The mission of this chapter is two-fold. First, we give a short description on the Social Network platform which has been used as the input data space for this research project and also explain its value as an information source. Then, we refer to the methods and tools that have been implemented or incorporated in our data collection and processing routines. Of course, this project is not about collecting data; however, our research would have been impossible without it.
Chapter 3. The Data – Characterisation, Collection and Processing

3.1 Twitter: A new pool of information

A blog, a term derived from synthesising the words web and log, is a – or occasionally a part of a – website supposed to be updated with new content from time to time.\(^1\) A blog’s content is based on posts identified primarily by their unique Uniform Resource Locators (URLs). Those posts can vary from random or opinionated articles to the embedding of songs, videos or other multimedia types.

Microblogging is a compact form of blogging, which can be seen as a ‘technological evolution’ of blogs, and surely is one the best complementary tools for the blogosphere. It became very popular when web services like Tumblr\(^2\) and Twitter\(^3\) kicked off. The basic characteristic (or limitation) of microblogs is that authors are restricted to a specific number of characters per post. For the experimental derivations of the work presented in this Thesis, we collect and use Twitter content.

Twitter was created in March 2006 and by 2011 the number of its users reached 100 million worldwide with a reported new accounts per day ratio equal to 460,000 for February 2011.\(^4\) The registered users of Twitter are allowed to post messages, commonly known as tweets (see Figure 3.1), which can reach a maximum of 140 characters. By default an account is public not private, \(i.e.\) everybody can see the tweets published by this account without any authorisation by its owner. Users have also the option to follow other users and therefore see their tweets on their time-line, the main part of Twitter’s web interface. For public accounts, becoming a follower does not need any type of approval from the followee; private accounts are able to authorise their followers. In that sense, each user – with a public or a private account – has a set of people who he follows (followees) and a set of people who follow him (followers). The only distinction is that connections in public accounts can be seen as one-sided relationships, whereas connected private accounts form a 2-sided – but hidden – relationship.

Users can mention or reply to each other in their tweets – by using the character ‘@’ followed by the username of the user to be mentioned – and therefore, conduct electronic conversations. They can also reproduce the tweet of another user, an action known as retweet.

Twitter incorporates topics in tweets; any word that starts with a hash (#) is perceived as something that denotes a topic (e.g. #OlympicGamesLondon2012).

3.1.1 Characterisation of Twitter – Why Twitter content is important

“If you think about our society as being a big organism this is just another tool to look inside of it”, says Noah Smith about Twitter [145]. In this section, we refer to several works which justify this statement in practice and highlight the importance of Twitter content.

Twitter’s social network is comprised by two underlying sub-networks: a very dense one made up by followers and followees and a sparser one, where more close and probably real friends participate. The latter one is more influential as expected, yet the former reveals the great degree of connectivity in Twitter [71]. In this space of data, three distinct categories of users have been identified: broadcasters, acquaintances (who are the vast majority) and miscreants (stalkers or spammers) [85].

Interestingly, Kwak et al. show that Twitter deviates from known human social networks having a non-power law follower distribution, a short effective diameter and low reciprocity\(^5\) [87]. Twitter users (in the United States) are also a highly non-uniform sample

\(^5\) Reciprocity is a quantity that characterises directed graphs by determining the degree of mutual connections – double links with opposite directions – in the graph [53].
of the population and represent a highly non-random sample of the overall race/ethnicity distribution [113].

An early work of Java et al. has shown that people use microblogging to talk about their daily activities and to seek or share information. An additional derivation presented in this paper was that users with similar intentions are more likely to be connected with each other [77]. As a result, in this environment, conversationality is also promoted. But Twitter is not only a means for public interaction; it finds use in collaborative scenarios as well [69]. By posting tweets or conducting conversations on Twitter a feeling of connectedness is sustained in working environments [172] and several aspects of educational schemes are improved [61]. Twitter encourages free-flowing and just-in-time interaction between and among students and faculty [36], finds great applicability during conferences [135] and is a great tool for promoting second language active learning methods [12].

The majority of topics on Twitter (approx. 85% in 2010) are headline news or persistent news in nature [87] and methods have been developed enabling the detection of breaking news [144]. Twitter also facilitates communication between communities with different political orientations. Similarly to real-life, an extremely limited connectivity between left- and right-leaning users is reported [28].

Furthermore, Twitter is a rich source for opinion mining and sentiment analysis [121]. Consumer opinions are public for companies to track and then adapt their overall branding strategies based on this intelligence source [76]. Polls can be easily conducted regarding consumer confidence and political opinion [27]. TV stations have started to combine their broadcasts with social networks to improve interaction. In particular, Twitter can provide a better understanding of sentiment’s temporal dynamics as a reaction to a political debate [33]. Predicting the result of elections based on Twitter content is a much harder problem to solve, still some preliminary approaches have already been proposed on this topic [54, 88, 102, 109, 164].

