PREPARATION OF MSC DISSERTATION

UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

STUDY ON THE TWITTER HASHTAG-HASHTAG CO-OCCURRENCE NETWORK AND KNOWLEDGE DISCOVERY APPLICATON

BY

NICOLA VITALE

FRIDAY, 2 SEPTEMBER 2016

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE DEGREE OF

MSC DATA SCIENCE

PREPARATION OF MSC DISSERTATION

UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

STUDY ON THE TWITTER HASHTAG-HASHTAG CO-OCCURRENCE NETWORK AND KNOWLEDGE DISCOVERY APPLICATON

BY

NICOLA VITALE

FRIDAY, 2 SEPTEMBER 2016

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE DEGREE OF

MSC DATA SCIENCE

Abstract

The emergence of social media has impacted the way most people communicate, learn, behave or conduct research. In recent years, a large number of studies have analysed and modelled this social phenomenon. Driven by social and commercial interests, social media have become an increasingly interesting subject for researchers. Following this thread new models, and applications have emerged with respect to different application domains. In this project we report a novel method for knowledge discovery based on the use of hashtags in tweets.

This paper presents an application of our model to a topic of general interest, smoking habits.

We ask the following question: can network analysis through data mining of social media (in particular Twitter) provide novel information about a conventional topic of social interest, such as smoking?

A summary of our model reveal possible trends in line with other papers as well as a potential evidence of topics discussed by users and mainly ignored by the existing smoking literature. Within other related works our approach represents an element of originality, the majority of papers within this area is in fact oriented towards using traditional surveys. Moreover we find potential evidence of a large discussion about electronic smoke products and cannabis consumption, topics almost completely ignored by the existing smoking literature that focus for the most part on traditional combustible tobacco products.

Table of Contents

ABSTRACT 3

TABLE OF CONTENTS 4

TABLE OF FIGURES 6

TABLE OF TABLES 7

CHAPTER 1 | INTRODUCTION 8

CHAPTER 2 | BACKGROUND AND SIGNIFICANCE 10

2.1 MOTIVATION 10

2.2 RELATED WORK 13

2.2.1 Data Mining and Knowledge Discovery 13

2.2.2 Network Modelling and Analysis 14

2.3 CONTRIBUTION 15

CHAPTER 3 | DATA COLLECTION FEATURES ENGINEERING, AND NETWORK CONSTRUCTION 16

3.1 KEYWORDS EXTRACTION 16

3.2 DATA COLLECTION AND NETWORK CONSTRUCTION 18

3.2.1 Collection of tweets containing hashtags in the set 19

3.2.2 Co-occurrence network construction 20

3.3 SENTIMENT CLASSIFICATION 24

3.3.1 Bayesian Approach to Classification 25

3.3.2 Pre-processing and features extraction 26

3.3.3 Training and test 27

3.4 FEATURES ENGINEERING 28

CHAPTER 4 | SOFTWARES AND TOOLS 29

4.1 INTRODUCTION 29

4.2 TWITTER API 29

4.3 PYTHON AND JUPYTER NOTEBOOK 30

4.4 R-SYSTEM 30

4.5 MONGODB 31

4.6 GEPHI 31

CHAPTER 5 | AN EXPLANATORY ANALISIS OF THE H-H SMOKING NETWORK 32

5.1.1 Network Characteristics 33

5.1.2 Smoking Discussion discovery 33

CHAPTER 6 | CONCLUSIONS AND FUTURE WORK 38

Table of Figures

Figure 1: Social Media Landscape 2016 10

Figure 2: Conditional probability distributions for the hashtags #vape and #tobacco. 21

Figure 3: Evolution of the network at each iteration of the snowball sampling algorithm. 22

Figure 4: Network of hashtags related to the smoking discussion 24

Figure 5: Force layout and nodes dimension proportional to their degree. 32

Figure 6: Distribution of network measures. 33

Figure 7: Nodes per module ratio with respect to all the whole network. 35

Figure 8: Coloured modules. 35

Figure 9: Set o bar charts indicating number of positive and negative tweets containing hashtags on the x axis. 36

Table of Tables

Table 1: Starting set of the 10 most occurring hashtags **(**hash key removed**).** 18

Table 2: Long format hashtags dataset (first 10 rows) 20

Table 3: Wide format data set of type feature-in-document. 20

Table 4: Snowball sampling. 22

Table 5: Hashtags degree an number of tweets collected. 23

Table 6: Emoticon replaced with unique key. 26

Table 7: Naive Bayes classifier performance. 27

Table 8: Classifier 10 most informative features. 27

Table 9: Set of tables containig mean retweet count and top 10 occurring words for hashtags #vape, #quitsmoking and #cannabis. 37

Chapter 1 | INTRODUCTION

<<The web is more a social creation than a technical one.>>

Tim Barners-Lee
Inventor of the World Wide Web

In recent years, social media (e.g., Twitter, Facebook, LinkedIn, etc.) had an heavy impact on everyday life. Social media are websites and mobile websites that allow people and other type of actors including companies, non-profit organisations and governments, to create, share and exchange various types of information such as pictures/videos, opinions, ideas or career interests. These networks create an unimaginable amount of data from individuals themselves across the globe. The result is a huge, unstructured amount of raw data that can be used for academic research and provide new insights on conventional, as well as newer topics in many fields. The catch is that these data requires data mining and new models to be interpreted.

The potential benefits that can derive from analysing social media data has made different type of actors interested in accessing it. For example, from a commercial point of view, organisations want to monitor the discussion about their products or services to improve their products. The value of social media data has become evident also in politics and the public sector: political scientists, public relations managers, campaign analysts are all interested in knowing the trends and preferences on a particular discussion topic. In Science the topic has opened new research domains for social scientists, physicists, computer scientists but also for mathematicians, biologists and medical scientists. In the different fields researchers are interested in the human behaviour on these websites, in finding solutions to new computational problems (such as improved web services and search engines optimisation) or in finding new path and organisation thanks to the enormous amount of human created data.

