sheskin, 2003), shown in eq. 1.

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{(n_1+1)n_1}{2} - \sum R_1 \\ U_2 &= n_1 n_2 + \frac{(n_2+1)n_2}{2} - \sum R_2 \end{aligned} \tag{1}$$

Where n_1 and n_2, R_1 and R_2 are respectively the sizes and ranks of the first and second group.

However, since we have a very large data set, there is no available critical value tabulated. Therefore, according to Sheskin (2003) we perform the normal approximation of the Mann-Whitney U statistic for large sample sizes.

In this test, the absolute value of parameter z,

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \tag{2}$$

is compared with the critical value for a confidence level of 0.05 shown in Altman's Z-Score table (Altman, 1968). This table displays the values of the normal distribution parametrized for different confidence levels.

If the absolute value of z is bigger than the critical value, for a one-tailed test, then the null hypothesis is rejected.

If the null hypothesis is rejected, then the median of group 1 is lower than the median of group 2.

D. Validity Evaluation

We have taken into account several validity threats in our research.

1) Internal: To ensure internal validity we have to be aware of all the factors that affect the increase of followers, our dependent variable. The experiment has been designed carefully in order to ensure that the measured effects are going to be influenced only by hashtag usage and no other control variable. As was explained in the experimental design section, we control for retweets, age, sex, gender, retweets and nationality by guaranteeing the same types of users on both our experimental and control group.

We have to be aware of what we can and can not show with our experiment. We work with a natural experiment rather than a true experiment. The reason behind this is because we are observing certain individuals, in our case users; and how some of their actions, in this case tweeting with hashtags; lead to a change in the number of followers. The users, already form part of one or other group, that is beyond our control since we are merely observers. We are dividing random people based on their use or not of hashtags, but we are not aware of the existence of other possible reasons for the increase in followers. In the end they are users and there is a limitation in the amount of information we can obtain. Therefore, with our results, we can not claim a cause-effect result. We can not know if the increase of followers is only achieved due to hashtags.

However, we have designed the experiment thoroughly so that our result is as closest as possible to a cause-effect claim. We have randomized in all possible aspects and made non-biased decisions of the picked users. What we can show with this experimental design is a correlation between hashtags and increase of followers. Based on the results, we would be able to present strong evidence of a possible connection between hashtags and followers.

2) External: For ensuring generalization in the data collection, we have tried to avoid bias choices in all possible cases. For that reason we obtain users from different periods of the day (during 24 hours), during a complete week. As has been explained these users are chosen randomly. Hashtags are also picked randomly, therefore no specific hashtags (e.g trending topics, specific category...) are used. To ensure context validity, we must highlight that our study is representative of the real population since we are gathering real data from real users from Twitter.

3) Construct: We need to show that we are really computing the increase of followers. For that reason, when building the script we made some deep tests that show that we are correctly measuring the number of followers. In this study is not difficult to assure construct validity, since we are just measuring a number through time, and Twitter provides that information easily.

4) Statistical Conclusion: To show that our conclusions are founded on an adequate analysis of the data we need to ensure that we are using the correct statistical analysis.

The main statistical test, Mann-Whitney U-Test has been chosen after checking that the data does not follow a normal distribution. As is stated in Sheskin (2003), if we want to measure the median between two groups, of independent samples and where the distribution is unknown or not normal, we should use the Mann-Whitney U-Test. At every moment we have based all our choices of statistical analysis on the book from Sheskin (2003).

V. Results and Analysis

We show a summarized description of the obtained data in Table I. Group 1 comprises users from the control group (users

TABLE I
SUMMARY OF GATHERED DATA

	Group 1	Group 2
Total Users	1,193,569	353,173
Mean of the difference of followers	2.177567	5.514176
Mean of the difference of followees	1.031156	2.617641
Mean of the total number of tweets	13,342.85	11,735.16

tweeting without hashtags) and group 2 comprises users from the experimental group (users tweeting with hashtags).

In Table II we show the top 20 users that have tweeted with hashtags. Considering top users as the ones that have the highest increase in the number of followers. For these users, we show the hashtags that were used by each user in the moment they were picked and the increase in the number of followees. In Table III we show, for the same top 20 users, the time they tweeted with that hashtag, the total number of followers, the total number of followees, the total number number of tweets, and the number of hashtags that were used in the tweet.

