Chapter 1

Introduction

Online social networks have exploded into the lives of millions of people worldwide over the last decade, and their use has dominated the communication highways and facilitated the interconnection of the world in ways never before perceived possible.

These social networks imitate real-world social networks. Although most such platforms each provide a different service to collaboratively satisfy an array of different use-cases, they tend to all be based around the idea of 'friendships' (i.e. links between the user nodes in the social graph) and the sharing of information amongst friends.

Social networks like these have been available for around ten years now (with MySpace[1] launching in 2003 and Bebo[2] in 2005), but it wasn't really until Facebook's[3] worldwide launch in 2006 that social networks became the staple, ubiquitous norm that they are today. More recently, we have seen the introductions of Google's social network grown from its Buzz service, Google Plus[4], Pinterest[5], App.net[6], and many more. They make up a large part of the basis and meaning behind the ideas of Web 2.0, which describes the web as being primarily formed from user-generated content and encourages the sharing of such content.

Another component that helped in the dawn of Web 2.0 was the rise of *blogging*. A

[1] http://myspace.com

[2] http://bebo.com

[3] http://facebook.com

[4] http://plus.google.com

[5] http://pinterest.com

[6] http://app.net

blog ('web-log') is a time-based series of posts consisting of continuous pieces of text, photos, or other media, and is generally contributed to by a single author. Blogs are often based around one or a set of topics and are usually public - meaning that they are written with the intention of being read by others. Despite this, they are often a way in which the author can look back at their history of posts, acting more as a diary recording snapshots of the author's life.
Various blogging services exist on the web today, such as Medium[7], Wordpress[8], and Tumblr[9].

1.1 Twitter as a Social Network

Twitter[10] is an online social network, which launched in the summer of 2006 [21]. Since then, it has rapidly gained in popularity amongst several different user groups - teens and young people, casual users, celebrities, reporters, and so on - and within eight months had around 94,000 registered users [19]. Although Twitter has never been a direct competitor with Facebook, users tend to use the two sites concurrently for different purposes: whilst Facebook's focus is on providing many services at once (such as photo-sharing, commenting/endorsing of information, messaging, pages for businesses, groups, events, etc.), Twitter's is more on simplicity.

More specifically than just being an online social network, Twitter is a microblogging website. Whilst a blog, as mentioned, typically contains long posts, Twitter only allows its users to post short pieces of text, up to 140 characters in length [21] [18], called 'Tweets'. Thus, Twitter is a hybrid social network and blogging service and whilst each Tweet may only realistically be able to hold a couple of sentences, this system facilitates quick, timely, and 'real-time' *live* information-sharing amongst its millions

[7] http://medium.com

[8] http://wordpress.com

[9] http://tumblr.com

[10] http://twitter.com

of users [43]. Its idea is that short pieces of news will 'travel' faster and will be seen by more people more quickly than traditional news stories.

Although Tweets are limited to 140 characters in length, the inclusion of URLs is allowed. This enables further extension of Tweets through external websites, and supports the inclusion of links to images and videos. Twitter has encouraged this use-case by providing 'share' buttons for developers to embed in websites, and direct support for photo and video applications, such as TwitPic[11] and Vine[12].

Its simplicity has also helped its growth into the mobile domain, in which smartphone users are able to very quickly post updates about their lives, a piece of information they want to share, or a photo or video, and be able to post it *as it happens* directly from the news source or geographical location [8]. This has been especially useful in emergency situations worldwide, including the Haiti earthquake in 2010 [26], and 2011's Egyptian protests [37] and Thai flood [20].

Indeed, [32] used Twitter to build an earthquake-reporting system for Japan that outperforms the Japan Meteorological Agency in terms of its promptness of notification.

Use of Twitter is based around 'timelines' of Tweets, to which new Tweets are prepended as they are posted by users. The *home* timeline is the default view, in which Tweets from all of a person's subscribed-to users are placed. Timelines of an individual user contain only Tweets from that user, and are known as a 'user' timeline. Customisation of timelines is also possible through the use of Twitter lists, in which different users can be placed to categorise streams of Tweets from different sets of users.

[11] http://twitpic.com

[12] http://vine.com

1.2 Twitter's Social Graph and Information Subscription

As with many social networks, the structure of Twitter lies within the users and their connectivity within its social graph. However, unlike Facebook, whose social structure is made up of bi-directional 'friendships' between users, Twitter's primary social graph is made up more of mono-directional links between its users [13]. A person using Twitter can elect to *follow* another user, which subscribes the person to receive all of that user's Tweets to their home timeline. The set of users that follow a person are known as that person's *followers*, and the set of users that the person follows are the person's *friends*.
Therefore, if two users both mutually follow each other, then the link between them is bi-directional.

Whilst bi-directional links are common amongst communities of similar interests, friends, colleagues, and so on, mono-directional links are found more in situations in which less-influential users follow more-influential users, such as celebrities.

1.3 The Problem

A user who follows a set of other users can *generally* be said to find that set of users to produce more interesting information than those users that the user does not follow. However, despite that, not *all* information produced by an 'interesting' user is likely to be interesting, and yet *all* information produced by a Twitter friend will be received onto the home timeline.
Noise is a common problem in Twitter, and is the uninteresting information one might receive that conveys little interest. It is likely that most of the information received on Twitter *is* uninteresting [2], and this makes it very hard to distinguish the interesting information from the uninteresting.