Following the existing literature on natural language processing of microblogging web texts and semantic networks, in this project we apply data mining and network analysis to answer the question: *can network analysis through data mining of social media (in particular Twitter) provide novel information about a conventional topic of social interest, such as smoking?*

This work presents a topic modelling and knowledge discovery methodology based on the use of hashtags and their co-occurrence in messages gathered from the famous microblogging service Twitter. We modelled the *complex* relationship between hashtags with a hashtag-hashtag *network* where a connection between hashtags (edge) exists if co-occurrence of the two connected hashtags within a message has been. The network has been constructed gathering tweets related to the topic of *smoking habits*.

We asked ourselves: what new information can the model deliver from the social media discussion on a conventional topic of research? The hash tagging practice is considered as a way to summarise the content of a message, as a social defined taxonomy.

We also want to study the distribution of sentiment and lexical measures within our data set, as each hashtag appears in a corpus of short texts interesting summaries are possible. The aim has been to develop a methodology that could possibly be used along or in place of some traditional methods based on surveys.

The Thesis outline is the following:

- Chapter 2 presents the motivations that brought us to investigate our specific method, it contains a review of the literature related to the topic of study and finally it summarises the contributions of this dissertation.
- Chapter 3 contains the methodology and techniques used in the implementation phase including the collection of a starting set of hashtags using keywords extracted from web pages, the sampling methodology and the construction of the network, the sentiment classification approach and the features engineering performed on the collected tweets.
- In chapter 4 we go through the tools and software used.
- Chapter 5 contains the results obtained in our case study through statistical and network analysis techniques.
- Chapter 6 contains a conclusive overview where our work is critically re-examined and where we suggest the possibilities for future work.

Chapter 2 | BACKGROUND AND SIGNIFICANCE

2.1 Motivation

As an application of the the internet, social media services have offered an environment where the exchange of information is easy for everyone. The 'shrinking world' phenomenon, first theorized by Milgram implies that any two individual in the world have a maximum of six degrees of separation between them in the chain of their connection through friends (Milgram, 1969) and is particularly evident within this context and the speed of trends diffusion has never been so high (Kirsch, 1995).

Social media have attracted the interests of researchers from different areas and Twitter in particular have been object of study more than any other. This is because of the open nature of the type of relationships that exists between users, which is particular of this platform, as well as its asynchronous but fast-paced communication.

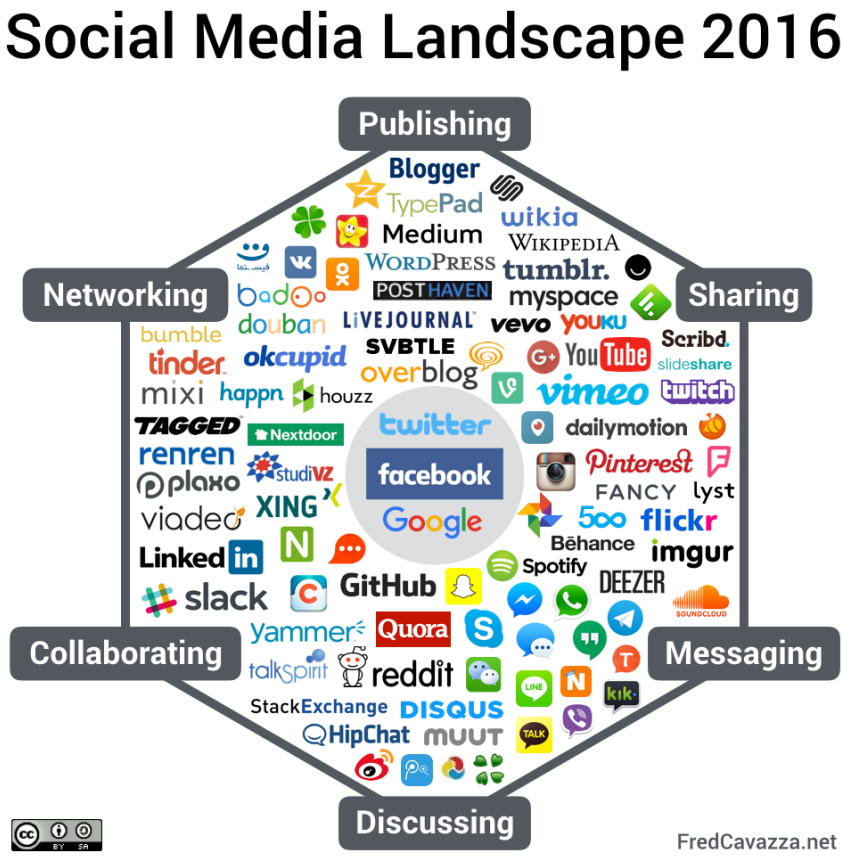


Figure 1: Social Media Landscape 2016

By definition, a tweet is a short text that cannot exceed 140 characters in length. Even if the amount of text is limited, this type of communication is usually very dense in information. Additionally, along with the texts, users can share multimedia contents, such as images, videos or links to external web pages. Further, there are other ways to overcome limitations such as the abbreviations "@" used to point at another user (user mention); "RT" to show that the tweet has been re-twitted and the user is sharing something originally posted by another user; and the hashtag "#". Among the various functions, a hashtag can be used to express an idea (e.g., #liberty); to promote an event or a group (e.g., #tobaccocontrol, #vapotage, #ukvapers); or even recommend a product (e.g., #eliquid). Moreover, people have personalised hashtags to express how they feel (e.g., #blessed, #goodvibes, #sorrynotsorry) or what they are doing (e.g., #pizzaandbeer, #gotmilk).