Due to the fact that microblogging usually expresses a real-time state of its author, tweets are considered more valuable than other media for connecting this information to personal experiences or situations of the users [172]. Exploiting this fact and Twitter’s deviating behaviour during mass convergence and emergency events [72], one can build applications to improve situational awareness during those events [166]. Alternatively, one can deploy methodologies that turn tweets to predictions about signals emerging in real life [91]. For
example, work is concentrated in detecting the occurrence and magnitude of disease out-
breaks in the population [30, 92, 93, 148] or natural disasters such as an earthquake [139].

From the above, one may naturally come to a conclusion about the high degree of im-
portance that Twitter content has. The interesting, one-sided nature of relationships, the
retweet mechanism which allows for a rapid information spread, the ability to conduct
online conversations and, on top of all that, the open Twitter API making this source of
information easy to crawl, offer an opportunity to study human behaviour [87].

3.2 Collecting, storing and processing Twitter data

At the early stages of this work, there was no Twitter data set to work with and try our
models. This is the main reason which forced us to create a pipeline that is able to collect
and store tweets. With hindsight, this was an essential step.6 In this short section, we
describe the data collection process with references to external libraries or tools that have
been used.

3.2.1 RSS and Atom feeds

Really Simple Syndication (RSS)7 or simply a feed is a family of web formats based on
Extensive Markup Language (XML) used for publishing and distributing web content in a
standardised manner. A more recent development, the Atom Syndication Format, is an
effort to fix the limitations of RSS by supporting, for example, XML namespaces.8 Twitter
uses Atom feeds to deliver its content in a structured format; one has the option to submit
a query on Twitter’s API and retrieve Atom feeds which contain sets of tweets matching this
query. An example of an Atom feed retrieved from Twitter is shown in Figure 3.2. The main
content – the actual tweet – of the feed lies inside the tags <entry> and </entry>. The
purpose of the encapsulated sub-tags is self explanatory, for example <published> holds
the posting date and time for a tweet.

---

6 In the recent past, Twitter data sets used in academic studies have become publicly available. However,
soon after they have been removed possibly because the open distribution of such content is against the com-
pany’s policy.

7 Or originally RDF Site Summary, where RDF stands for Resource Description Framework.

8 From Wikipedia: “XML namespaces are used for providing uniquely named elements and attributes in an
3.2.2 Collecting and storing tweets

For data collection, a basic crawler has been implemented in Java programming language; Java was preferred to other object-oriented programming frameworks mainly because it is a well-established programming interface for which many useful methods have already been made available by third-parties. Tweets are collected by querying Twitter's Search API.9 Based on the fact that information geolocation is a key concept in our studies, we are only considering geolocated tweets, i.e. tweets for which the author's location is known to Twitter.

In order not to exceed the request limit set by Twitter, something that usually causes suspensions in the data collection process, we limit our data collection to tweets geolocated

---

9 Twitter Search API, https://dev.twitter.com/docs/api/1/get/search.
within a 10Km radius of the 54 most populated urban centres in the UK.$^{10}$

Here is an example query on Twitter’s Search API:

http://search.twitter.com/search.atom?
geocode=51.5166,-0.1,10km&
lang=en&
rpp=100&
result_type=recent

With this query, we are able to retrieve the most recently published (result_type=recent) 100 tweets (rpp=100) geolocated in a 10Km radius from a location with latitude = 51.5166 and longitude = −0.1 (geocode=51.5166,-0.1,10km – this is an approximation for the centre of London) written in English language (lang=en). Twitter’s response will produce an Atom feed that follows the format shown in Figure 3.2. We collect and parse the XML content on this feed by using the Java libraries in ROME$^{11}$ and then store its contents – if they have not crawled before – in a MySQL database.$^{12}$ Therefore, for each request we make, we can at most retrieve 100 new entries (tweets) to be stored in the database.

From the above, it becomes apparent that a sampling strategy has been followed during data collection. Every $n$ minutes, we form a search query for each one of the 54 locations and retrieve the 100 latest tweets. The value of $n$ varied from 1 to 10 minutes during this project and was mostly dependent on the total number of tweets we aimed to collect on a daily basis. Obviously, the shortest the collection period ($n$), the more tweets are being collected; as the number of Twitter users reached very high figures, we were able to reduce this frequency and still collect a large amount of tweets per day. Nevertheless $n$ is kept the same for all urban centres and therefore, possible sampling biases of this type are reduced.

465,696,367 tweets (geolocated in the UK) have been collected by the crawler and then stored in the database from the 19th of June 2009 to the 31st of December 2011 (926 days).


Those tweets have been published from at most \textbf{9,706,618 unique users}; the latter number is an approximation as we did not track users that might have continued to post from the same Twitter account but with a changed username.\footnote{The research in this project did not study the behaviour of individual users, but was focused only on sets of published content.} Within those dates, a rough average number of tweets per user is approx. 50; 502,912 tweets have been collected per day on average. However, this is not a very good representation for the daily data collection figures as the number of registered Twitter users has been constantly increasing together with the published volume of content. To provide the reader with a better picture, we have plotted the daily volume of collected tweets in Figure 3.3(a) as well as its cumulative equivalent in Figure 3.3(b). From the figures, we observe that half-way through the considered period the number of collected tweets starts to increase with a higher acceleration and at a point we start collecting more than one million tweets per day. To reduce possible side-effects induced by the increasing volume of tweets and maintain an equal representation of the Twitter corpus between different dates, we tend to introduce a normalisation factor in our vector space models, which is usually equal to the number of tweets considered (see for example the methods in the Chapters 4 and 5).

