UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science

José Santos

Analysis of Retweeting Behavior Using Topic Models

Master's Thesis
(30 eap)

Supervisor: Professor Marlon Dumas

Author: …….. "….." November 2011
Supervisor: …… "….." November 2011

TARTU 2011

Acknowledgements

I would like to thank my advisor Prof. Marlon Dumas for his support and promptness throughout the whole project. I would as well like to thank André Karpištšenko and Luciano García-Bañuelos for helping me put the initial idea together.

I thank my family, friends and co-workers for the their love and understanding.

Thank you Ülle for being always there for me.

Abstract

Social networks are nowadays a constant presence in our lives and increasingly have a role in important social and commercial phenomena. Microblogging services such as Twitter appear to play an important role in the process of information dissemination on the Internet making it possible for messages to spread virally in a matter of minutes. In this research work we study the mechanism of re-broadcasting (called “retweeting”) information on Twitter; specifically we use Latent Dirichlet Allocation to analyze users and messages in terms of the topics that compose their text bodies and by means of ANOVA we are able to show that the topical distance between users and messages is shorter for tweets that are retweeted than for those that are not. Using Decision Tree learning we build several models in order to assess the accuracy and usefulness of our topic-based model of retweeting. Our results show that our topic-based model slightly outperforms a baseline prediction measure, so we conclude that such model is indeed a valid option to consider for predicting retweet behavior with possibilities open for improvement.

Contents

Chapter 1
Introduction

Nowadays we observe the increasing adoption of information technologies for replacing traditional communication channels such as mail, telephone, newspapers or news television programs. The Internet has had a particular important role in making many of these communication media inexpensive, customizable and interactive, promoting a revolution in the way people communicate and handle information.

The advent of online social networking has given a new layout to this, allowing people to hold their identity both in terms of what they stand for and their social relationships, creating a platform that permits the average citizen to have an audience. One of such networking models is designated by *microblogging* which, simplistically, consists of publishing short sentences or stories that other people can have access to (typically after subscribing or befriending the originator). Microblogging has become a center of attention in the area of social networking due to the amount of users it has attracted and the load of messages it commands daily [1], as well as the role that it allegedly had in certain major social events, such as the Iranian election protests of 2009-2010 [2,3,4,5].

This research is motivated by a desire to understand what drives users of social networks to disseminate information they come across. One important factor in developing friendships is certainly the interests that people have in common but it is unlikely that one finds someone with whom they agree on everything. Therefore, the existence of a relationship between two individuals alone is definitely not enough to explain why someone will (or not) forward a piece of information they were exposed to. For this purpose we focus on a social network known as Twitter (www.twitter.com) that allows users to exchange short text messages, called *tweets*; occasionally, a certain tweet becomes viral when the users exposed to it keep re-broadcasting it throughout the network (in the Twitter universe, this is called “retweeting”).

We want to test the possibility that the topical distance between a tweet and the users who are exposed to that tweet (“topical distance” roughly being a measure of how close two text bodies are in terms of the words that compose them) has a direct relation with

the action of retweeting. Formally, we want to understand the relation between the likelihood of a follower *F* retweeting a tweet *T,* and the distance between *T* and *F*. More specifically, we would expect that the lower the topical distance between *T* and *F*, the more likely it is that *F* will retweet *T*. Further we evaluate the usefulness of a predictive model that takes topical distance as a feature.

This question is answered utilizing different techniques for text and statistical analysis such as Latent Dirichlet Allocation (LDA) [6], Analysis of Variance (ANOVA) [7] and Decision Trees [44]. We begin by learning which topics define the network, where a topic might be the set (“cat”, “dog”, “food”, “vet”, “vaccine”). After classifying both tweets and users using those topics we measure the topical distance between them. This makes it possible to determine if indeed the distance is related to the acts of retweeting or ignoring a tweet. Finally, several Decision Tree models are tested for evaluating the accuracy of the topical distance in predicting retweets of messages.

The thesis is organized in the following way: in chapter 2 the Twitter social network is described and related work is evaluated; in chapter 3, a description of the Predictive Model is made, followed by the detailed explanation of the an experimental evaluation, in chapter 4. We finalize with the conclusions stating what was achieved and suggesting a direction for future work.

Chapter 2
Background

Recent years have witnessed an emergence of web systems that facilitate the exchange of information in real-time as well as the creation and maintenance of channels through which communication takes place. The appearance of websites such as Blogger, Wikipedia or YouTube fomented an environment where Internet users can easily publish and access user-generated content in the form of written articles, photographs, videos and others. *Social media* was the term coined to refer to this novel approach to information sharing [8]. In addition, those platforms often allowed members to connect with each other by establishing links, typically *subscription* relationships. The evolution of such paradigm resulted in popular *social networks* such as MySpace, Orkut or Facebook where participants keep their *friendship* circles and share news, opinions and links to different types of media. Friends can then join the conversation by commenting, approving or forwarding to their own social networks.

A common characteristic of these social networks is the feature of writing (usually) short sentences, known as *status updates*, that users post on their *homepages* and that are viewable by their friends. The act of publishing status updates is often referred to as *microblogging*. [9]. One of such microblogging systems is Twitter that, despite being regarded as a social network, presents its features in a rather distinct manner and therefore becomes a very peculiar and eccentric ecosystem [10]. Nevertheless, the popularity of these services in general gives researchers the opportunity to study online social networks and the communities that exist within them [12].

2.1. Twitter

Twitter launched in July 2006 as a service that allowed users to blog “on the go” using traditional telephone short message services (SMS). With the call for action being “What are you doing ?”, the original idea was that users would send messages, commonly called *tweets*, to the system and have them posted on their personal

microblog timeline [11]. Their *followers* (Twitter's terminology for *subscriber*) would then get the messages on their own timelines [13] and possibly reply to the tweet, thus engaging in a conversation. [12].

However the use cases of Twitter have gradually changed and, although the SMS functionality is still intact, it is not the core purpose of the service currently. Twitter went from being mostly utilized for sharing thoughts with relatives and co-workers informally [14] to a place where people raise visibility to events, gather information about topics of interest, or obtain help and opinions on specific subjects. [12]. Zhao et al. concluded through a systematic survey that users look for real-time information and utilize it as a people-powered RSS feed[1] [14]. This flexibility in usage may be attributed to the fact that unlike on most online social networking sites a relationship requires no reciprocation i.e. a user can follow any other user, and the user being followed needs not follow back[2], hence employing a *following model* as opposed to the common friendship model. [10, 11].

As Twitter has grown in popularity, it has become a tool for public discussion in many domains of society and *twitterers* (name for Twitter users) are actively exploring its potential to serve other purposes [15]. Research has shown that approximately 19% of tweets do mention an organization or product brand, indicating that microblogging is viable for implementing viral marketing campaigns or doing customer relationship management, and is a place where electronic word-of-mouth takes place [16]. Another study acknowledges that over 85% of tweets revolve around headline news or news that are persistent in nature [10]. This deviation from its original use was in fact acknowledged by the company when in late 2009 it changed the motto of the platform into “What's happening ?” [17]

Twitter.com currently ranks 9th on the Alexa Traffic Rank[3] and its relevance in the Internet scene is testified by search engines like Google and Bing that are now including feeds from Twitter in their search results. [9]. The popularity of Twitter amongst Internet users is also confirmed by the fact that as of September 2011, over 100 million active users posted about 200 million tweets each day, equaling over 2300 tweets sent each second [18, 19, 20].