The use of these new forms of communication has opened new interesting possibilities of study for linguists. They are experts in the study of languages and their structure, and through systematic analyses, they can infer various aspects about individual, or event societies, identity, attitudes and values towards various objects, among many other aspects. To this group of scholars, Twitter is a "linguistic marketplace" and where they have explored the phenomena of self-branding and micro-celebrity (Page, 2012). It is shown that opinion leaders tend to reinforce their position in the online world occupying a considerable share of the participation in Twitter. Within other studies hashtags communities have been linked back to their most active users (M. D. Conover, 2010). Here researchers wanted to study how users behave with respect to a political hashtag. It is suggested that the network of users interested in a common topic grows gradually with time and more and more users join the network constructed around a single hashtag.

As the popularity of hashtags grew, they started covering more and more topics, actions, places, public figures. The network of connected hashtags could then be seen as a collective brain where different neurons point to locations, events, places or moments. For this particular research, we argue that modelling the connection between hashtags, characterizing them with specific features (such as the number of times they are shared or the main sentiment associated with them) and exploring their network could revel unexpected information about the object being tagged. However, the vast and unstructured amount of data containing the objects of our interests needs a computational effort to be modelled and used.

Developing a methodology to find and group significant hashtags require a mix of data mining techniques, network theory and creativity. In recent years hashtags related phenomena have

gathered the attention of scientist. Lehmann et al. (Lehmann, 2012) analysed hashtags to identify classes of collective attention in Twitter. These are the online discussion mirroring the offline information. They have then focused their attention on the dynamic nature of these classes linking them to real events. Their findings include the presence of peaks in the popularity of hashtags. Moreover, they tracked hashtags propagation and suggested that the fast growth popularity of a hashtag is mainly related to exogenous factor outside the online world, i.e. the real discussion is strongly connected to what happens on the web. Weng et al. conducted a study on the virality of trends within social networks using Twitter data (Weng, 2013). Their results suggest that the virality is related to the type of contagion and to the structure of the network, they used hashtags to group users interested in a common topic.

Sentiment analysis has been largely applied to studies on Twitter. Among these studies the contribution of Bollen et al. (Johan Bollena, 2010) contains a method to link the sentiment around topics of social interests to the variations within the stock market. An approach to classify tweets sentiment is given in by Agarwal et al (Agarwal, 2010). The main lesson in this paper is that a correct choice of the features plays a crucial role and special considerations are made for tweets containing emoticons.

An application that uses Twitter to examine smoking behaviour and perception of emerging tobacco products is presented by Myslin et al. (Myslín, Shu-Hong Zhu, Wendy Chapman, & Mike Conway, 2013). However, the authors of this paper do not consider any network aspect of Twitter. The study focuses on implementing a good classifier for tobacco-related tweets, and it is mainly a natural language processing contribution.

An interesting series of smoking habits related studies using Twitter data have been published on the international journal “Tobacco Control” between the years 2012 and 2016. The first study (Freeman, 2011) consider the ways tobacco related products are promoted through the internet through summary statistics on online tobacco advertise and promotion. This detailed overview of the various online channel used for this marketing activity suggests fresh approaches to regulating tobacco industry marketing are needed. Two of these papers include surveys where quantitative results show the influence of social media on behaviour of smokers belonging to different demographic groups (Sherry L Emery, 2014) (Cornelia Pechmann, 2016). The impact of online marketing campaign of electronic cigarettes products and the possibility of helping smokers to quit through additional online help have been investigated. Huang et al. (Jidong Huang, 2014) collected tweets containing one or more keywords related to the specific topic of electronic

cigarettes and found strong evidence that Twitter is heavily used to promote electronic smoke products.

2.2 Related Work

The methodology developed in this dissertation draws on the two areas of: data mining and complex network modelling and analysis.

2.2.1 Data Mining and Knowledge Discovery

Data mining is a fundamental of this project. Web pages mining for keywords extraction is the starting point of our methodology. Sometimes hashtags are created ad-hoc, in other cases they emerge from a fast process among the social media users. As a consequence, it is not possible to find a set of topic related hashtags a priori. Nevertheless, it is possible to find keywords related to a topic considering web pages containing discussion about smoking habits regulations, news and forums. This is a classic problem in data mining and both supervised and unsupervised approaches can be used.

Supervised approach requires human intervention on the data which usually consist in classification in two or more classes, while unsupervised techniques try to extract information from raw data without human intervention.

Decision trees learning is experimented on scientific documents (Turney, 1999) with better results than unsupervised techniques. Another supervised approach is based on logistic regression, Rengarajan and Galhotra (Galhotra, 2015) used it on web pages with very good results. On the other hand, an unsupervised method may suit better to the ever-changing language we find on the internet for possible live applications. We used the Rapid Automatic Keyword Extraction (RAKE) algorithm, an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents described in (Stuart Rose, 2009). Recently the RAKE algorithm has been improved introducing a dictionary of synonyms and the authors (Dutta, 2016) show that results of this variant are closer to the manually selected list of documents key phrases. Another popular unsupervised method is TextRank, a graph-based ranking model (Tarau, 2004).

Sentiment classification of documents is another very popular data mining technique used in social media studies (Lee, 2002). Also here alternatives are many among supervised and

unsupervised machine learning algorithms. The classic Naïve Bayes classifier performs very well on tweets as shown by Go et al. (Alec Go, 2005). We trained this learning algorithm on a manually labelled dataset of general tweets to classify our smoking related tweets. In our implementation we considered emoticon (e.g., :) and :( ), with extreme attention to improve the performance of our classifier as suggested by Wang (Wen Zhang, 2010).