Since people tend to use Twitter most in short sporadic moments, looking for a quick news fix, they do not have time to filter out noisy information. Thus, the presence of noise can dampen the experience of the user, making it much more difficult to find interesting information.

In addition, Twitter users typically exist within an information ‘bubble’. This is similar to the notion of the Google search bubble, in which the search engine uses previous results and search terms to only return information to a user based on what *it thinks* the user would find the most interesting and useful.
This results in the users not knowing which information exists beyond the confines of their bubble, and if they do not know it exists, they cannot know if it is of interest to them. Similarly, a Twitter user cannot follow all of the users he/she may find interesting, since he/she will not *know* of all the interesting users existing on the social graph.

How can users be exposed to *interesting* and *relevant* information, but without them having to know about it or look for it first?

1.4 Contributions

This thesis focuses on understanding information propagation, and how this combined with knowledge of the social structure of Twitter can assist towards solving the problem of identifying interesting and relevant information and determining it from the noise on Twitter. Whilst other work in the area has also looked into the notions of relevance and interest in online social networks, and Twitter in particular, none has addressed the problem in such a way as this.

Part of the outcome of this research are methods for effectively inferring interesting information and, indeed, ranking information by interestingness. The methods are validated in various ways to help highlight their strengths and weaknesses in performing inferences and appropriate use-cases.

The work addresses the problem area in that it helps towards solving the goal of identifying *globally* interesting information in Twitter. In addition, certain measures are taken in an attempt to address the idea of information relevance, which denotes how information interestingness is subjective, and thus different from user to user.

1.5 Thesis Structure

The rest of this thesis is structured as follows.

A background is provided as an introduction to some of the ideas behind the main research, which immediately follows this chapter, and includes a review of relevant literature across the range of topics addressed in the thesis.
Following this are chapters that contain research on Twitter's information propagation characteristics and its interesting and useful behaviours, the social structure of Twitter and the ways in which this is important for understanding the spread of information, and then on the research of the methodologies themselves, including validation and analysis of the results of this work.

The thesis ends with a general analysis and conclusion, and a discussion of potential future work in this area and leading on from this research.

Chapter 2

Background

One of the most widely-used features of Twitter is its inbuilt function for facilitating the spread of information within its social structure. This phenomenon is the basis for much of the research in this thesis and, when combined with the characteristics of Twitter's user graph, has many interesting attributes and behaviours associated with it.

2.1 Domain Context

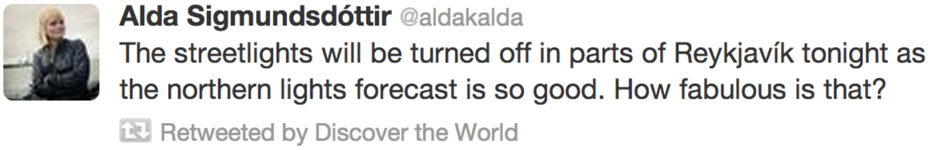
2.1.1 Information Propagation through Retweeting

The function of propagation in Twitter is known as *retweeting*, and is carried out by the Twitter users themselves. When a user views a Tweet that they believe to be particularly interesting, and believe it to also be interesting to his/her followers, then he/she can elect to retweet it, and thus pass it further through the social graph to that user's followers also. A Tweet that has been retweeted is known as a *retweet*, and it is clear that a Tweet which is retweeted will be made available to significantly more users than a Tweet that isn't retweeted [36] [22].

A retweet can be carried out in one of two ways: either through the use of Twitter's native retweet button, or manually.

The retweet button is displayed along with each Tweet in a Tweet timeline which, when clicked, immediately creates a new retweet containing the verbatim content of the original Tweet and automatically sends it on to the retweeter's followers.

The user who created the original Tweet is credited as the author on the recipients' timelines, with an indication of who carried out the retweet itself. Thus, users who follow the retweeter will see a Tweet appear in their home timeline from someone that they may not directly follow.



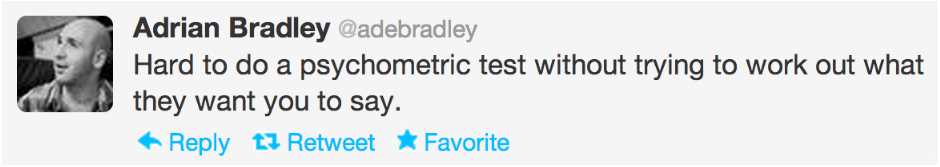
Alda Sigmundsdóttir @aldakalda

The streetlights will be turned off in parts of Reykjavík tonight as the northern lights forecast is so good. How fabulous is that?

Retweeted by Discover the World

Figure 2.1: ***A retweeted Tweet.***

The manual approach involves physically copying the content of the Tweet to be retweeted and pasting it into a new Tweet, usually with the text '`RT @<username>:`' pre-pended, where `RT` stands for **ret**weet and `<username>` is the username of the author of the original Tweet. This method allows for annotating the original content of the Tweet (for example, to provide an opinion on the Tweet contents), producing a *modified* Tweet, which can sometimes be pre-pended with `MT` rather than `RT`.



Adrian Bradley @adebradley

Hard to do a psychometric test without trying to work out what they want you to say.