TABLE II
20 TOP USERS THAT HAVE THE HIGHEST INCREASE OF FOLLOWERS

Username	Followers increase	Hashtag(s) used	Followees increase
RealTonySimon	176023	#Twitter, #TonyRocha	1
ah3761	114343	Arabic language	0
WiseTheGoldBoy	41707	#buendia, #ff	16
MLMGods	27424	#Rippln	2555
multiverse60	17048	#TRUEF4F	16815
oOoSaudoOo	16680	Arabic language	16240
Meditation1Deeb	15575	Arabic language	4244
XL123	14144	#XLAwards	2
RealTonyRocha	13162	#OnlineBusiness, #TonyRocha	1
Rassd_News	12979	Arabic language	0
FABAClD	12934	#MTVHottest	88
KavalonThatsMe	12265	#TFR, #Follow	4246
NBmusicpromo	12247	#promo	11298
RT_RT511	10880	Arabic language	840
fferawanti	10378	#Ajak_teman	2
D3M79	9550	Arabic Language	-10366
KutipanDAHSYAT	9218	#RECOMMENDED	0
GirIFeelings	8685	#hurry	-4
SindiranJenius	6134	#TanyaYuk	0
Shorouk_News	6089	Arabic language	0

An interesting fact that can be observed from Table II, is that 35% of the top users have tweeted with hashtags in Arabic language.

In Tables IV and V we show the 20 worst users that produce the highest decrease in the number of followers.

For these 20 users, we show the same details as before: followers increase, followees increase, total number of tweets, total number of followers, total number of followees, number of hashtags and finally the hashtags that they tweeted with.

If we look into the differences between the worst and top users, we can see that while the top users have tweeted with one or two hashtags, the worst users have tweeted with two, four even with seven hashtags. Furthermore, there is a decrease in the number of followees for the worst users while there is no such decrease presented in any of the top users.

In order to show possible additional relationships between the features from the different users, we provide some graphical representations.

Figure 1 displays a comparison between the increment of followers and the increment of followees as a function of the total number of tweets posted. This comparison is made

TABLE III
20 TOP USERS THAT HAVE THE HIGHEST INCREASE OF FOLLOWERS

Total Tweets	Date gathering	Total Followers	Total Followees	Number hashtags
48214	2013-07-20 02:14:05	207689	70	2
14224	2013-07-20 14:40:46	124725	2	1
100713	2013-07-23 22:52:01	49552	1306	2
183	2013-07-23 14:11:32	270111	83153	1
64017	2013-07-23 19:30:45	17048	16815	1
5269	2013-07-18 19:39:22	16684	16240	1
7051	2013-07-20 23:22:47	15577	4249	1
59031	2013-07-20 21:50:39	395716	49008	1
278898	2013-07-20 19:51:12	1794686	699	2
171916	2013-07-24 05:40:42	882520	4	5
27442	2013-07-21 02:52:07	12935	88	1
266949	2013-07-21 20:26:03	335033	288228	2
68682	2013-07-22 20:37:34	12248	11299	1
62252	2013-07-22 18:51:22	15423	16967	3
34358	2013-07-24 02:15:58	19769	227	1
27363	2013-07-19 22:53:02	703595	683290	1
6385	2013-07-21 18:40:30	704787	1	1
587	2013-07-21 01:07:38	345653	55	1
137	2013-07-18 16:02:17	99969	0	1
285315	2013-07-18 14:17:57	1039074	3	2

TABLE IV
20 WORST USERS THAT HAVE THE HIGHEST DECREASE OF FOLLOWERS

Username	Followers decrease	Hashtag(s) used	Followees increase
ReneeGiraldy	-55469	#Winner, #moved , #blog, #Wordpress	57
TDF_Pronostics	-42369	#tdf	0
finestbieber	-36737	#SelenaK104	-185
SVorreiter	-14347	#RT #FB	-12425
moanshoran	-10414	#mtvhottest	-398
theHAIRspy	-8262	#FashionFiles	0
TopRockNews	-5931	#TRNBackstage	-1
DunhillTot	-5592	Arabic language	430
VegasJai	-5476	#300aDay #Vegas, #StripClub, #SapphireLV	0
maclove_lv1	-5340	#JnfVr	0
FracesFelices	-3939	#RT	-1308
lustfulouis	-3436	#AtSunsetFollowSpree	-1823
theluanz	-3430	#luanz	-1
IAMTONYNEAL	-3388	#GROUNDBREAKING	0
RT2GAINALWAYS	-3220	#TFBJP, #FOLLOW-BACK, #RT2supergain, #TEAMFOL-LOWBACK, #OPENFOLLOW, #FOLLOW, #F4F	715
AriesWeAre	-2959	#Aries	-25
TravelSquire	-2958	#books #Architecture	-1
zialldior	-2856	#mtvhottest	-460
ALOMMAHQ8	-2453	Arabic Language	6
junqnjiw	-2338	#openfollow	-2318