---

[1] http://en.wikipedia.org/wiki/RSShttp://en.wikipedia.org/wiki/RSS

[2] in fact Twitter allows users to decided who can follow them but that is not the common usage.

[3] http://www.alexa.com/siteinfo/twitter.com

2.1.1. What is a *tweet* ?

As noted above, status updates or messages in Twitter are commonly designated as *tweets*. [9,22]. The fact that the original concept was based on the mobile phone short message service determined the maximum length of 140 characters per message, twenty less than the 160 allowed in standard SMS; this reduction was needed in order to reserve space for the username of the sender [13,17] therefore constraining tweets to a few words per transaction [11,12,21]. This restriction, often aggravated by the inclusion of web links in the payload [9], compelled users to communicate their messages in dialects analogous to the ones employed in SMS and instant messaging [22]. While the use of such abbreviations allows tweets to contain more word-tokens, it makes them harder to mine for information, due to its lack of standardization. Table 1 shows some examples [22]. Below we present the main area of a Twitter.com homepage.

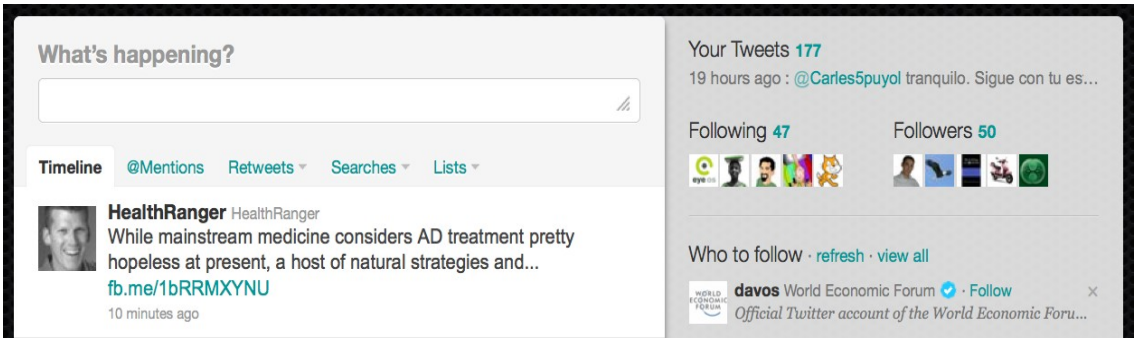

Figure 1 – Main section of Twitter.com homepage.

2.1.2. The tweet protocol

Given that a tweet is merely a text message (which means the system does not provide a means for distinguishing particular purposes) users created conventions for referencing others as well as differentiating message types; this resulted in a specific markup language [10,11]. While a tweet can be simply a general statement broadcasted by a user to all her followers, it may also refer particular users by applying the '@' symbol together with the recipient(s) username(s) (ex.: *@khoesant* for referencing the author of this research work). Also, conventions exist for specifying the type of the message, as illustrated in Table 1.

An *@username* reference in a tweet is generically called a *mention*. A mention is used for signaling that the entity being mentioned is as well registered on Twitter. A particular case of a mention, the *reply*, occurs when one user addresses another directly or responds to another's previous tweet by placing *@username* in the beginning of the message [11,23].

(generic) tweet	mention(s)	reply	retweet
"Just got out of the shower"	"By the way, that catchy tune you hear in the video is by **@FreelanceWhales**. We love them! #NewTwitter"	"**@jasongroupp** I think you meant shoot=take pix... Right? ;)"	"RT: **@google** LIFE Magazine (1936-1972) now on Google Books: http://bit.ly/3AJxxg "
"Coding in C++ and python is like the difference between watching a James Bond movie and being James Bond."	"Congrats to **@alpjor** & **@Trammell** on your new gigs! Excited to hear about the great things you'll both be doing for **@facebook** & **@twitter**!"	"**@Adamcrombie** lol what are you referring to? The ad? Lol"	"http://twitpic.com/iuyzk - I had a great meeting with Joe Perry and his wife Billie - limit#Aerosmith" (via **@SenJohnMcCain**)

Table 1 – The four Twitter message types. Twitterers have to use certain conventions in order to make it clear what type of message it is. Note that a reply or a retweet are particular cases of a mention and all are tweets.

In Twitter, a *retweet* consists of re-broadcasting someone else's tweet and normally involves starting the message with 'RT' followed by '@username'; the sender may include additional comments also [10,11]. This mechanism empowers users to spread out information that they consider interesting beyond the reach of the originator of the message. [10].

Boyd et al. point out the inconsistency of retweet formats and show several more syntaxes: 'RT: @', 'retweeting @', 'retweet @', '(via @)', 'RT (via @)', 'thx @', 'HT @', 'r @', and '♻ @'; different syntaxes may depend on personal preference or the third-party client used to access Twitter. Table 2 shows several prototypes of a retweet. As of this writing, Twitter has incorporated the retweeting functionality into the system allowing it to be done in one click. However such retweets are still harvestable using the aforementioned syntaxes. [24]. Another non-standardized item of a retweet, is which

sources to reference in the payload: the originator, the last user in the chain, or all involved.

tweet	*A*: Hello World!	original message by **A**
retweet alternative 1	*B*: Hello my People **RT @A**: Hello World!	retweet made by **B** who is a follower of **A**
retweet alternative 2	*B*: Hello World! **(via @A)**	retweet made by **B** who is a follower of **A** utilizing a different method for referencing
retweet alternative 3	*C*: RT **@B**: RT**@A**: Hello World!	retweet made by **C** who is a follower of **B** but not of **A**
retweet alternative 4	*C*: **RT @A**: Hello World! **(via @B)**	retweet made by **C** who is a follower of **B** but not of **A**, utilizing a different method for referencing

Table 2 – Different styles of retweeting, all widely acceptable.

The *retweet* is considered the feature that has made Twitter a new medium of information dissemination [10] and is, in fact, the most relevant type of tweet in the context of this research. Generally speaking, the interest of studying retweeting behavior is that retweeting is a form of information spread. Thus, the study of retweeting behaviour may shed insights into the general mechanisms for information spread in social networks.

Although not relevant in defining the type of a message, it is common practice to summarize the topic of the message (called a *hashtag*) by writing '#' followed by the classifier e.g. "#sports" or "#ilovefridays". It has been found that 5% of tweets in general and 18% of retweets contain hashtags [11]. The use of such hashtags is exemplified in Table 1 (bottom-rightmost cell).

2.2. Related Work

The increasing popularity of online social networking services has attracted research by the scientific community. Here we overview several works recently published focusing on Twitter.