\subsection*{3.2.3 Software libraries for processing textual and numerical data}

Several software libraries have been implemented for applying basic methods and algorithms driven by the fields of IR, Data Mining and Machine Learning. A very common software package, used in most projects that deal with the analysis of textual content is Apache Lucene,\footnote{Apache Lucene, \url{http://lucene.apache.org/}.} a set of libraries – initially programmed in Java – able to implement a text search engine. We use Lucene to index tweets, \textit{i.e.} to tokenise and identify unique words in the textual stream, but also for counting word frequencies in documents (a document can be a superset of tweets). Optionally, Lucene can also create indices of stemmed text by applying for example Porter’s algorithm \footnote{Based on the fact that we already use the Java connector (or driver)\footnote{MySQL Connector/J is the official JDBC driver for MySQL, \url{http://dev.mysql.com/downloads/connector/j/5.0.html}.} to query the database, it is almost straightforward to embed Lucene’s libraries in our source code and create several types of indices.} [129]. Based on the fact that we already use the Java connector (or driver)\footnote{MySQL Connector/J is the official JDBC driver for MySQL, \url{http://dev.mysql.com/downloads/connector/j/5.0.html}.} to query the database, it is almost straightforward to embed Lucene’s libraries in our source code and create several types of indices.
Another useful library, which is again a project of Apache Software Foundation, is Mahout.\textsuperscript{16} Mahout is capable of handling large scale data sets and implements the most common Machine Learning algorithms for clustering and classification or linear algebra operations such as Singular Value Decomposition. The key about Mahout is that its methods are compatible

\textsuperscript{16} Apache Mahout, \url{http://mahout.apache.org/}.  

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.3}
\caption{Volume and cumulative volume plots for tweets collected from the 19th of June, 2009 to the 31st of December, 2011.}
\end{figure}
with Apache Hadoop\textsuperscript{17} in that they are also using the map-reduce paradigm,\textsuperscript{18} and therefore, can perform operations using a cluster of computers; when a cluster is not available, Mahout can also operate using a single node. We use Mahout to retrieve vector space representations directly from a Lucene index, as there exist already implemented interfaces for this purpose. The VSM can be comprised by term TF or TF-IDF weights as well as other neat options which allow us to remove stop words or rarely used words. In some occasions, Mahout is also used for simple operations such as the computation of cosine similarities between pairs of large vectors.

Given the fact that sometimes we needed customised operations with our data which were not – at least explicitly – implemented in Lucene or Mahout, we have also created our own software libraries (in Java) which can be used to create a text index, perform stemming, remove stop words, compute vector space representations and finally perform our algorithms. Methods from MATLAB\textsuperscript{18}’s native toolboxes as well as from PMTK3\textsuperscript{19} have been also used in this work.

### 3.3 Ground truth

Ground truth is a necessary set of data for our work since many of our proposed methods are based on supervised learning. Hence, ground truth is not only needed for validation, but most importantly for learning the parameters of a model during the training process. We have used three types of ground truth:

- Influenza-like Illness (ILI) rates from the Health Protection Agency (HPA)\textsuperscript{20}, the Royal College of General Practitioners (RCGP)\textsuperscript{21} [46] and the QSurveillance scheme (QSur)\textsuperscript{22}.

- Precipitation (rainfall) measurements from amateur weather stations in the UK. As no official authority could be found to provide us – timely – with rainfall indications,
we used information collected by independent weather stations in Bristol, London, Middlesbrough, Reading and Stoke-on-Trent.

- Voting intention polls prior to the 2010 General Election in the UK from YouGov.\textsuperscript{23}

The first two sets of data (flu and rainfall rates) are used for the experimental process of Chapters 4 and 5, whereas voting intention polls are used in the preliminary work presented in Section 7.3.

\section*{3.4 Summary of the chapter}

In this chapter, we started by defining the terms blog and microblog. We moved on by describing the main operations and characteristics of Twitter, a microblogging service which serves as the main source of information for this project. Twitter has attracted the focus of the academic community in the last few years as it allows researchers to study human behaviour using massive amounts of data. This platform has found also use in many other applications such as opinion mining, event detection and education. Next, we described in detail the way we use Twitter’s Search API in order to retrieve tweets geolocated in 54 urban centres in the UK; we also provided the reader with figures reporting the average number of collected tweets per day and their cumulative distribution. Then, we referred to some of the most important software tools that we have used to process text, form vector space representations and carry out statistical operations. Finally, we gave a short description of the various versions of ground truth used throughout this project.

\textsuperscript{23}YouGov Document Archive, \url{http://labs.yougov.co.uk/publicopinion/archive/}. 