TABLE V
20 WORST USERS THAT HAVE THE HIGHEST DECREASE OF FOLLOWERS

Total Tweets	Date gathering	Total Followers	Total Followees	Number hashtags
19053	2013-07-19 10:20:50	84132	8025	4
6260	2013-07-18 16:02:17	50278	158	1
0	2013-07-20 05:41:11	1	1	1
40341	2013-07-18 12:31:00	90266	15	2
0	2013-07-20 23:22:55	3	13	1
1431	2013-07-18 19:39:11	11632	148	1
194729	2013-07-20 03:58:02	79891	993	1
7938	2013-07-20 00:29:17	12890	5287	1
288040	2013-07-20 02:14:05	39133	433	4
244245	2013-07-20 00:29:10	3817	9	1
0	2013-07-23 19:22:41	0	15	1
1	2013-07-22 12:20:51	2	79	1
93513	2013-07-20 03:58:06	111732	38	1
175815	2013-07-20 02:14:12	231491	1344	1
23037	2013-07-19 10:20:04	4273	1587	6
47295	2013-07-22 20:37:35	117120	65429	1
19695	2013-07-24 20:13:31	161737	6454	2
0	2013-07-20 16:23:30	0	0	1
4468	2013-07-18 16:02:17	99969	0	1
0	2013-07-25 10:25:18	0	0	1

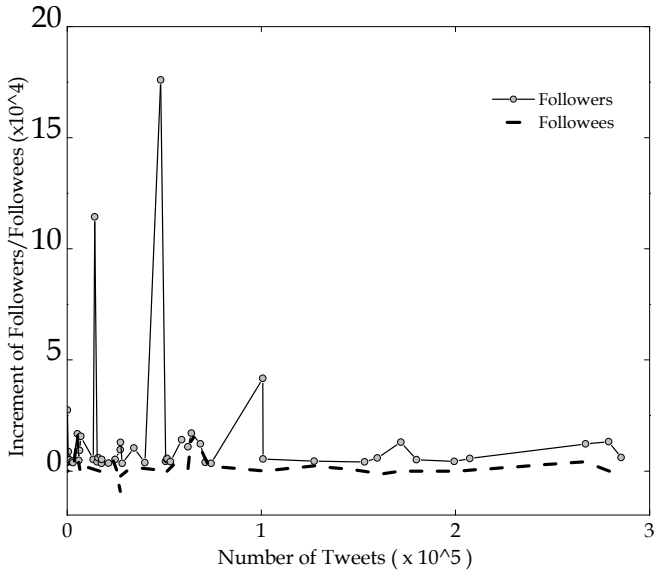

Fig. 1. Increase of followers and increase of followees as a function of total number of tweets.

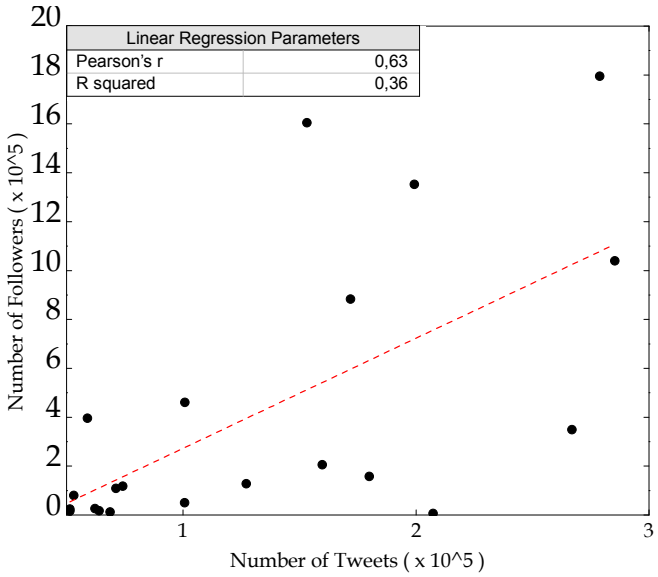

Fig. 2. Total number of followers against total number of tweets.