Kwak et al. [10] quantified different aspects of the network from a massive dataset containing nearly forty-two million user profiles, about 1.500 million relations and 106 million tweets. Their analysis found that the top users by number of followers in general correspond to celebrities and mass media who in principle, do not follow back. Also, the

general level of reciprocity was of 22.1% i.e. user pairs connected is 77.9% of the cases one-sided. In addition to that, 67.6% of users are not followed by any of their *followees*[1] (term used to refer to someone who's followed by others) which leads to conjectures that Twitter is used predominantly as a source of information and less as a social network. The same work acknowledges the phenomenon of favoritism: people only retweet from a small number of followees and only a subset of a user's followers actually retweet; this factor indeed foments our interest in understanding if a tweet has higher probability of being retweeted if its content is closer to followers' interests.

Still regarding "retweeting", the same study claims that independently of the number of followers the remittent of the message has, the message is to reach an average of 1000 users. In addition, half of the retweets take place within an hour and 75% under 24 hours; this information is relevant given that we need to arbitrate a time interval within which we accept a tweet as being a legitimate retweet. This is important in the context of our experiment, namely for arbitrating a time interval for considering retweets.

Weng et al. [25] centered their work on identifying topic-sensitive *influentials* in a dataset of over 6500 Singapore based users, defining an influential as a user with a certain authority within her social network. Amongst other applicabilities, the influence exerted by a certain member within a community is a relevant area of research in many social sciences, including marketing which is indeed one of Twitter's uses (see section 2.1.). As an alternative to the common metric of node in-degree, they propose a novel method for measuring influence that borrows from the concept of PageRank and adds to it the topics that define users in terms of their tweeting content. This work differs from our research in the sense that it attempts to understand the strength of social bonds and authority rather than directly relating a tweet and a user.

In the same research, topics are extracted with a Latent Dirichlet Allocation (LDA) implementation, defending that hashtags occur only sporadicly and are therefore insufficient; given that the aim of such research is to classify the overall behavior of users, tweets are thus concatenated into a single document per user.

[1] this term does not exist in the Oxford English Dictionary but it is widely used by academic to refer to the person whom a user follows.

In terms of reciprocity the same work obtains an antagonistic result to [10] in that 72.4% of users follow more than 80% of their followers and 80.5% of users have 80% of followees following back. These opposing conclusions may be representative of the previously mentioned shift of use cases (refer to section 2.1.) even more so when [25] regards a single country only.

A study by Canini et al. introduces a technique to enhance the ability of finding relevant users given a topic. It combines topology properties with a content-algorithm that, as well, distills topics by using LDA.

Huberman et al. [39] reports that the number of *friends* (i.e. bidirectional relationships) is actually smaller than the number of followers or followees; furthermore, it finds that the number of friends is an indicator of tweeting activity which means that the network of actual interactions (called *activity network*) is more representative than the declared network of friends.

More recently, Gonçalves et al. studied Twitter's conversations in a massive dataset of 380 million tweets and found evidence of Dunbar's [27] theory that correlates brain neocortex size of primate species with group/community size [26]. For humans, it averaged 150. The paper claims that Twitter users will at most maintain interactions with 100 to 200 other members, regardless of how big their declared social network is i.e. one may follow 2000 other accounts but address only a fraction of those when tweeting. The suggestion that the declared friendships are not a very accurate indicator of real interactions is an interesting one. Instead, the level of authority of the source of a tweet and/or the proximity of a tweet's topic relative to the topics of interest of the follower are likely to drive the interactions between tweeters, and particularly their retweeting behaviour.

As far as the author of this thesis is concerned, there has been no research studying specifically the direct response of users to tweets by taking into account the content similarity of the two.

Chapter 3
Topic-based Model of Retweeting

As discussed in Chapter 1, the goal of this work which is to study the phenomenon of retweeting. More precisely, the goal of this research is to create a model that is able to, given user U (and by extension the set of her followers F_u) and a tweet T sent by U, predict the response in terms of retweeting of each member of F_u with a certain degree of accuracy.

Generally, our approach consists of measuring how distant a tweet and a user (exposed to that same tweet) are in terms of the content they carry. Further, we study the correlation of such distance with the behaviour of the user regarding retweeting.

In the following sections, we introduce and justify each the techniques chosen to build the predictive model of retweeting.

3.1. Topic Identification

Several research works (refer to section 2.2.) apply Latent Dirichlet Allocation (LDA) successfully for characterizing tweets and twitterers in terms of the topics they are "composed" of. In [42] Kireyev et al. argue that traditional natural language processing (NLP) techniques that rely on syntactic models are ill-prepared for dealing with microblogged messages because of their non-standardized nature (see sections 2.1. and 2.2.). The same work defends that LDA is a more adequate option due to the "bag-of-words" model which does not rely on syntactic structure or word order in the text, thus likely better for handling Twitter's irregular grammar. In addition, topic models infer latent relationships between elements in the data, making them more robust for handling misspellings, acronyms, terminology and other variations of messages; another advantage is that it can represent statistical knowledge as homogenous numerical vectors that ease comparisons, visualization as well as mathematical manipulations. Following, we briefly describe the technique.

LDA (Blei, Ng, & Jordan 2003) is an unsupervised machine learning technique utilized to identify latent topics in a set of documents. Formally, in LDA a collection of D documents is associated with a multinomial distribution over T topics, symbolized by $\boldsymbol{\theta}$. Each topic is, in turn, associated with a multinomial distribution over words, denoted as $\boldsymbol{\varphi}$. $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ have Dirichlet prior with hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. For each word in a document d, a topic z is sampled from the multinomial distribution $\boldsymbol{\theta}$ associated with the document, and a word w from the multinomial distribution $\boldsymbol{\varphi}$ associated with topic z is sampled consequently.

Document-topic distributions $\boldsymbol{\theta}$, and T topic-word distributions $\boldsymbol{\varphi}$ are inferred from the data. By learning the two parameters, it is obtained the information of *a*) which topics each document is about and of *b*) the representation of each document in terms of these topics. In practice this is called a "bag of words" model that basically processes documents as vectors of word counts i.e. a document is represented by a probability distribution over topics and a topic is represented by a probability distribution over words. The LDA algorithm performs a generative process for assign the topics of each document, in the following way:

1. for each document, pick a topic from its distribution over topics,
2. sample a word from the distribution over the words associated with the chosen topic,
3. repeat the process for all the words in the document.

Figure 2 shows the LDA model. Arrows represent a conditional dependency between two variables and boxes represent repetitive sampling with the number of repetitions given by the variable at the bottom of the box. Shaded and unshaded shapes represent observed and latent variables, respectively.