Figure 2.2: ***The retweet 'button' in context.***

Each Tweet has a retweet count associated with it, which is the raw representation of the number of times that the Tweet has been retweeted using the retweet button method. Since the manual retweet technique is more community-driven, there is no official way to include these as part of the retweet count of the original Tweet. However, since the manual method is typically only really used with the aim to annotate or modify the Tweet in some way, the resultant 'retweet' is no longer a real representation of the content of the original Tweet anyway, and so should not be counted as such.

It should be noted that Twitter users may choose to make their account ‘protected’. A person who has a protected account will still have a publicly-visible profile (displaying a name, username, bio, and so on), but their Tweets and other information (such as the followers and friends lists) are hidden from users that aren’t followers of the person. Potential followers of a protected account must *request* a followship, which can then be accepted or rejected by the protected account holder.
Since Tweets from a protected account are only visible to approved followers, the retweet button is unavailable for them to disseminate the Tweet any further than the author’s immediate local follower network. However, since the manual retweet method does not rely on the button and isn’t governed by Twitter, a protected account’s Tweets can still be retweeted in this way.

In a similar way to Facebook supporting the endorsement of information found on its site by inviting users to ‘like’ a piece of content, retweeting is effectively a *vote* or endorsement for a Tweet on Twitter. In both cases, the number of likes and number of retweets is visible to the platforms’ respective users, and so this provides some insight into the *popularity* of the information.

2.1.2 Retweets and the Social Graph

The social graph of Twitter is constructed, like in other online social networks, by edges between users, partially emulating real-life social interactions between humans. The growth of social media has encouraged more dense communication between users all over the world, who would not previously be able to be in direct contact with one another in this way.

Derived from this, Stanley Milgram’s finding of “six degrees of separation” [25], which defines that people are usually no more than six hops away from each other on the ‘real-life' social graph, was found to be an overestimate when it comes to the analysis of the structure of OSNs by [5], who found that the average 'distance' observed in Facebook's entire 721 million-node graph in 2011 was only around 4.7 hops. This implies that denser links between users and larger communities that apparently manifest themselves in OSNs create a smaller 'world' than that experienced in reality.

In each of Milgram's experiments participants passed a message to one another, at each stage only passing to other people that they actually *know*, in the hope of the it reaching a single intended recipient. This meant that people could use acquaintances in other geographic locations to transfer the message from community to community.
Twitter supports a similar propagation mechanism in the fact that retweets can themselves be retweeted; this is a focus of some of the earlier research in this thesis.

This behaviour provides further penetrative 'depth' of the information through the social network away from the source user in addition to the spread 'width' made by the initial retweets. Although retweeting is not carried out with the aim of information reaching any particular final user (or set of), as with Milgram's experiment, this phenomenon allows retweets to 'travel' between 'online communities' of users.

As with real-life social networks, communities of users in OSNs are also a common feature [34].
In Twitter, these communities are typically small to begin with and are based on a topic of interest or around a more influential user. As more Tweets are produced from within the community, further links are made to interconnect the community's users, producing a growing 'swarm' of interest around the initial topic or user [19].
As further users begin associating themselves with this community, its audience becomes more widespread and the community grows. This concept is discussed in greater length by [19], who also experiment further with communities and describe them as compact groups of users connected by dense follower links.

In more dense communities, Tweets can be made available to many users immediately after they are published, since many of the links between users are shared. This means

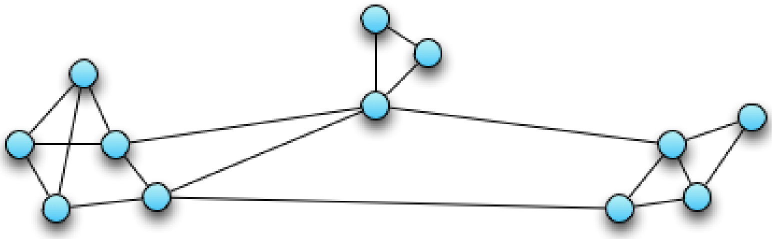


Figure 2.3: ***A hypothetical group of user communities.***

that any retweets that occur within communities are likely to have a lot of *redundancy*, in that many of the retweets will be sent to users who have already seen the Tweet. Although Twitter prevents this information duplication by not showing the retweets of Tweets that have already appeared on a user's timeline, it does increase the chance of the Tweet making its way out of the community.

Retweets amongst users within a community are likely to be common, due to the shared-interest nature of communities, and some users can provide 'bridges' by being active in more than one community. In these cases, Tweets can be passed between the communities through retweets by the bridging user. If there are many users sharing communities, then there are many more avenues available for propagation to occur down, causing a high level of information throughput. If there are fewer bridges, then there is more of a bottleneck between the communities, hindering the information spread.

[19] also finds that communities can be formed from different types of people, such as those who Tweet frequently and have many followers, and those who contribute very little and have few followers. Those with many followers and many friends receive lots of information and have the potential to spread information further than those with fewer inward and outward edges. Studies in the behaviour of different types of users in Twitter is done more thoroughly in [21], which defines 'broadcasters' (users with many followers and few friends) and 'miscreants' (users with few followers but many friends) and their roles in information propagation.

Users that retweet the interesting information from a source user to others, who do not follow the source user and so would not naturally receive the information, are effectively acting as information *filters*. By not following the source user, a person might still receive the interesting information through these filters, but will not receive any of the 'noise'. Thus retweeting means that friends of a user become useful filters of information for users further 'downstream' and retweeted information can be said to have a higher *credibility* than Tweets that aren't retweeted [8].