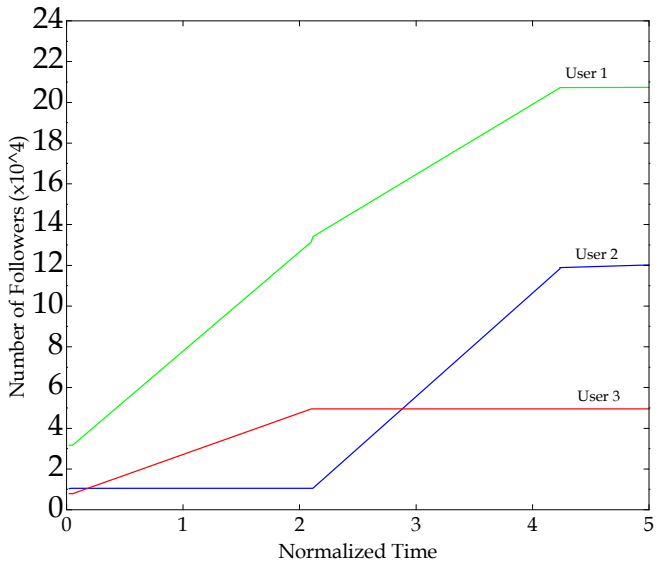

Fig. 3. Increment of followers for the top 3 users against normalized time.

between the top 20 users tweeting with hashtags. The top users are the ones with the highest increase on followers. These users then have been ordered as function of the number of the posted tweets. Two significant remarks can be pointed out over this figure: First of all, as the number of tweets increases the increment of followers/folowees is almost constant, not being influenced by the increase of the number of tweets. Secondly, we can notice that the envelope of the followers and followees curve is quite similar.

In figure 2 we show the relationship between the total number of followers and the total number of tweets. This plot is also made for the top 20 users and ordered as a function of the number of tweets. In order to provide an analytical description of the curve trend, we have performed a linear regression statistic. The Pearson correlation coefficient value, r=0.63, shows that the average trend of the total number of followers is to increase linearly. However, it must be noticed that the curve poorly fits the experimental data. In fact, the high fluctuations of the number of followers produce a low value of R^2.

Figure 3 displays the variation as a function of time of the number of followers for the three best users. These users, represented by $User_1$, $User_2$ and $User_3$ are the top 3 users with the highest increase on the number of followers that tweeted with hashtags. The plot clearly shows that the increment of followers does not occur smoothly and constantly, but abruptly variates and saturates. For each user, we have plotted the number of followers for each time that user has been updated. Hence, in order to provide a continuous timeline

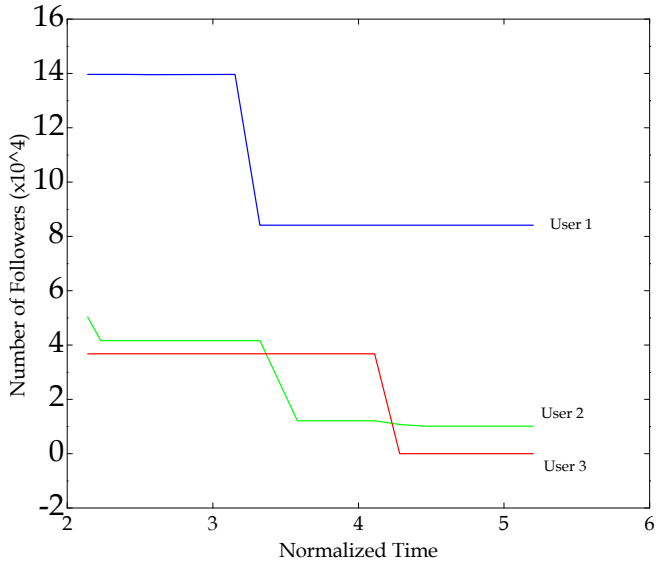

Fig. 4. Decrease of followers for the worst 3 users against normalized time.

TABLE VI
KOLMOGOROV - SMIRNOV TEST ON CONTROL AND EXPERIMENTAL GROUP.