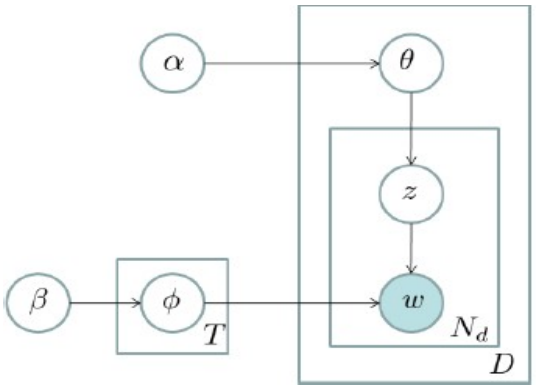

Figure 2 – LDA model

The result of applying LDA produces *DT* which is a *DxT* matrix, where *D* is the set of documents classified, and *T* is the set of topics. Each row in *DT* represents the probability distribution of a document over each topic. Therefore, DT_{ij} contains the proportion that a certain topic has in defining the document.

In the specific context of modelling retweeting behaviour, we're interested in having *D* representing users and tweets in order to classify them according to their topical preferences. For example, the tweet "Good morning World", classified by a 10-topic model could potentially be characterised by the vector of topics shown in Table 3.

topic0	topic1	topic2	topic3	**topic4**	topic5	topic6	**topic7**	**topic8**	topic9
0,10	0,05	0,00	0,00	**0,30**	0,00	0,05	**0,20**	**0,20**	0,10

Table 3 – Example of a document classified in terms of a vector of topics

Note that the topics are simply concepts and not mutually exclusive sets of words. Table 4 describes how topics can be composed, assuming descending order of relevance of the words within a topic. The word "good" occurs in topics four and seven, but with different importance in each. Basically the distribution of topic probability indicates the tendency of a document.

LDA generates the sample topics shown below by learning from a training set of concatenated tweets (this process is approached in detail in section 4.3.3). The sets of words that composed a topic are constructed based on the patterns of words that consistently co-occur in the same documents.

topic4	topic7	topic8
morning	**good**	news
evening	stuff	round
good	lot	planet
afternoon	lord	**world**
night	eat	society
sun	people	organization
cold	church	club

Table 4 – Example of topics used to classify the document of Table 3.

The LDA inference process exemplified here is to be applied to both tweets and users.

3.2. Topical Distance between a Tweet and a User

As previously described, LDA classification yields vectors of topics as a means for describing documents. We need to calculate the distance between those vectors in order to measure the distance between users and tweets. [29] states that the distance between LDA documents can be calculated using the Squared-Chord distance. In the specific context of Twitter, the distance between users (defined as topics vectors as well) has been calculated with methods such as the inner product [21] or the Jensen-Shannon Divergence [25] with satisfactory results.

$$d_{JS} = \frac{1}{2}\left[\sum_{i=1}^{d} P_i \ln\left(\frac{2P_i}{P_i + Q_i}\right) + \sum_{i=1}^{d} Q_i \ln\left(\frac{2Q_i}{P_i + Q_i}\right)\right]$$

Formula 1 – Jensen-Shannon formula

In Formula 1, P and Q represent the two vectors from which the distance is calculated where i denotes the position in the vector. Besides these two measures, we can use other measures of distance between vectors such as cosine distance and others, as defined for example in [30], including those belonging to the Minkowski, Squared and Shannon's entropy families [31].

Chapter 4
Experimental Evaluation

In this chapter we start by describing the data utilized followed by the reasoning behind the options taken in the design of the experiment. We then detail each step of the experiment and conclude with the results.

4.1. Data Sources

We utilize a fraction of a dataset published by the *Stanford Network Analysis Project (SNAP)* from Stanford University [36]. The original dataset covers a 7-month period from June 1 2009 to December 31 2009 and according to SNAP it contains approximately 20-30% of all public tweets published on Twitter during that particular time frame. The data is provided in separate files, aggregated per month. For this experiment, we picked the month of September given that it shows the highest rate of tweets per second of all; meaning the harvesting process was able to collect more tweets for this time interval than for the overall period in a total of 94.176.126 tweets, 35.213.716 of which including URLs. Each tweet is described by: author, timestamp and content of message. As of this writing, Twitter has requested SNAP to not have the data available for download from the website.

```
T       2009-09-01 01:13:24
U       http://twitter.com/cipherbreaker
W       If you want to know what is
really going on in America....watch
this video! http://bit.ly/QhMKf
```

Table 5 – Example of a tweet in the dataset. Source [36]

As presented in Table 5, *T* indicates the timestamp of when the tweet occurred, *U* represents the twitter address of the author and *W* is the body of the message in the tweet.

In addition, we downloaded a twitter graph dataset from [37] which is linked to at SNAP's website. This dataset contains the graph of declared following relationships

between users, meaning that it contains an edge from user u_1 to u_2 if user u_1 follows user u_2. It represents following relationships between users in this way:

1	34
1	68
2	3
2	15
2	39
2	72
3	34

Table 6 – Extract of the graph dataset (user on the right follows user on the left). Source [37]

4.2. Experiment design and considerations

In this section, a list of considerations is made that delimit and give context to the way in which the experiment is laid out.

4.2.1. Activity Network versus Declared Network

As found out by [39], Twitter users tend to narrow down their activity to a fraction of their declared relationships i.e. they follow or are followed by users with whom they don't interact. This the activity network and in this section we want to verify that phenomenon. This is important in order to understand the usefulness of the graph, especially given the fact that the datasets were collected by different research teams in potentially disparate points in time. Since the declared followers graph is only representative of a specific point in time, it's possible that the relationships implicit in the tweets dataset do not correspond.

To execute this step, we randomly picked a set of 1000 users and their respective followers from the tweets dataset, where a user is a follower of another if it has mentioned it in at least one of her tweets. We gathered the declared followers of the same 1000 users, based on the graph, and performed a match (per user) of the two sets. The degree of overlap was negligible, therefore rendering the graph unusable for this experiment. We believe that this fact goes in accordance with the findings reported by Huberman et al. [39] (see section 2.2.) which claim that the actual activity network in Twitter is far smaller than the declared followers graph network. We therefore opt by arbitrating a **follower as a user that has mentioned another at least once** meaning that

we build the graph based on the activity network rather than on the declared graph.

4.2.2. Definition of User

We are interested in two distinct entities: tweets and users. Tweets are structurally defined in the dataset (see 3.1.) and therefore require no further treatment. However, the definition of user must be arbitrated. Similarly to what [25] defend, we create the text body of a user by concatenating of all its tweets.

Also, for this experiment a user whose tweets' will be tracked is called an *author*. Given the various purposes for which people use Twitter, it is expected that a portion of the users behave very differently from the average user. Activities that take place in the social network such as marketing, customer relationship management or celebrity presence are not the type of phenomenon intended for analysis in this study; in addition, the existence of spam must be considered. Similarly, users that rarely write tweets provide little value for this experiment and so it is adequate to ignore those. This factor is an important one in the implementation of the experiment and will be revisited in section 4.3.2. of the Experiment Implementation.