2.1.3 User Influence

Just as there are different types of user *behaviours* on Twitter, as mentioned in the previous section, the are also users of different *influence* levels [31].

Much research has gone into user influence, including on how this might be detected [40], and influential users are generally found to be those that have a greater impact on Twitter's social network [6] and that usually have significantly more followers than an average user. Influential users tend to have a high persuasion over other users, relating *influtentials* in Twitter to those who are also influential in the real world as part of traditional communication theory [9], and therefore many Twitter influentials are the accounts belonging to real-world celebrities.

As with real-world celebrities, Twitter influentials are those with many 'influenced' followers, or fans, which are the users who have the strongest agreeable opinions of the influential. As a result, an influential user has a greater number of followers who are interested in the information produced by the user, and is therefore more likely to receive more retweets than less influential users.

Although influence level is partly derived from the follower count of the user, it should be noted that a user with high in-degree on the social graph[1] does not necessarily imply a high level of influence. An 'active' audience of users who reply, retweet, and interact

[1] In-degree: many followers

are more indicative of an influential user [7]. This is especially true since a user can gain more followers through campaigns such as '#teamfollowback'[2] or by following 'out of politeness', in which a user will follow another user back as an act of politeness, but these users tend to have *both* high in- and out-degree and invoke less interactivity amongst their followers, which are not necessarily characteristics of an influential user [9].

Klout[3] is a web service that attempts to review a user's social media influence by assigning users a Klout Score. Their website declares that this score, which ranges from 0 to a maximum of 100 and whose generation algorithm is kept private and unpublished [13], is determined from a variety of 400 sources taken from eight different social media platforms, and which *also* seems to take interactivity between users as the primary indicator [3]. Additionally, the service indicates the topics a user is influential about, with the general idea being for organisations to check up on which users are influential for marketing purposes, but also to highlight the users that should be replied-to at a higher priority.

2.1.4 Twitter as an Information Retrieval System

From a high level, Twitter is essentially just a variety of information-retrieval system, which people can utilise to produce and consume information when required. In traditional information-retrieval systems, such as search engines and library systems, keywords and search terms are common ways for describing the type of information the user would like to receive back. The system would then search a database or archive for what it believes is relevant information, *based* on these 'retrieval parameters', and return results to the user ordered usually by the estimated relevance of the articles [4].

Information quality is also reliant on the expected reading effort of the returned documents. The character precision-recall metric was introduced by [4] by way of demon-

[2] Users associate themselves with #teamfollowback to imply they will return all followships.

[3] http://klout.com

strating the tolerance-to-irrelevance ratio. The general mechanism for this ratio is to do with users reading a document passage; the point at which this ratio is reached is when the user stops reading the particular passage and moves to the next whole document, since they assume the rest of the document is also irrelevant to them.

Therefore, the more effective the information retrieval system is in displaying high-quality information, the lower the chance that this ratio is reached by the user.
It is comparable that a Twitter user viewing Tweets from a user they are following may get to the point where he or she reaches this ratio (i.e. is beginning to get bored or find the Tweets irrelevant) and decides to unfollow the friend. Similarly, the more effective the user is when selecting people to follow in the hope of receiving interesting information, the less likely it is that the user will remove these friends.

Whilst Twitter does not support the use of keyword searching for its primary information delivery method, it does lend its users some control over the type of information they wish to receive. As mentioned previously, users receive all of the Tweets from everyone that they follow onto their home timelines. Thus, by selecting users to follow, a person is effectively describing and implicitly indicating the type of information he/she would like to receive, and by editing their friends list (either by adding new followers or pruning existing ones) he/she can alter this indication.

Despite this control, it is still unlikely that users will achieve a perfect Twitter experience due to the presence of *noise* [2]. As discussed in the Introduction, this problem stems from that although a person follows users they consider to be interesting, it is often the case that not *all* information produced by interesting users will be interesting itself.

2.1.5 Information Quality, Popularity and ‘Interestingness’

Information-retrieval systems typically use some measure of information *quality* when determining which documents to return to a user and also when deciding on the *order* the documents should be displayed in. This ‘quality’ is subjective in that different systems use a variety of different algorithms for deducing quality, usually based on the level of *interest* in each of the available documents (such as Google’s Page Rank algorithm and Amazon’s recommendation algorithms), but also in that the level of quality itself depends on the user itself requesting the information.

In the case of Google’s Page-Rank, the algorithm uses multiple cues to determine who the user is, their interests, past searching habits, links clicked, and so on, to return *relevant* information, which is incidentally one of the causes of the aforementioned Google search bubble.

Amazon’s recommendation algorithms analyse a user’s past item views and purchases and cross-matches these against trends based from users who also looked or bought similar items. Amazon is then able to accurately determine the type of items a customer are interested in purchasing, and can send emails to that customer with personalised recommendations.

Thus, information quality is essentially a function of information interestingness and information relevance, which are both related to the concept of *effective stimulation* [38] discussed later.