Twitter data mining applications are documented in vast amount of papers as documented in the comprehensive literature review by Steiger et al. (Enrico Steiger, 2015). This paper shows that the large majority of Twitter research have been conducted on event detection applications and methodologies. Papers related to the study of human behaviour through Twitter data investigated the connectivity of suicidal users (Bridianne O'Dea, 2015) the detection of tensions within online communities (Pete Burnap, 2013), systems for users profiling for market analysis (Kazushi Ikeda, 2013).

In general, the data mining field contains a wide number of applications to several different fields.

2.2.2 Network Modelling and Analysis

Using the terminology present in Newman (Newman, 2003) we could characterize the entire network of Twitter hashtags as a complex cyclic information network. A complex network because it is not possible to characterize the state of such system at each point in a deterministic way, because there are too many variables into play and their relationship is not easy measurable. Cyclic refers to the absence of order between its components. As such we model it through the instruments of graphs theory, a set of techniques used to empirically characterize and analyse networks. In the same paper, Newman provides a list of measures that can be computed to characterize components, of a complex network such as degree distribution and clustering. In our case the co-occurrence H-H network is limited to the smoking related hashtags, but by applying network community detection techniques, such as collaborative filtering type algorithms (Sergey Brin, 1998), may reveal previously ignored knowledge.

Co-occurrence of words network have been applied in computational linguistics to try and compare characteristics of different languages. In one of this cases (Yuyang Gao, 2013), analysing co-occurrence networks of words of different languages, researchers quantify some characteristics and show how their relationship is different from language to language.

The co-occurrence of hashtags within tweets network, which we use to model the smoking discussion on Twitter, is the major characteristic of novelty in our methodology. As far as we know

there are no implementation of a H-H network applied to the Twitter smoking discussion in the literature.

2.3 Contribution

Our contribution in this dissertation is the suggestion of a possible novel system for tobacco control surveillance using hashtags that may be used to drive or complete traditional survey based techniques. The approach we present is based on network modelling of the relationship between hashtags and on the sentiment characterisation of these hashtags through the classification of tweet texts they compare in.

The models aim to answer the following questions: can we discover unknown behaviour and attitudes of smokers towards specific topics or products? Can we learn unreported information about their habits?

The network is then used as a rich data model for exploration through network mining techniques.

Chapter 3 | DATA COLLECTION FEATURES ENGINEERING, AND NETWORK CONSTRUCTION

The implementation of our project included the following phases:

1. Keywords extraction from smoking related web pages;
2. Collection of tweets through a snowball sampling algorithm;
3. Network construction;
4. Sentiment classification of tweets;
5. Hashtags features engineering.

3.1 Keywords Extraction

The first objective is to collect an initial set of keywords used to perform a first collection of tweets from which a set of starting hashtags will be extracted. We considered three different types of web pages according to three main different linguistic registers[1]. It would be biasing to collect keywords from the same website as the characteristics of this may affects the results and we do not want to start with a biased sample. The type of web pages mined for frozen, consultative and occasional register is reported in the following list:

- **Frozen:**
    - Governmental pages
    - Regulation of public locations (such as universities and airports)
- **Consultative**
    - Different type of news (trends, health, industry, advertising)
    - Informative web pages on health and prevention
- **Occasional**
    - Pro smoking forums where users discuss about tobacco related products
    - Quit smoking support discussions

We first obtained the text removing HTML tags using Python ***beautiful soup*** library. We then applied the RAKE algorithm to extract the keywords.

[1] From Wikipedia: https://en.wikipedia.org/wiki/Register_(sociolinguistics

RAKE take as input a set of stop words to be filtered from the text. A score is assigned to candidate keywords based on the on the co-occurrence of words between a phrase which is identified by the full stop. Keywords are ranked using the ratio

$$Score_{(w)} = \frac{Deg(w)}{Freq(w)}$$

where $\boldsymbol{Deg_{(w)}}$ represent the total frequency of the candidate key word in the entire document plus the number of occurrences of the other candidate keywords with it, and $\boldsymbol{Freq_{(w)}}$ is the frequency of the the key word in the entire document. Candidate keywords are ranked through this measures and returned. In addition, if required, n-grams can be returned as, once the unigram keywords are extracted through a second the ranking procedure can be reiterated.

This technique is very sensitive to non textual characters present in the text and to additional web repetitive features especially in occasional register web pages where for example names of users or dates of the posts are repeated.