4.2.3. Focus on URLs

Suh et al. [24] found that URLs occurrences in non-retweeted tweets is 19% whereas in retweeted tweets it is 57%. In addition to that, tweets may contain very little context in their content. As the sample tweet in table 7 shows, the author tweets an enigmatic statement from which one can only retain two words with real content: “America” and “video”. This is common practice in Twitter so, in order to better understand what the author is actually writing about, one would have to open the link. Hence we decide to build the topics model based on tweets that contain URLs and add their titles to the content of the tweets.

```
T       2009-09-01 01:13:24
U       http://twitter.com/cipherbreaker
W       If you want to know what is
really going on in America....watch
this video! http://bit.ly/QhMKf
```

Table 7 – Example of a tweet with very little context in its payload. Source [36]

4.2.4. Removing stopwords

[38] apply different natural language processing techniques to analyze Twitter's trending topics. One of the approaches utilized concerning the definition of *stopwords* is to count occurrences of each word in the documents and remove any word that appears in at least 75% of the cases. We applied the same restriction but it was not itself sufficient, resulting in LDA topics which top words were meaningless (such as “lol” or “wow”); for this reason we used an iterative process where we added the top words (if considered useless by visual inspection) of each topic to the list of stopwords and repeated the LDA estimation until the words in the topics appeared more meaningful.

It is arguable how adequate it is to remove stopwords from tweets given that loss of detail is guaranteed, particularly in a context where abbreviations are frequent; however, we believe this tradeoff will in most cases be an advantage. A clear example of loss of detail:

Word “wow”	
Abreviation for *World of Warcraft* online game	**Interjection** used for expressing surprise.

Table 8 – example a stopword candidate which is relevant in other context.

We understand this tradeoff but it is our conviction that it's a necessary part in the experiment.

4.3. Experiment Implementation

As defined in chapter 3, the experiment consists of four main separate phases that require different tools. The following is a list of all the software packages utilized:

- **Topic Identificaton** – an implementation of the Latent Dirichlet Allocation (LDA) that uses Gibbs sampling[4] for parameter estimation and inference (GibbsLDA++)
- **Topical Distance** – Python's[5] library of scientific tools SciPy[6], well-known open-source software for mathematics, science, and engineering.
- **Correlation of Topical Distance and Retweeting** – again we use SciPy for this task.
- **Construction of the predictive model** – Weka[7], a collection of machine learning algorithms implemented in Java by the University of Waikato, New Zealand.

In addition, the transformations of the datasets from the output of one phase to the input of the next are done with custom Python scripts.

These four core phases are however organized and executed with a higher degree of granularity, as the following sections detail.

4.3.1. Defining and harvesting the *Trackable Tweets*

The selection of the tweets to be tracked, we call them *trackable tweets*, requires the presence of URLs in the message and also the word "iphone"; the justifications are shown on table 9:

Requirement	**Justification**
URL present	URLs are an important factor in the phenomenon of retweeting and occur in more than half of the retweets. (see section 4.2.3.)
"iphone" word present	Given the unstructured nature of tweets, one needs to arbitrate a definition for what a retweet is. Identifying a common word in the original tweet and its corresponding retweets considerably raises the correctness of the match.

Table 9 – Requirements for a tweet to be eligible for the experiment.

The requirement for a specific word to occur in both a tweet and its corresponding retweets is rather critical. In section (2.2.2.) it is stated that tweets lack a rigid structure which therefore obliges the observer to arbitrate what a tweet in fact is; the following fictional sequence of tweets illustrates this fragility:

[4] http://gibbslda.sourceforge.net

[5] http://www.python.org

[6] http://www.scipy.org

[7] http://www.cs.waikato.ac.nz/ml/weka

	username	timestamp	message
1	@angel	19:02	Just had access to inside info. Next iPhone model: http://myiphone.me
2	@brown	19:03	RT **@angel** Just had access to inside info. Next **iPhone** model: http://iphone-secrets.info
3	@carlo	19:03	Cool!! RT **@angel** Just had access to inside info. Next **iPhone** model: http://iphone-secrets.info
4	@angel	21:43	Watching Barcelona vs Real Madrid... Barça is the new dream team!
5	@dennis	21:59	RT **@angel** Just had access to inside info. Next **iPhone** model: http://iphone-secrets.info

Table 10 – Chronology in Twitter

The chronology presented shows a perfectly plausible Twitter scenario where, given the rate at which the author produces tweets, the reactions of the followers occur in an "unordered" way. The last retweet (row 5) made to @angel's first tweet (row 1) happens after a second tweet (row 4) by @angel. The requirement for a word to be present increases the probability of true positives in the process of matching tweets with retweets. The choice for this rather popular word was made in order to not excessively reduce the amount of data to work with, for the remainder of the experiment.

A full scan of the tweets dataset was performed in order to gather all tweets that fulfill the two requirements and we call them *trackable tweets*. We arbitrate nine hours as the time interval allowed for a tweet to be considered a retweet (refer to section 2.2.)

4.3.2. Exposure Graphs

After isolating the trackable tweets, a check is made on the activity levels of the authors in order to discard those that do not comply with the requirements defined in section 4.2.2. The authors are validated so that only authors that tweet a minimum of thirty times and a maximum of 210 times during a month (equivalent to a minimum of one tweet and maximum of seven tweets per day) are taken into account.

At this point, it is necessary to find all users that are exposed to each of the trackable tweets; the network graph (recall section 4.2.1.) is queried by authors that passed the aforementioned check, and their followers retrieved. Note that the activity filtering is made exclusively to authors of tweets. Finally we obtain over 7.000 pairs (author, tweet) and the roughly 170.000 followers of those authors.

Each tweet is then tracked along the tweets dataset and all acts of retweet are registered, resulting in what can be expressed by Table 11 (RT stands for ReTweet whereas NORT stands for NO ReTweet).

(Author, Tweet)	Follower	Reaction to tweet
(A1, T1)	**F_1**	NORT
(A1, T1)	**F_2**	NORT
(A1, T1)	**F_3**	RT
(A2, T2)	**F_1**	NORT
(A2, T2)	**F_2**	NORT
(A2, T2)	**F_3**	NORT

Table 11 – Representation of an exposure graph. A tweet *t* is written by an author *a* and retweet or ignored by each follower *f* of *t*.

At this point in the process, all entities of interest are well defined: tweets, users that retweet and users that do not retweet. Note that most of the tweets are not retweeted which results in about 2.500 tweets that are retweeted.

4.3.3. LDA Model Estimation and Inference

In this part of the experiment, we build a topic model (estimation) for classifying tweets and users (inference).

We collect all the tweets (the text string signaled by "W") of 100.000 randomly chosen users, thus obtaining what we define as *user histories* (as described in section 4.2.2.). The user histories are further enriched with the titles of the URLs they link to, for enhancing data quality (section 4.2.3.) and the resulting data are lowercased plus cleaned from stopwords (see 4.2.5.)

```
T       2009-09-01 01:13:24
U       http://twitter.com/cipherbreaker
W       If you want to know what is really going
on     in     America....watch     this     video!
http://bit.ly/QhMKf
```

Table 12 – Example of a tweet with very little context in its payload. Source [36]

Photo: Lars Top-Galia does it again..! Very cool! http://tumblr.com/xnq2ybrng Photo: Finally! New shit on the office walls! http://tumblr.com/xnq2yhmer RT @signaldigital Imorgen er der fernisering på vores Silverthorne-udstilling, og vi glæder os som sindsyge! http://bit.ly/sd20silverthorne Photo: Spotify Premium now available at Pressbyrån and 7-Eleven (in Sweden) http://tumblr.com/xnq2ylk34 Photoset: Had a feeling it wouldn't be – Still pretty disappointed though.. http://tumblr.com/xnq306yqb Photo: This is funny as hell.. http://tumblr.com/xnq307x3q RT @cphtwestival Copenhagen Twestival 2009 » New location and new sponsors http://bit.ly/KbYZB RT @cphtwestival: We have just revealed parts of the program for @cphtwestival: http://bit.ly/6nAWZ. Grab your ticket: http://bit.ly/1CwkdO

Table 13 – example of user historic, aggregation of all tweeting history.