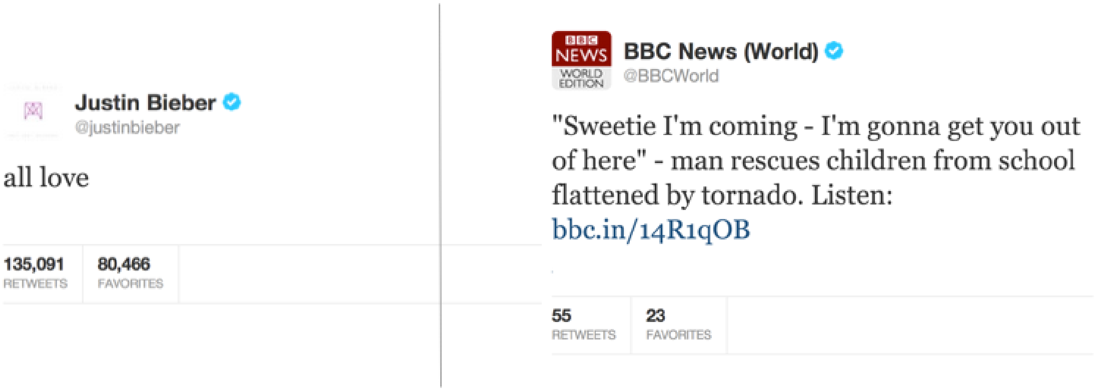
Twitter uses no such metrics to deliver information to its users, relying on the users themselves to implicitly ‘choose’ the information they want to receive - it is an information retrieval system and not a recommendation system. Additionally, information is always displayed in chronologically-ordered timelines, with new Tweets being continuously inserted at the top as they occur. Twitter does not try to indicate interesting Tweets on the timeline which means that the interesting information is shown at equal value alongside the ‘noisy’ Tweets, causing the difficulties in identifying the interesting information as has been mentioned previously.

Indeed, the recent TechCrunch article from October 2013, “Twitter Quitters And The Unfiltered Feed Problem”[4] talks at more length about this particular phenomenon, and

[4] http://techcrunch.com/2013/10/05/sorry-my-feed-is-full

helps highlight the problem area of this work more clearly.

The retweet count of a given Tweet is a useful metric in inferring a Tweet's *popularity*. If a Tweet is retweeted 10 times, then ten people have taken the time to read that Tweet, decide it is worth sharing, and then actually retweet it [35]. This user (and the other nine retweeters) may have found the Tweet interesting, yet it should be noted that although the count can be used as a measure of popularity, as a function of the influence of the Tweet's author, the retweet count alone cannot be used as a measure of how interesting the Tweet actually is [27]. For example, it is inappropriate to say that the first Tweet in Figure 2.4 is so significantly more *interesting* than the second, although it is clearly more popular since Justin Bieber is an extremely influential Twitter user.



Justin Bieber
@justinbieber

all love

135,091 RETWEETS | 80,466 FAVORITES

BBC News (World)
@BBCWorld

"Sweetie I'm coming - I'm gonna get you out of here" - man rescues children from school flattened by tornado. Listen: bbc.in/14R1qOB

55 RETWEETS | 23 FAVORITES

Figure 2.4: ***Example of Tweets with significantly different retweet counts.***

Whilst the work in this thesis does not aim to build an accurate retweet-predictor, this does become a basis for some of the work in later chapters.

[35] identifies the same problem of 'noisy' Twitter timelines and discusses methods for predicting *popular* Tweets using a J48 decision tree classifier, based on the likelihood of the Tweet being retweeted by a particular user. Although the authors address information relevance from a user-centric point of view, the validations of whether a prediction of a retweet occurring for a given Tweet is actually indicative of the *interestingness* of said Tweet do not perform particularly well.

A retweet-prediction model based on a factor graph model is introduced by [39] to determine how retweetable a Tweet is on a global scale. A precision of just under 29%

is achieved in predicting if a Tweet will be retweeted, but no mention is made of how this relates to how *interesting* the information is.
Another study into retweet prediction was carried out by [42], in which a trained probabilistic collaborative filter model (named 'Matchbox') was used to determine the useful features in making the predictions. As with the previous study, the research focuses on a retweet *probability*, which is a binary decision made by one particular user. The methodology is not aimed at the inference of interestingness, and simply determines that the most relevant features for accurate decision predictions are the author of the original Tweet and the retweeter.

Inversely, [33] and [17] predict the *type* of messages that are likely to be retweeted further, the latter using a logistic regression to both predict an individual retweet decision and a retweet *volume*. The methods do not apply these notions to how interesting the information actually is, achieve low recall and the multi-classifications seems only to perform well on very unpopular or very popular Tweets. It is made clear, however, that the retweet volume of a Tweet is useful in denoting Tweet *popularity*.

[30] uses a passive-aggressive machine-learning algorithm to make binary predictions on retweet decisions and cited that social features - for example, number of followers of the author, frequency of Tweeting, etc. - were the largest factors in the performance, and [27] uses a logistic regression, partly using a dataset published as part of another paper by the same authors as [30], to predict retweet decisions in order to address information interestingness. However, little effort is made to define interestingness or, indeed, validate that the inferences towards this are accurate and correct.
A logistic regression is again used by [44] for predicting binary retweet behaviours with the focus on information propagation in disaster scenarios, and [29] showed that conditional random fields can perform better than logistic regressions than when modelling retweet behaviour in the same way.