For each set of web pages, the keywords extracted have been manually double-checked to arrive at the final list; the process is illustrated in the following:

```
# All significant unigrams and ngrams (from first manual selection)
["tobacco", "tobacco products", "marijuana", "medical marijuana", "smoke", "tobacco
advertising", "tobacco control", "cigarette companies", "tobacco companies", "tobacco
industry", "cigarette packs", "cigarette advertising", "cigarette brand", "tobacco
marketing", "tobacco sponsorship", "tobacco advertisements", "smoking", "smokeless",
"smokeless tobacco", "smoking prevalence", "shisha", "smoker", "cigarette tax",
"cancer", "cancer society", "health", "public health", "health care", "smoking ban",
"smoker lobby", "smoking habit", "smokers", "marlboro", "snus", "heroin", "hookah",
"oral tobacco", "quit smoking", "bronchial", "asthmatics", "asthmatics smoking", "cigs",
"vaping", "vaping devices", "electronic cigarettes", "cheapest cigarette", "reduce
smoking rates", "tobacco illegal", "pipe", "antismoking", "cigarette sales", "smoke
anymore", "reducing smoking", "tobacco smoke", "smoke worry", "diabetes risk", "heart
attack", "drug", "kretek", "hooka", "stay clean", "smoking crystal", "menthol", "toke",
"acid trip", "inhalation", "pipe smoking", "inhale", "sputum", "asthma", "bronchial",
"ecig", "eliquids", "vaporizer"]

# After second manual selection
["tobacco product", "medical marijuana", "tobacco advert", "tobacco control", "cigarette
companies", "tobacco companies", "tobacco industry", "cigarette pack", "cigarette
advert", "cigarette brand", "tobacco market", "smoking", "smokeless", "smoking
prevalence", "shisha", "smoker", "cigarette tax", "lung cancer", "cancer society",
"public health", "health care", "smoking ban", "smoker lobby", "smoking habit",
"marlboro", "snus", "heroin", "hookah", "oral tobacco", "quit smoking", "cig", "vaping",
"electronic cigarette", "cheapest cigarette", "reduce smoking", "tobacco illegal",
"antismoking", "cigarette sale", "smoke anymore", "reducing smoking", "smoke worry",
"diabetes risk", "heart attack", "drug", "kretek", "hooka", "stay clean", "smoking
crystal", "menthol", "toke", "acid trip", "inhalation", "inhale", "sputum", "asthma",
"bronchial", "ecig", "eliquid", "vaporizer"]

# Final selected unigrams
['drug', 'marijuana', 'nicotine', 'tobacco', 'smoking', 'smoke', 'smokeless', 'smoker',
'cig', 'ecig', 'vaping', 'vaporizer', 'shisha']
```

The final key unigram contains 13 features. Notice the distribution of these unigrams in three macro class:

- Drugs related (2 elements)

- Combustible tobacco related (7 elements)
- Electronic smoking (3 elements)
- Alternatives (1 element)

Using the final list, the Twitter streaming API has been queried and in the time interval of approximately 4 hours 26,000 tweets have bee collected. Queries have been issued for each word separately and the same number of tweets (2,000 tweets) have been obtained for each keyword. This means we collected 2,000 tweets containing the word ‘drug’ at least once, 2,000 tweets containing the word ‘marijuana’ at least once, 2,000 tweets containing ‘nicotine’ at least once and so on.

This set of tweets represented our starting point to find a starting set of hashtags to begin our sampling.

Hashtag	Count
vape	469
vaping	291
cannabis	284
ecig	239
vapeporn	230
marijuana	209
eliquid	205
sorrynotsorry	200
vaporizer	200
ecigs	199

Table 1: Starting set of the 10 most occurring hashtags (hash key removed).

In Table 1 we show the result of the count of the top 10 occurring hashtags. It can already be noticed how the discussion seems to be skewed towards electronic smoking devices, evidence which will emerge even stronger in the network. An important action that we performed has been lowering down all the hashtags. Twitter users often use the convention of pasting two or more words inside an hashtags substituting spaces with upper case letter. This practice is very common, in our starting set for example ‘#ecig’ appeared also as ‘#ECig”, ‘#Ecig’ and ‘#eCig’. This simplification seemed very reasonable since we are interesting in the meaning of hashtags with respect to one another.

3.2 Data Collection and Network Construction

With our set of hashtags, we could start our sampling methodology. We implemented a sampling strategy of type “snowball”.

This procedure starts gathering a number of tweets containing at least once each hashtag present in our starting set (as we did with the keywords).

While querying for keywords we obtained data quite fast using the streaming API this time the streaming appeared slow. A query such as '#vaporizer' seems to be slower than the single word query because Twitter returns tweet that contains exactly the same string. It in fact interprets blank spaces as OR conditions on the strings. For this reason, we decided to mix search and streaming APIs following a search with a streaming period (and a sleeping time before issuing another query because the search API has not only query length constraints but also query frequency constraints). We noticed that in the "worst" case (limits wise because for us the more tweets the better) we issue approximately 3 queries per hashtags. Considering the API limit of 180 queries per 15 minutes time window (as to say 0.2 queries per second or 1 every 5 seconds) we tried to tune streaming time and sleeping time accordingly to optimize the number of tweets retrieved per unit of time.

The network is reconstructed at each iteration of the snowball sampling and new hashtags are added to the set "branching out" with the most co-occurring hashtags with the set added at the previous iteration. In the following we illustrate the procedure more in detail; the components of each iteration are:

1. Collection of tweets containing hashtags in the set;
2. Co-occurrence network construction;
3. New most co-occurring hashtags are added to the set.
4. A new set branching out from the previous one is created.

3.2.1 Collection of tweets containing hashtags in the set

Our starting set is represented by the hashtags contained in Table 1. For each of those a number of tweets is collected using a mix of search and streaming API. This time we could not set an amount of tweets to be collected because the search API goes 1 week in the past returning a varying number of tweets matching the query and the streaming time has been limited (for certain hashtags the streaming is faster due to their popularity). For this reason the co-occurrence matrix, through which the network is constructed, has been normalised.

3.2.2 Co-occurrence network construction

Given the set of tweets containing our hashtags a co-occurrence matrix is constructed counting the co-occurrences of each hashtags with the others.

An efficient way to do this in R software is with the following procedure:

1. We create a long format dataset where all the tweets appear and a tweet _id may therefore be repeated.

	tweet_id	hashtag
1	770256597654179840	ecig
2	770255700937302016	ecig
3	770255606603284480	ecig
4	770254731981811712	ecig
5	770252466730467328	vapenews
6	770252466730467328	vaping
7	770252466730467328	ecigs
8	770252466730467328	ecig
9	770250904368975872	ecig
10	770250713335148544	vaper
	…	…

Table 2: Long format hashtags dataset (first 10 rows) .