The dataset of histories is then provided as input to the LDA implementation for buildinging the model. According to the documentation of GibbsLDA++, 100 is the default value for topics to be yielded. In [21], the experiment on Twitter is run for fifty topics. We thus use these two values as a reference point and extend it to two more, resulting in four different models considered: for fifty, 100, 200 and 300 topics. Having the topic models ready, we used them to classify the tweets and the users exposed to them.

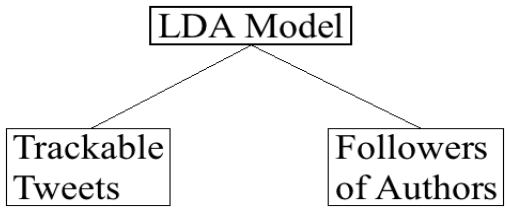

Figure 3 – LDA classification of tweets and followers datasets.

Recall from section that that the output of the inference process is a set of probability distribution vectors where each vector represents each text document.

4.3.4. Topical Distance

At this stage we have all followers and tweets classified in terms of vectors of topics, allowing the distances between a tweet_vector and a follower_vector to be calculated easily.

The table below provides a visual idea of how the separation between retweeters and non-retweeters is approached. Suppose that a certain author (A) has only three followers

(F_1, F_2 and F_3) that are by default exposed to A's tweet's.

Author *A* publishes one tweet and, of the three followers, two ignore the tweet and the remainder one finds it interesting and retweets it – this is indicated by "NORT" (did NOt ReTweet) or "RT" (ReTweeted) in Table 15.

Tweet by Author A

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
0,10	0,05	0,05	0,00	0,30	0,00	0,05	0,15	0,20	0,10

Follower F_1 (NORT)

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
0,03	0,07	0,20	0,12	0,08	0,05	0,15	0,00	0,30	0,00

Follower F_2 (NORT)

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
0,00	0,40	0,00	0,15	0,05	0,12	0,08	0,03	0,07	0,10

Follower F_3 (RT)

topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
0,00	0,40	0,00	0,15	0,05	0,12	0,08	0,03	0,07	0,10

Table 14 – Representation in terms of LDA vectors of the three followers and the tweet they're exposed to.

The measures referred in section 3.2. are used to calculate the distance between topic vectors of followers and tweets; in our hypothetical example this scalar value is represented in the middle column of Table 15.

Follower	Topical distance to tweet	Reaction to tweet
F_1	13	NORT
F_2	15	NORT
F_3	7	RT

Table 15 – Summarization of Table 14 where the topical distance is calculated and represented as a scalar value.

4.3.5. Correlation of Topical distance and the act of Retweet with Analysis of Variance

The step of assessing the existence of a correlation between topical distance and retweeting behaviour (see 3.2.) is critical in the making of the predictive model since it determines if the model can or not rely on the topical distance.

To test this correlation, we use a statistical hypothesis test called Analysis of Variance (ANOVA). ANOVA tests the *null hypothesis* that samples in two or more groups (possibly with differing sizes) are derived from the same population by estimating the variance of their means [32,33,34]. This test fits our goal of testing if the distinct sets of retweeters and non-retweeters do have the same topical distance to a tweet that both sets are exposed to.

The ANOVA method produces two output values: the F-ratio and the *p-value*. If the difference between the means is due to chance, the expected value of the F-ratio is one (1.00) [34]. The p-value represents the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If ANOVA yields a p-value lower than the *significance level* α, the null hypothesis is rejected which means the results is considered "statistically significant". Despite the *significance level* being arbitrary, it is conventionally used at 5% (0.05) or 1% (0.01) [35].

We therefore define $\mu_{retweet}$ as the mean topical distance of pairs <tweet, user that retweeted> and $\mu_{noretweet}$ as the mean topical distance of pairs <tweet, user that did not retweet>. The null hypothesis is then H_0: $\mu_{retweet} = \mu_{noretweet}$ and the alternative hypothesis is H_1: $\mu_{retweet} < \mu_{noretweet}$.

It should be noted that ANOVA requires that the samples are independent, have approximately equal variance and that they are from a distributed population. We assume here that the act of retweeting is an individual decision and therefore the samples are independent from each other [40]. We further verify that the variances of both samples are approximately equal. Regarding normality, we follow the criteria set by [41] which states that if kurtosis is within [-0.5, 4] the population can be considered normal.

Following, the results of ANOVA for fifty, 100, 200 and 300 topics. Shaded rows mean that the p-value is lower than the target significance level $\alpha = 0.01$.

Distance measure	F-ratio	p-value
Braycurtis	2.182	0.140
Canberra	2.182	0.140
Chebyshev	1.212	0.271
Cosine	6.460	0.011
Dot-product	29.950	$4.446e^{-08}$
Euclidean	1.604	0.205
Hamming	0.479	0.489
Jensen-Shannon	0.548	0.459
Manhattan	2.182	0.134
Pearson	63.915	29.141
Squared Euclidean	1.723	0.189

Table 16 – ANOVA for 50 topics.

Distance measure	F-ratio	p-value
Braycurtis	1.774	0.183
Canberra	1.774	0.183
Chebyshev	1.903	0.168
Cosine	22.537	$2.064e^{-06}$
Dot-product	43.892	$3.491e^{-11}$
Euclidean	1.672	0.196
Hamming	0.941	0.332
Jensen-Shannon	0.3704	0.5427
Manhattan	1.7739	0.1829
Pearson	77.3333	62.1455
Squared Euclidean	1.0890	0.29668

Table 17 – ANOVA for 100 topics.

Distance measure	F-ratio	p-value
Braycurtis	1.622	0.203
Canberra	1.622	0.203
Chebyshev	5.668	0.017
Cosine	31.675	$1.831e^{-08}$
Dot-product	20.616	$5.621e^{-06}$
Euclidean	6.090	0.013
Hamming	0.055	0.814

Jensen-Shannon	5.339	0.020
Manhattan	1.62	0.203
Pearson	32.36	46.69
Squared Euclidean	7.90	0.005

Table 18 – ANOVA for 200 topics.