Since the above papers only effectively consider a prediction of retweet outcome, which is a binary decision, it is hard to relate this to more of a global interestingness, aside from stating that a retweet implies the retweeter's relative interest in the Tweet. However, a retweet count, as mentioned above, is inappropriate as an indicator of *magnitude* of interest, and so the research into predicting individual retweet decisions cannot be used as a basis for this. Additionally, not much emphasis is placed on how well the techniques work 'on-the-fly'; many of the methodologies discussed require several features that may take a long time to collect and compute, making them unsuitable for use as part of quick and useful interestingness evaluations.

The idea of Tweet scoring and retweet *count* predictions is introduced by [15], who used their methodologies to produce a system[5] enabling users to compile Tweets in ways that are predicted to achieve the most retweets. The predictions are based on averaging the score, derived through a linear regression, of different components of a user's Tweets (such as the inclusion of a particular hashtag), so that when a Tweet by the same author is next constructed, the various components of the new Tweet can be compared against the scores of the counterparts seen in previous Tweets. The value produced through this method is then used to generate an expected retweet count as part of a comparison to the user's average ('baseline') achieved retweet count at this point in time, and was shown to perform well on influential Twitter users.

However, the methods described do not take into account fluctuations in the social graph, particularly in the case of less-influential Twitter users, who's local networks are prone to more frequent changes. Additionally, they rely on enough previous Tweet and temporal information on the user to be evaluated, and do not relate the resultant score to any type of interestingness metric in the context of highlighting it from amongst noise.

Alonso et al. ([2]) also use 'scoring' to address interestingness, focusing more on determining *uninteresting* content, by assigning Tweets an integer score out of five. Although the authors initially attempted to train a decision tree classifier on a set of 14 features, they settled on classifying a Tweet as 'possibly interesting' if it simply

[5] https://sites.google.com/site/learningtweetvalue/home

contains a URL, and otherwise classify it as ‘not interesting’. Although the authors did then further classify the possibly interesting Tweets, by studying the magnitude of the crowdsourcees used to evaluate the Tweets that found them interesting, and then classifying Tweets based on them containing a particular type of named entity - for example, a person’s name, a place or brand name, and so on - the categorisation system is too coarse and is not capable of representing the many different types of Tweets seen on Twitter.

Additionally, despite achieving relatively high accuracy in this particular area, the methods are not suitable for assessing Tweets on a general or user-specific level, especially since Tweets that don’t contain URLs might still contain interesting content.

An interesting study is described by [23], in which a clustering algorithm is used, taking into account the retweet count of a Tweet and how this is related to the popularity of the source user, to determine information quality. Although this work is more similar to the research discussed later in this thesis than others, the scoring is quite simple and the author’s use-case seems limited to that of identifying the most important Tweets surrounding a particular event (such as the death of Michael Jackson).

Additionally, the authors do not make any effort to verify their results in any way, aside from comparing the Tweets determined to have a high quality by each of their two assessed methodologies.

2.1.6 Precision and Recall

Precision and recall are two metrics that are often used simultaneously to verify the performance of a method or procedure, with the usual goal being to maximise both. The metrics are used for validating *accuracy* in different ways, yet they can be applied to other purposes also and are useful in describing the notion of interestingness in Twitter.

The precision and recall measures are talked about somewhat in Twitter- and retweet-based literature. These pieces tend to only analyse the measures on their own work when applied to Twitter rather than on any more global scale. Certainly, there is less in the literature on the subjects of precision and recall with regards to retweeting in general.

The idea of assessing the credibility of information is introduced in [8], in which the authors demonstrate methods of measuring the credibility of ‘news’ and ‘chat’ Tweets. In this case, retweeting is seen as a possible measure of a Tweet’s credibility, since users typically only retweet information they see as interesting or useful. The authors use a logistic regression on a set of features derived from each Tweet in order to classify its credibility.

The precision and recall metrics are used to verify the different aspects of the paper’s results. In particular, they are applied to the classification of assessing credible information (and users) in order to calculate how well classified the information is. A higher precision, therefore, shows that their model has accurately classified most of the total information classified as either credible or non-credible.

$$Precision = \frac{\text{Number of correct classifications}}{\text{Number of total classifications made}}$$

$$Recall = \frac{\text{Number of correct classifications}}{\text{Total number of potential classifications}}$$

On a similar note, [17] discusses the notions of precision and recall more generally. The authors discuss the problem regarding the balance of information received by Twitter users. Having too few friends reduces the number or interesting posts received (i.e. low recall); having too many friends may cause information overload and is likely to include a lot of noise (i.e. low precision). This issue is used, instead of to validate results, as a basis for the work; predicting the Tweets that are most popular and will be retweeted the most.

In addition, precision and recall are used to compare the method to two other baselines; the TF-IDF score and *Retweet Before*, which uses the fact that if a Tweet in the training

data has been previously retweeted, then it's likely to be retweeted again. The two metrics are also used to compare results when certain features are removed from the classifier. For example, showing that without using a 'user retweet' feature, the precision and recall remain significantly higher than when removing other features, meaning that this feature does not contribute highly to the performance.
More specifically, precision and recall are used in a similar way to in [8]; except rather than looking at the number of classifications made, the authors use the number of predicted retweets.