Value	Control	Experimental
P-Value	$< 0.16e^{-17}$	$< 0.16e^{-19}$
Confidence Level	0.05	0.05
Final Result	Rejected: not normal	Rejected: not normal

we normalized the value of the date and hour of the update to the first one, to achieve a vector of ordinate numbers that could be plotted.

Finally, Figure 4 shows the same information as Figure 3 but in this case for the 3 worst users. We plot the number of followers for each time the user has been updated. We can see the same behavior as before, the decrease of followers does not occur smoothly but abruptly.

All these plots and figures are based on the top and worst users, not the complete dataset of users.

Moving on to the statistical tests, the first test that we performed was a Kolmogorov-Smirnov test for normality. The results are shown in Table VI.

The null hypothesis is: "the values measuring the difference of followers from the control and experimental group are drawn from a normal distribution". As we can see from table VI, the p-value is less than 0.05 in both groups therefore the null hypothesis is rejected for both groups. Based on this we can conclude that the difference in the number of followers for both the control and the experimental group does not follow a normal distribution.

The second test that we performed was the Mann-Whitney U-Test. The numerical values used to compute this test are presented in table VII.

We performed the Mann-Whitney U-Test, obtaining a value of:

$z = -115.0944$

TABLE VII
VALUES FOR COMPUTING THE MANN-WHITNEY U-TEST

Parameters	Group 1	Group 2
Group Size	n_1 = 1,193,569	n_2 = 353,173
Group Rank	$\sum R_1$ = 896,244,124,590	$\sum R_2$ = 299,962,056,062
U parameter	U_1 = 237,596,295,512	U_2 = 1.8394 e^{11}

TABLE VIII
MANN-WHITNEY U-TEST BETWEEN CONTROL AND EXPERIMENTAL GROUP

Value	Result
P-Value	$< 2.2e^{-16}$
Confidence Level	0.05
Final Result	Null hypothesis rejected

The critical value for a confidence interval of 0.05 for a one-tailed test is $z_{0.05} = 1.65$, obtained from the Z-Score standard table (Altman, 1968). The absolute value of z is bigger than $z_{0.05}$ therefore the null hypothesis is rejected and the alternative hypothesis is supported.

Based on these results, the median value of the control group is lower than the median value of the experimental group. Hence, we conclude that the increase in the number of followers for the users tweeting with hashtags is significantly higher than the increase in the number of followers for the users tweeting without hashtags.

To validate our analysis we have performed the test in R, obtaining the results shown in table VIII.

The achieved results allow us to answer the research question (RQ) showing that there is a correlation between tweeting with hashtags and the increase in the number of followers.

VI. CONCLUSIONS

The goal of our study was to determine whether the addition of hashtags to tweets produces new followers. For this reason we performed an experiment. In this experiment we gathered random users that tweeted with hashtags and random users that tweeted without hashtags; for a period of 7 days. Then we preprocessed the data so that we computed, for each user, the difference in the number of followers in 30 minute slots. We ended up with a total of 1,546,742 users. Having collected the experimental data, we wanted to see if the change in the number of followers followed a normal distribution, therefore we conducted the Kolmogorov-Smirnov test. This test showed that the samples of both groups are not drawn from a normal distribution.

Given that, we decided to perform the Mann-Whitney U-Test on both data sets (users using hashtags and users not using hashtags); to test if the mean in the difference of followers of the users tweeting with hashtags is significantly different than the mean from the users tweeting without hashtags. We obtained positive results, that show that users tweeting with hashtags have a significant increase in the number of followers

compared to the ones tweeting without hashtags. The results of this test therefore answer our research question: There is in fact a correlation between hashtags and followers.

These results can significantly help marketing companies to correctly target their customers. Knowing the reasons behind increase of followers can lead them to target more accurately their audience.

This work provides additional contribution to the research field of online social networks by presenting the first correlational analysis between hashtags and followers in Twitter. This may be valuable for researchers, since it has opened a new research direction, that can be continued in order to investigate new facts, such as the type of hashtags that produces the increase of followers.

VII. Future Work

This work opens new horizons for the research over the reasons behind the increase of followers. We believe that this thesis opens two main research directions. First of all, we suggest that an interesting work could be made to discover which hashtags attract new followers and which do not. Right now we know that hashtags and increase of followers are correlated, but we do not know precisely the type of hashtags that are responsible for this phenomena. For that reason, one option could be to apply machine learning techniques to group hashtags into different types, and discover if there exists specific type of hashtags that produces an increase of followers. Moreover, another option could be to apply Natural Language Processing techniques in order to morphologically analyze each hashtag.