Distance measure	F-ratio	p-value
Braycurtis	3.23	0.07
Canberra	3.23	0.07
Chebyshev	8.03	0.004
Cosine	53.50	$2.62e^{-13}$
Dot-product	32.96	$9.47e^{-09}$
Euclidean	9.34	0.002
Hamming	0.04	0.84
Jensen-Shannon	11.09	0.0008
Manhattan	3.23	0.07
Pearson	53.55	138.16
Squared Euclidean	10.83	0.001

Table 19 – ANOVA for 300 topics.

Note that as the number of topics increases, more methods yield p-values below α. This suggests that an LDA-based topical distance can be a good characteristic for modeling retweet behavior. Also, the results show that cosine distance consistently outperforms other methods. Therefore we reject the null hypothesis and accept the alternative hypothesis. We try to further validate this fact.

4.3.6. Topical Distance as a Predictor of Retweeting

Given that ANOVA indicated a relation between topical distance and the act of retweeting, we will use a Decision Tree to compare the topical distance against more basic features such as the *overall retweet ratio* (ORR) and the *retweet ratio with regard to the author of the tweet* (RRA). For this, we use the cosine distance measure as a value for topical distance, given that it is the one that evidences the best results.

We use the C4.5 algorithm to generate the decision tree.[43]. Decision trees are a form of multiple variable analyses. All forms of multiple variable analyses allow to predict, explain, describe, or classify an outcome. An example of a multiple variable analysis is a probability of sale or the likelihood to respond to a marketing campaign as a result of the combined effects of multiple input variables, factors, or dimensions [44].

In line with traditional methods for statistical classifier validation, we measure the accuracy of the classifiers obtained based on the confusion matrix, sketched in Table 20.

		Predicted	
		positive	negative
Actual	positive	TP	FN
	negative	FP	TN

Table 20 – Confusion Matrix

- TP – true positive: instances that are "positive" that are classified as "positive",
- TN – true negative: instances that are "negative" that are classified as "negative",
- FP – false positive: instances that are "negative" that are classified as "positive",
- FN – false negative: instances that are "positive" that are classified as "negative".

From the confusion matrix it is possible to calculate the precision and recall of the classification:

- precision = TP / (TP + FP)
- recall = TP / (TP + FN)

It would also be possible to calculate the F-measure, which combines the precision and recall into a single number, but in the context of this study we deem this unnecessary given that the number of experiments is small.

Hence we build different combinations of inputs based on the three variables available: *overall retweet ratio* (ORR), *retweet ratio with regard to the author of the tweet* (RRA) and *topical distance* (TD).

In general, learning from imbalanced data is often reported as being a difficult task. Decision Trees specifically may need to create many tests to distinguish the minority class cases from majority class cases when in presence of class overlapping [45].

To overcome this limitation, the number of elements for each class (retweet/noretweet) was evened out i.e. the proportion of retweets and non-retweets instances was made one (1). Roughly 5000 acts of retweet/noretweet were used and random sampling was applied to separate 75% of the data for training and the remainder for testing. We calculate precision and recall, as described in 2.4.

Features combinations	Precision	Recall
ORR & RRA	0.731	0.727
ORR & RRA & TD	0.719	0.715
ORR	0.546	0.532
ORR & TD	0.587	0.587
RRA	0.735	0.732
RRA & TD	0.719	0.716
TD	0.577	0.571

Table 21 – Results of the Decision Trees Models using different combinations of features (ORR, RRA and TD).

Table 21 shows the results of running the decision tree algorithm on all possible combinations of the three features considered. The strongest result is the *retweet ratio with regard to the author of the tweet* (RRA) feature that is able to classify tweets with a precision of 73.5% and a recall of 73.2%. Furthermore, the results show that indeed the *topical distance* feature (TD) does not improve the results of RRA and, in addition to that, TD individually is also weaker than RRA individually.

From another angle, it can be said that TD is stronger than ORR given that it individually has higher precision (+3.1%) and recall (+3.9%); TD is also able to improve ORR alone by 4.1% precision and 5.5% recall. The improvement that the topical distance is able to create when utilized together with the overall retweet rate is not very significant but it opens room for speculation as to whether topical distance might or not be able to be a good starting point in understanding retweeting behavior in terms of content of messages.

The fact that TD is individually much less accurate than RRA (roughly 16% less for both precision and recall) but when used together lowers both precision and recall only in 1.6% may imply a correlation between the two features.

Conclusions

The results of this research allow us to conclude that indeed topical distance based on LDA topics and the retweeting phenomenon are related. For each LDA model tested (with fifty, 100, 200 and 300 topics) ANOVA was able to consistently yield p-values below the significance level $\alpha = 0.01$, obtaining its best results for the 300-topic model. The results show that the topical distance between a tweet and a user is lower for retweeted tweets than for non-retweeted ones.

This correlation is further confirmed by the results of the classification with decision trees which show that topical distance can be used as a feature to predict whether or not a tweet will be retweeted.

The topical distance feature, taken individually, has better precision and recall than the *overall retweet ratio* of a user for predicting retweets.. The topical distance is however considerably less accurate than the *retweet ratio with regard to the author of the tweet* when each are tested separately but not when the model is built using both features. This fact may imply that the two are correlated in some way. In other words, the *retweet ratio with regard to the author of the tweet* might encapsulate the topical distance feature in itself because when a follower f retweets user's u tweet t, that action reflects on the history of follower's f tweets and also on f's retweet ratio.

The most important contribution of this research is that indeed we find evidence that people consider the affinity that they have with a certain piece of information before they decide to re-broadcast it to their personal social networks. Namely, the textual content of the message directly influences the act of retweet if it relates to the history of messages of the recipient.

Therefore, it is our conviction that there is reason for further studying information dissemination in social networks based on topical content. It must be noted that the options taken in the design of the experiment were made having strong scientific support but do not necessarily correspond to an ultimate answer to the problem. Different techniques may be tested to potentially improve the insights brought by the current work. It is also important to refer that a higher degree of variability shall be applied as well in future research, for instance, the number of topics tested or the arbitration of what a retweet is, are two examples of where further investigation can explore.

We are as well aware of limitations in this research. For instance, restricting the experiment to tweets about one specific theme (“iphone”) is clearly a limitation of this work that affects the generalizability of the results. The arbitration of the network graph is also an important fragility given that it acknowledges only users that interact with each other, potentially leaving out great part of the links in the graph.

Resümee

Magistritöö (30AP)

José Santos

Igapäevase eluga põimunud virtuaalsed sotsiaalvõrgustikud omavad üha kasvavat rolli sotsiaalsetes ja ärilistes nähtustes. *Microblogging* teenused nagu Twitter mängivad olulist rolli Interneti infovahetuses, muutes võimalikuks sõnumite leviku minutitega. Käesolevas uurimuses analüüsitakse korduvalt edastatavate sõnumite (*retweet*) levikut Twitteris. Kasutades Latent Dirichlet Allocation mudelit teemade eristamiseks näitame, et kasutajate ja sõnumites sisalduvate teemade vaheline suhteline kaugus on lühem korduvalt edastatavatel sõnumitel. Kasutades otsustuspuid hindame teemapõhise *retweet* mudeli täpsust ja kasulikkust. Töö tulemusena näitame, et teemapõhine mudel on tugevama ennustusvõimega võrreldes baseline mudelitega, millest lähtuvalt väidame, et antud lähenemine on sobiv korduvalt edastavate sõnumite ennustamiseks ning edasiseks arenduseks.