As we can see the tweet whose _id ends in 328 contains four hashtags: #vapenews, #vaping, #ecigs, #ecig (tweet like this one are usually promotion tweets that, as we will observe in the network analysis part, are posted by companies to promote their name reaching possible costumers and represent a large share of the electronic smoke tweets; from this point of view it seems that the results found by Huang et al. (Jidong Huang, A cross-sectional examination of marketing of electronic cigarettes on Twitter, 2014) about this practice in 2012 are still true.

2. At this point the data is transformed in a wide format $\boldsymbol{hashta\_in\_tweet} \in \boldsymbol{\mathcal{M}_{hxt}}$, where $\boldsymbol{h}$ equals the length of the vector of unique hashtags and $\boldsymbol{t}$ is the number of tweets we got. The rows ca be seen as the features characterising our documents (tweets, appearing as columns).

	tweet_id			
hashtag	766837611256094720	766848734827192320	766855086840700928	
ecig	1	1	1	
ecigs	0	0	0	
vapenews	0	0	0	
vaping	0	0	0	
cloudchaser	0	0	0	…
ukvaper	0	0	0	
ukvapers	0	0	0	
vape	1	1	0	
vaper	0	0	0	
vapor	0	0	0	
		…		

Table 3: Wide format data set of type feature-in-document.

The table contains the first 10 rows of the first 3 tweets (this time hashtag are taken uniquely).

3. The co-occurrence matrix is now obtained by the matrix (partially) represented in Table 3 by its transpose.

$$cooc\_matrix \in \mathcal{M}_{hxh}$$

$$cooc\_matrix = hashta\_in\_tweet * hashta\_in\_tweet^{T}$$

Since we are interested in the relationship between hashtags the diagonal is set to zero. Given the naïve assumption that the hashtags are independent we can normalise the matrix dividing each column by the total of is summed values. This is an estimation of the conditional probability of hashtags co-occurrence. From this data we construct an undirected network considering the matrix upper part.

4. We now have a distribution of the conditional probability of our set of hashtags(Figure2).

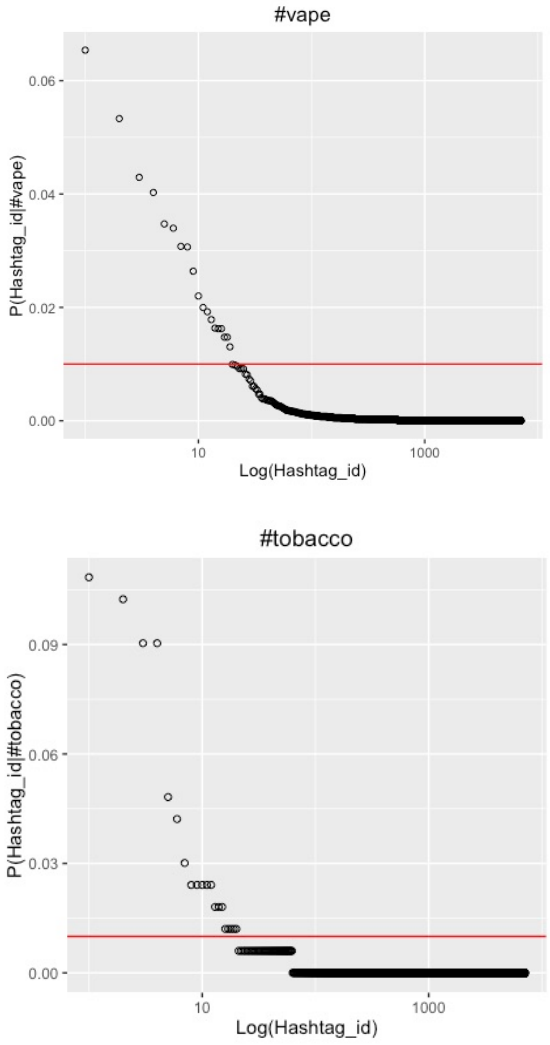


Figure 2: Conditional probability distributions for the hashtags #vape and #tobacco.

We add to our set of hashtags those ones whose conditional probability is grater than an arbitrarily chosen threshold of 0.01 probability. The hashtags added to the list are considered only once stripping out repetitions.

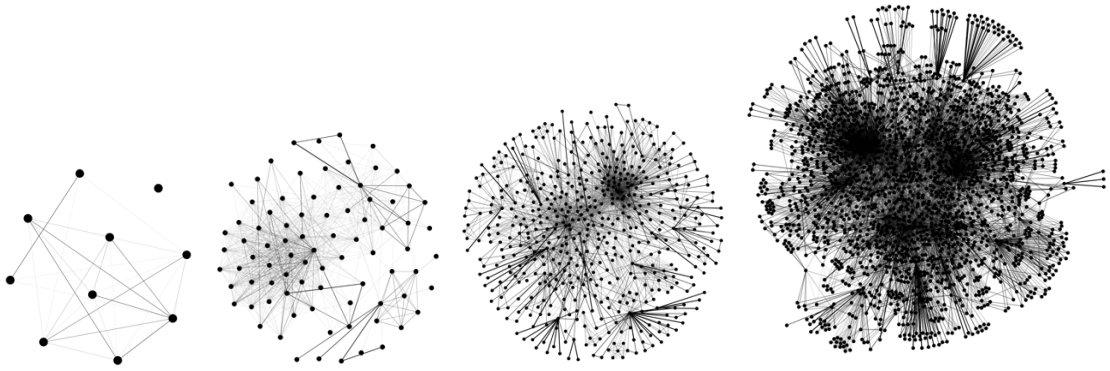


Figure 3: Evolution of the network at each iteration of the snowball sampling algorithm.