[7] discusses a proof of concept for detecting influential users in one of two categories; evangelists or detractors. Precision and recall, in this case, are used slightly differently:

$$Precision = \frac{\text{Number of influential users retrieved}}{\text{Number of users retrieved}}$$

$$Recall = \frac{\text{Number of influential users retrieved}}{\text{Total number of users}}$$

The concept is taken further through the use of another metric, the *Mean Average Precision*, which is used to denote an influential user as being a detractor or an evangelist. A high precision, in this case, would imply a large proportion of influential users are classified correctly and a high recall means that most of the influential users existing in the entire dataset have been classified. The final results then show the precision and recall values for detecting evangelists and detractors in both follower/following networks and interaction networks. Both precision and recall improved when the size of the set of highest classified influentials increased (i.e. the top set of influential users).

[28] presents a method for the automatic classification of Twitter information to determine if a document is positive, negative or neutral in sentiment. In this case, the authors replace precision with *accuracy* and recall with *decision*, since they are using many classes instead of a binary classification, and define them as the following:

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Number of all classifications}}$$

$$Decision = \frac{\text{Number of retrieved documents}}{\text{Number of all documents}}$$

The accuracy is measured across the classifier's decision, and the $F_{0.5} - measure$ is then calculated based on these values instead in order to show that the classifier works well when the dataset size is increased.

As well as a good news source, Twitter is also used as an informational, user-contributed source on world events. [24] introduces a system, TwitInfo, which can be used for detecting, summarising and visualising events from Tweets. The authors looked at football match footage, web content, and earthquake survey data, and manually annotated major events in each to produce ground truth sets. These would be use to compare and contrast the results produced by their event detector using the following definitions of precision and recall:

$$Precision = \frac{\text{Number of events detected were from ground truth set}}{\text{Total number of events}}$$

$$Recall = \frac{\text{Number of events detected}}{\text{Number of events in ground truth set}}$$

With these definitions set, the authors were then able to easily calculate precision and recall for their algorithm.

For the work in this thesis, interestingness of information is the performance metric used to describe information quality, and thus precision and recall for any particular user in the scope of this thesis can be defined as follows:

$$Precision = \frac{\text{Number of interesting Tweets received}}{\text{Total number of Tweets received}}$$

$$Recall = \frac{\text{Number of interesting Tweets received}}{\text{Total number of all interesting Tweets}}$$

where *received* means that the Tweet has arrived on the user's home timeline, but does not imply that the user has *read* the Tweet.

Therefore, a user following many other users will receive lots of interesting information onto their home timeline in amongst lots of noise; resulting in a reduced precision and higher recall. Another user might follow a very select few other users who are of direct interest, and thus will experience high precision, but low recall.
These metrics are therefore useful in describing the concepts of noise and interestingness, and are consistent with their respective definitions in that users will achieve an optimum Twitter experience if both precision and recall are maximised.

Zadeh et al. ([41]) defined bespoke definitions of precision and recall, yet also in the domain of interesting information on Twitter. Although the authors identify the need for users to be able to discover other users of interest and declare that Twitter does, in fact, have a 'high precision' of interesting information, they admit to using a very coarse set of possible interest categories and is only based on *overlapping* interests rather than addressing the interest-noise ratio more concerning the research in this thesis. Additionally, clicks on URLs by users are the only means by which to measure this interestingness, and Tweets with URLs are usually the most interesting type of information [2].

2.2 Collecting Twitter Data

Most of the analytical work in this thesis relies on various data being collected from Twitter. Twitter provides an API for developers in order to facilitate the production of applications for its platform, but also for research purposes. It permits interfacing with many components of Twitter's service, such as posting and retrieving Tweets, interacting with other users (e.g. creating new friendships), and most of the features that Twitter's service itself provides to its users.
The API encourages use of the OAuth[6] authorisation framework to handle access[7],

[6] http://oauth.net

[7] https://dev.twitter.com/docs/auth

allowing Twitter to keep track of applications and each application's access privileges and rate limits[8].

Twitter's traditional REST API, v1[9], provided many useful endpoints for data collection and allowed each OAuth-authenticated application 350 hourly POST and GET requests[10].

In June 2013 Twitter officially deprecated v1 of its REST API, forcing use of its new v1.1 API[11]. The new version contains many of the same resources[12] as the original, but workarounds are required to get the results as some of the endpoint requests possible through v1. Additionally, new rate-limit policies were introduced, allowing more limited and controlled access to most of the available resources.

Since the work in this thesis was ongoing over this switch-over date, the initial work utilised API v1, and the latter work API v1.1, causing some changes to some of the data-collection methodologies as the thesis progresses. Descriptions of the data-collection in each relevant part of the thesis reflect this change, where appropriate.

2.3 Research Motivation

The motivation for the work in this thesis lies in the need to distinguish interesting information from noisy Tweets in Twitter, the latter of which is the problem area identified over the previous sections of this thesis.

It has been made clear that the retweet count of a Tweet cannot reliably be used as a measure of interestingness, especially in the context of influential users, who naturally achieve significantly more retweets than average users, but which does not imply that the information they produce is of a higher quality or interest level.

[8] https://dev.twitter.com/docs/rate-limiting/1.1

[9] https://dev.twitter.com/docs/api/1

[10] https://dev.twitter.com/docs/rate-limiting/1

[11] https://dev.twitter.com/blog/api-v1-retirement-date-extended-to-june-11

[12] https://dev.twitter.com/docs/api/1.1

As a result, the retweet count alone cannot be useful in distinguishing interesting information from noise in a timeline of mixed Tweets from different users with different levels of influence - some further metric is required to make this distinction.