As a second future investigation, we could conduct an experiment similar to this one. The goal would be to investigate if there is a correlation between the increase of followers and retweets, one of our control variables explained beforehand. The factor that we investigate here is when U_a tweets something, then U_b retweets that tweet; and a third user, friends with U_b sees that tweet; and for that reason starts to follow U_a.

References

D. Terdiman. (2012, Oct.) Report: Twitter hits half a billion tweets a day. [Online]. Available: http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1031–1040.

A. Pentland, "Honest signals: how social networks shape human behavior," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 583–584.

A. S. Pentland, P. Hinds, and T. Kim, "Awareness as an antidote to distance: Making distributed groups cooperative and consistent," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1237–1246.

S. Nikolov, "Trend or no trend: A novel nonparametric method for classifying time series," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 1155–1158.

K. Makice, *Twitter API: Up and running*. Oreilly & Associates Incorporated, 2009.

R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link Mining: Models, Algorithms, and Applications*. Springer, 2010, pp. 337–357.

A. E. Mislove, *Online social networks: measurement, analysis, and applications to distributed information systems*. ProQuest, 2009.

K. Lerman and A. Galstyan, "Analysis of social voting patterns on digg," in *Proceedings of the first workshop on Online social networks*. ACM, 2008, pp. 7–12.

M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.

P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.

H. Jeong, Z. Néda, and A.-L. Barabási, "Measuring preferential attachment in evolving networks," *EPL (Europhysics Letters)*, vol. 61, no. 4, p. 567, 2003.

J. Lang and S. F. Wu, "Anti-preferential attachment: If i follow you, will you follow me?" in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 339–346.

A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *Proceedings of the first workshop on Online social networks*. ACM, 2008, pp. 25–30.

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*. ACM, 2010, pp. 851–860.

J. Ritterman, M. Osborne, and E. Klein, "Using prediction markets and twitter to predict a swine flu pandemic," in *1st international workshop on mining social media*, 2009.

L. Qiu, H. Rui, and A. Whinston, "A twitter-based prediction market: Social network approach," *ICIS 2011 Proceedings*, 2011.

A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *Discovery Science*. Springer, 2010, pp. 1–15.

A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, 2010.

M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 406–418, 2011.

E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *ICWSM*, 2011.

A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, pp. 1–12, 2009.

N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1195–1198.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 177–184.

N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," 2011.

S. A. Macskassy and M. Michelson, "Why do people retweet? anti-homophily wins the day," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 209–216.

S. Ye and S. Wu, "Measuring message propagation and social influence on twitter. com," *Social Informatics*, pp. 216–231, 2010.

O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 643–652.

M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 71–79.

D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 2010, pp. 1–10.

S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter," *Prof. of AAAI on Weblogs and Social Media*, 2011.

J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," *Proc. ICWSM*, 2010.

B. Huberman, D. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *Available at SSRN 1313405*, 2008.

A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 56–65.

M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *4th international aaai conference on weblogs and social media (icwsm)*, vol. 14, no. 1, 2010, p. 8.

D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Machine learning and knowledge discovery in databases*. Springer, 2011, pp. 18–33.

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*. ACM, 2010, pp. 591–600.

K. Lerman, "Social information processing in news aggregation," *Internet Computing, IEEE*, vol. 11, no. 6, pp. 16–28, 2007.

A. Avnit, "The million followers fallacy," *Pravda Media Group*, 2009.

A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dellIstituto Italiano degli Attuari*, vol. 4, no. 1, pp. 83–91, 1933.

H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

D. S. Moore and G. P. McCabe, "Introduction to the practice of statistics," *AMC*, vol. 10, p. 12, 2011.

W. Cochran, "G.(1977); sampling techniques," *New York, Wiley and Sons*, vol. 98, pp. 259–261, 1977.

W. G. Cochran, *Sampling techniques*. John Wiley & Sons, 2007.

D. S. Starnes, D. Yates, and D. Moore, *The practice of statistics*. Macmillan, 2010.

L. M. Given, *Qualitative Research Methods*. Sage, 2008, vol. 2.

E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The journal of finance*, vol. 23, no. 4, pp. 589–609, 1968.