Figure 2 illustrate the evolution of our network of hashtags at each iteration of the sampling algorithm. We can see how the number of connections grow rapidly.

In the following table we include, the number of nodes (unique hashtags), edges and tweets (containing at lest one of the hashtags) at each iteration of the "branching-out" procedure.

Table 4: Snowball sampling.

	Starting set	**1st iteration**	**2nd iteration**	**3rd iteration**
Nodes	10	83	535	2,073
Edges	26	759	3,877	15,245
Tweets (containing hashtags)	3,896	16,155	104,190	(arrested)

For time constraints reasons we decided to focus our analysis on the network formed by the 535 hashtags collected up to the second iteration. In fact, as the number of new hashtags grows the query to be issued to the Twitter API to collect new tweets is longer and longer. For the above reported constraints applied by Twitter on the length of the query that can be issued and on the number of tweets that can be collected per time unit, the time required to collect tweets for our new hashtags grows exponentially. We therefore decided to stop the sampling and consider the network constructed with 535

nodes (2nd iteration). We then spent more time on collecting more tweets containing our set of 535 hashtags to reduce the bias that derives from a low number sample.

In Table 5 we report the number of tweets collected for each of our hashtags (this number correspond to tweets containing the hashtag), the values are reported for those hashtags with more than 50 connection (degree) and the data is sorted by decreasing degree.

Table 5: Hashtags degree an number of tweets collected.

Hashtag	**Degree**	**Collected tweets**
vape	163	4202
vaping	110	2075
vapelife	105	1976
cannabis	90	3143
ecig	88	1170
vapefam	87	1592
vapeporn	83	1311
weed	83	1652
marijuana	82	1765
vapelyfe	81	887
vapor	74	641
vapenation	73	783
vapeon	71	1147
ejuice	68	711
dabs	66	526
itsmyfriday	64	44
mod	63	458
vapecommunity	62	619
eliquid	59	816
vaper	59	327
vaporizer	56	485
smokefree	56	258
hemp	56	685
ecigs	55	806
notblowingsmoke	55	550
mmj	54	824
subohm	53	257
stoner	51	791
smoke	51	605
monday	51	809

As we can observe, with our sampling we tended to collect more tweets for hashtags with more connection in the network. Figures 3 represent the network we are going to analyse (circular layout).

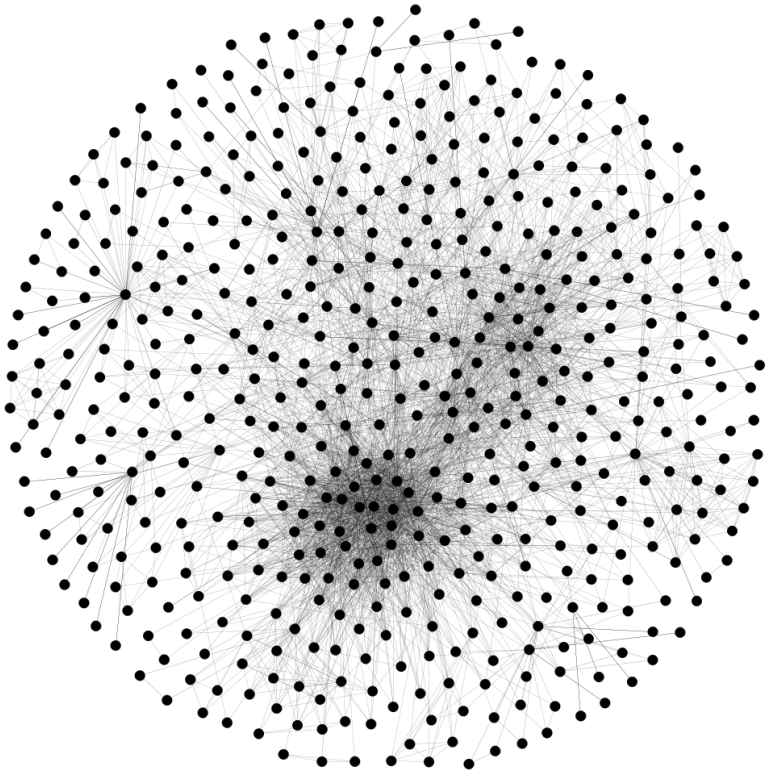


Figure 4: Network of hashtags related to the smoking discussion

3.3 Sentiment Classification

An important part of our study is the sentiment characterization of hashtags through the classification of the tweets they compare in. For this purpose we trained a Naïve Bayes classifier on a dataset[2] of 10,000 labelled tweets. To improve the performance of a tweets sentiment classifier it is fundamental to clean the texts properly; the feature selected by the classifier are strongly influenced by this phase.

In the following we are going to briefly describe the principle of the Bayesian approach to classification, we describe how we proceeded in the filtering phase and eventually we report the training and test phases presenting the most informative features found by the classifier.

[2] The dataset has been downloaded from: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

3.3.1 Bayesian Approach to Classification

The Bayesian classifier perform a classification of statistical type based on the calculation of the a-posterior probabilities i.e. given that an event has happened, we determine the probability of the causing feature. The classifier directly comes from the famous Bayes theorem.

The theorem derives from two fundamental laws of probability: conditional probability theorem, composed probability theorem.

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

the probability of event A conditioned by event B is defined as the probability that event A occurs given that B has occurred.