This thesis covers the procedure and research behind a methodology that determines and ranks information on Twitter through inferences of interestingness that allows the more interesting information to be brought forward.

Bibliography

[1] Stuart M. Allen, Gualtiero Colombo, and Roger M. Whitaker. Uttering: social micro-blogging without the internet. In *Proceedings of the Second International Workshop on Mobile Opportunistic Networking*, MobiOpp '10, pages 58–64, New York, NY, USA, 2010. ACM.

[2] Omar Alonso, Chad Carson, David Gerster, Xiang Ji, and Shubha U Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[3] Isabel Anger and Christian Kittl. Measuring influence on twitter. In *I-KNOW*, page 31, 2011.

[4] Paavo Arvola, Jaana Kekäläinen, and Marko Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010. 10.1007/s10791-010-9133-9.

[5] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. *CoRR*, abs/1111.4570, 2011.

[6] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Identifying 'influencers' on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*, 2011.

[7] Carolina Bigonha, Thiago N. C. Cardoso, Mirella M. Moro, Virgílio A. F. Almeida, and Marcos A. Gonçalves. Detecting Evangelists and Detractors on Twitter. In *WebMedia*, 2010.

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[9] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

[10] MJ Chorley, GB Colombo, SM Allen, and RM Whitaker. Better the tweeter you know: social signals on twitter. In *International Conference on Social Computing, ASE/IEEE*, pages 277–282, 2012.

[11] A. Clauset, C. Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *ArXiv e-prints*, June 2007.

[12] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[13] Chad Edwards, Patric R. Spence, Christina J. Gentile, America Edwards, and Autumn Edwards. How much klout do you have? a test of system generated cues on source credibility. *Computers in Human Behavior*, 29(5):A12 – A16, 2013.

[14] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, pages 3–3, Berkeley, CA, USA, 2010. USENIX Association.

[15] Sean Gransee, Ryan McAfee, and Alex Wilson. Twitter retweet prediction, 2012.

[16] Oliver Hein, Michael Schwind, and Wolfgang König. Scale-free networks. *WIRTSCHAFTSINFORMATIK*, 48:267–275, 2006. 10.1007/s11576-006-0058-2.

[17] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM.

[18] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *ArXiv e-prints*, December 2008.

[19] Akshay Java, Xiaodan Song, Tim Finin, and Belle 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*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[20] Alisa Kongthon, Choochart Haruechaiyasak, Jaruwat Pailai, and Sarawoot Kongyoung. The role of twitter during a natural disaster: Case study of 2011 thai flood. In *Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET'12:*, pages 2227–2232. IEEE, 2012.

[21] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, WOSP '08, pages 19–24, New York, NY, USA, 2008. ACM.

[22] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[23] H Lauw, Alexandros Ntoulas, and Krishnaram Kenthapadi. Estimating the quality of postings in the real-time web. In *Proc. of SSM conference*, 2010.

[24] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI'11*, pages 227–236, 2011.

[25] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[26] Sidharth Muralidharan, Leslie Rasmussen, Daniel Patterson, and Jae-Hwa Shin. Hope for haiti: An analysis of facebook and twitter usage during the earthquake relief efforts. *Public Relations Review*, 37(2):175 – 177, 2011.

[27] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *In: Proceedings of the ACM WebSci'11*, pages 1–7. ACM, June 2011.

[28] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[29] Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. Retweet modeling using conditional random fields. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 336 –343, dec. 2011.

[30] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

[31] D. Quercia, J. Ellis, L. Capra, and J. Croweroft. In the mood for being influential on twitter. 2011.

[32] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[33] Bongwon Suh, Lichan 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*, pages 177 –184, aug. 2010.

[34] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[35] Ibrahim Uysal and W. Bruce Croft. User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2261–2264, New York, NY, USA, 2011. ACM.

[36] W. Webberley, S. Allen, and R. Whitaker. Retweeting: A study of message-forwarding in twitter. In *Workshop on Mobile and Online Social Networks (MOSN), IEEE*, pages 13 –18, sept. 2011.

[37] Christopher Wilson and Alexandra Dunn. Digital media in the egyptian revolution: Descriptive analysis from the tahrir data sets. *International Journal of Communication*, 5:1248–1272, 2011.

[38] Yunjie Xu. Relevance judgment in epistemic and hedonic information searches. *Journal of the American Society for Information Science and Technology*, 58(2):179–189, 2007.

[39] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1633–1636, New York, NY, USA, 2010. ACM.

[40] Aron Yu, C Vic Hu, and Ann Kilzer. Khyrank: Using retweets and mentions to predict influential users, 2011.

[41] Reza Bosagh Zadeh, Ashish Goel, Kamesh Munagala, and Aneesh Sharma. On the precision of social and information networks. In *COSN'13*. ACM, 2013.

[42] Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, volume 104, pages 17599–601. Citeseer, 2010.

[43] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 243–252, New York, NY, USA, 2009. ACM.

[44] Jiang Zhu, Fei Xiong, Dongzhen Piao, Yun Liu, and Ying Zhang. Statistically modeling the effectiveness of disaster information in social media. In *Global Humanitarian Technology Conference (GHTC), 2011 IEEE*, pages 431 –436, 30 2011-nov. 1 2011.