References

[1] 200 million tweets per day; Published online at http://blog.twitter.com/2011/06/200-million-tweets-per-day.html. Last visited on November 1, 2011

[2] Iranian Election Protests; Published online at http://en.wikipedia.org/wiki/2009%E2%80%932010_Iranian_election_protests. Last visited on November 1, 2011

[3] Iran Revolution twitter account; Published online at http://twitter.com/#!/iranrevolution

[4] Iran's Twitter Revolution; Published online at http://www.washingtontimes.com/news/2009/jun/16/irans-twitter-revolution. Last visited on November 1, 2011

[5] Evaluating Iran's Twitter Revolution; Published online at http://www.theatlantic.com/technology/archive/2010/06/evaluating-irans-twitter. Last visited on November 1, 2011

[6] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 '03, pages 993-1022.

[7] Rupert G. Miller, Jr.. Beyond ANOVA: Basics of Applied Statistics. Chapman & Hall, ISBN 0412070111, 1997

[8] Social Media; Published online http://en.wikipedia.org/wiki/Social_media. Last visited November 1, 2011

[9] Anand Karandikar. Clustering short status messages: A topic model based approach. Masters Thesis, University of Maryland, 2010

[10] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue B. Moon: What is Twitter, a social network or a news media? WWW 2010: 591-600

[11] Danah Boyd, Scott Golder, Gilad Lotan: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS 2010: 1-10

[12] Daniel Sousa, Luís Sarmento, Eduarda Mendes Rodrigues. Characterization of twitter @replies Network: Are User Ties Social or Topical ? SMUC '10 Proceedings of the 2nd international workshop on Search and mining user-generated contents, pages 63-70.

[13] How Twitter was born; Published online at http://www.140characters.com/2009/01/30/how-twitter-was-born. Last visited November 1, 2011

[14] Dejin Zhao, Mary Beth Rosson: How and why people Twitter: the role that micro-blogging plays in informal communication at work. GROUP 2009: 243-252

[15] Kevin R. Canini, Bongwon Suh, Peter Pirolli. Finding relevant sources on Twitter. University of Berkeley, 2010

[16] Bernard J. Jansen, Mimi Zhang, Kate Sobel, Abdur Chowdury: Twitter power: Tweets as electronic word of mouth. JASIST 60(11): 2169-2188 (2009)

[17] What's happening; Published online
http://blog.twitter.com/2009/11/whats-happening.html
Last visited November 1, 201

[18] 200 million tweets per day; Published online
http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.
Last visited November 1, 2011

[19] Twitter co-founder Jack Dorsey rejoins company; Published online at
http://www.bbc.co.uk/news/business-12889048
Last visited November 1, 2011.

[20] Twitter has 100 million active users; Published online at
http://mashable.com/2011/09/08/twitter-has-100-million-active-users
Last visited November 1, 2011

[21] Kriti Puniyani, Jacob Eisenstain, Shay Cohen, Eric P. Xing. Social Links from Latent Topics in Microblogs. Carnegie Mellon University (2010)

[22] J. M. Kaufmann. Syntactic Normalization of Twitter Messages. In REU Site for Artificial Intelligence, Natural Language Processing and Information Retrieval Research Projects, 2010

[23] What are replies and mentions; Published online at
https://support.twitter.com/entries/14023-what-are-replies-and-mentions
Last visited November 1, 2011

[24] Bongwon Suh, Lichan Hong, Peter Pirolli, Ed H. Chi: Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. SocialCom/PASSAT 2010: 177-184

[25] Jianshu Weng, Ee-Peng Lim, Jing Jiang, Qi He: TwitterRank: finding topic-sensitive influential twitterers. WSDM 2010: 261-270

[26] Bruno Gonçalves, Nicola Perra, Alessandro Vespignani: Validation of Dunbar's number in Twitter conversations CoRR abs/1105.5170: (2011)

[27] Dunbar, R.I.M. Neocortex size as a constraint on group size in primates. Journal of Human Evolution 22 (6): 469–493 (1992).

[28] Akshay Java, Xiaodan Song, Tim Finin, Belle L. Tseng: Why We Twitter: An Analysis of a Microblogging Community. WebKDD/SNA-KDD 2007: 118-138

[29] Julia Hockenmaier. Lecture 6:Topic Models(Latent Dirichlet Allocation); Published online at cs.illinois.edu/class/sp10/cs598jhm/.../Lecture06HO.pdf.
Last visited November 1, 2011.

[30] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Meth Appl Sci 1: 300–307, 2007

[31] Pavel Zezula, Giuseppe Amato, Vlatislav Dohnal, Michal Batko. Similarity Search: The Metric Space Approach. ISBN 0387291466. Birkhäuser, 2006

[32] A New View of Statistics; Published online at
http://sportsci.org/resource/stats/ttest.html
Last visited November 1, 2011

[33] Analysis of Variance; Published online at
http://simon.cs.vt.edu/SoSci/converted/ANOVA/activity.html
Last visited November 1, 2010

[34] ANOVA; Published online at
http://www.psychstat.missouristate.edu/introbook/sbk27m.htm
Last visited November 1, 2011

[35] Ronald Aylmer Fisher. Statistical Methods for Research Workers, Macmillan Pub Co; 14th edition (June 1970)

[36] Stanford University SNAP; Published online at
http://snap.stanford.edu/data/twitter7.html
Last visited November 1, 2011

[37] What is Twitter, a Social Network or a News Media?; Publised online at
http://an.kaist.ac.kr/traces/WWW2010.html
Last visited November 1, 2011

[38] J. Benhardus. Streaming trend detection in twitter. (2010)

[39] Bernardo A. Huberman, Daniel M. Romero, Fang Wu: Social Networks that matter: Twitter under the Microscope. First Monday 14(1): (2009)

[40] Independent test samples; Published online at
http://www.wellesley.edu/Psychology/Psych205/indepttest.html
Last visited October 27, 2011

[41] Rupert G. Miller, Jr.. Beyond ANOVA: Basics of Applied Statistics. Chapman & Hall, ISBN 0412070111 1997.

[42] K. Kireyev, L. Palen, and K. Anderson. Applications of topics models to analysis of disaster-related twitter data. In NIPS Workshop on Applications for Topic Models: Text and Beyond, December 2009.

[43] John Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[44] Decision trees for business intelligence and data mining: using SAS Enterprise Miner, Barry De Ville, chapter 1, page 8, SAS Institute, SAS Publishing, 2006

[45] Nitesh. V. Chawla. C4.5 and imbalanced datasets: Investigating the e#ect of ampling method,probabilistic estimate, and decision tree structure. In Proceedings of the ICML'03Workshipkshop on Class Imbalances, 2003.

[46] Weka Mailing List; Published online at https://list.scms.waikato.ac.nz/pipermail/wekalist/2005-January/003301.html Last visited November 1, 2011