Composed probability:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

Bayes theorem in is simplest form:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Given a class of interest C and a set of features $F_1, \dots, F_n$ of a document we want to know with which probability this document belongs to C i.e. we want to compute $P(C|F_1, \dots, F_n)$. Applying the composed probability rule we obtain the Bayes theorem form:

$$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

Ignoring the denominator (defined as normalisation constant) and applying n times the conditional probability formula we have:

$$P(C|F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C, F_1) \dots P(F_n|C, F_1, \dots, F_{n-1})$$

The generic term $P(F_i|C, F_1, \dots, F_{i-1})$ becomes $P(F_i|C)$ for Naïve assumption of independence between the features (tweet terms, in our case). For the composed probability theorem, we finally have:

$$P(C|F_1, \dots, F_n) = P(C)P(F_1|C)(F_2|C) \dots . P(F_n|C) = P(C) \prod_{i=1}^{n} P(F_i|C)$$

In our case we classify our tweets within the positive or negative classes a document $\hat{d}$ is then assigned to the positive class C_p or to the negative class C_n on the basis of the following maximum a posteriori decision rule, $\hat{d} = C_k$ for some k as follows:

$$\hat{d} = argmax_{k\in\{p,n\}} P(C_k) \prod_{i=1}^{n} P(F_i|C)$$

This classifier is said to be naïve because it is based on the simplifying assumption that all the features that describe an instance are independent given the category of the instance. Even if this assumption is violated in the majority of real problems, such as texts classification, the Naïve Bayes classifier is effective. The assumption of independence between features also allows to learn the parameters of each class separately simplifying the learning process.

We use the accuracy and error measures to test our classifier:

$$\hat{A} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\hat{E} = 1 - \hat{A}$$

There $TP + TN$ (true positive and true negative) represent the number of documents correctly classified.

3.3.2 Pre-processing and features extraction

The pre-processing phase aims at filtering all those textual components that are not useful for the classification so that it can be more efficient.

In particular, we used detection techniques through *regular expressions* to:

- Search and elimination of: URL (http…), hashtags, RT (tag that indicates the retweet), @user (mention of a certain user in the tweet);
- Search and substitution of emoticon. In particular, those that indicates a a positive sentiment are replaced with the string “posmaremo” (positive marker emoticon), while those that indicates a negative sentiment are replaced with “negmaremo” (negative marker emoticon); the emoticon found and replaced are reported in Table 4.

Table 6: Emoticon replaced with unique key.

Key	Emoticon
posmaremo	(: :) :] [: :-) (-: (; ;) ;] [; ;-) (-; :D :p XD xd
negmaremo	:( ): :-( )-: ;( ); ;-( )-; :-[ ]-: :/ :| |: D:

- Search and elimination of letters repetitions, for example “I looooove youuu” becomes “I love you”;
- Elimination of stopwords and punctuation;
- Elimination of repeated blank spaces.

In the python notebook file relative to the tweets sentiment classification all the regular expression used are reported.

3.3.3 Training and test

For the training and test phase a subset of 10,000 tweets have been extracted from the labelled dataset making sure to include 5,000 positive labelled tweets and 5,000 negative labelled tweets. Our dataset has then been randomly shuffled and divided in a training set of 7,000 tweets and a test set of 3,000 tweets used to measure accuracy and error of our classifier.

We trained the Naïve Bayes classifier implemented in Python ***nltk*** package which before starting the actual training required us to transform the data rom csv to a convenient Python dictionary structure.

The training phase is typically computationally expensive. Even with this reasonably small training set this phase took around 15 minutes on our machine with 2.9GHz Intel Core i5 processor an 8 GB RAM.

When tested on the set of 3,000 tweets the classifier showed performed as indicated in Table 5.

Table 7: Naive Bayes classifier performance.

Training tweets	Testing tweets	Accuracy	Error
7,000	3,000	0.756	0.244

In Table 6 is include the list of the 10 most informative features spotted by the classifier. The first columns indicate the feature and the second the ratio of tweets containing that feature classified as positive or negative.

Table 8: Classifier 10 most informative features.

```
          contains(sad) = True              0 : 1      =     44.2 : 1.0
        contains(quoti) = True              1 : 0      =     40.8 : 1.0
   contains(musicmonday) = True             1 : 0      =     36.5 : 1.0
 contains(followfriday) = True              1 : 0      =     21.8 : 1.0
      contains(youquot) = True              1 : 0      =     20.6 : 1.0
      contains(missing) = True              0 : 1      =     16.8 : 1.0
      contains(anymore) = True              0 : 1      =     14.2 : 1.0
        contains(hurts) = True              0 : 1      =     14.0 : 1.0
       contains(thanks) = True              1 : 0      =     11.4 : 1.0
```

For example from the first row we see that “sad” has been classified as negative (0) the majority of the time and only a small fraction (1/44.2) of the times the feature appeared in positive labelled tweets (1).

3.4 Features Engineering

Since each of our hashtags corresponds to a set of texts (in which it compares in) a series of summaries measures of these texts can be computed.

In particular, to each of the 535 hashtags we tracked, we computed:

- **Positive and negative tweets count**

  After having processed the tweets with the classifier we counted the number of positive and negative ones for each hashtag.

- **Retweet count mean and variance**

  One of the interesting metadata that come with a tweet is the number of time it has been retweeted, i.e. shared to the date. We attached mean and variance to an hashtag computing them with respect to its corpus of tweets.

- **Lexical diversity and average number of words**

  The first measure is defined as the ratio of the number of unique words in a text (or in a corpus, of tweets in our case) and the total number of words in the text. It is a popular measure in computational linguistic.

  Average number of words is simply the mean length of a tweet in words given a set of tweets.

Chapter 4 | SOFTWARES AND TOOLS

4.1 Introduction

In the development of our project a rich variety of computational tools has been used.

In a first instance it we had to familiarise with the Tweitter API (Twitter Applications Interface) promoted by Twitter itself among developers that create applications with embedded Twitter functionalities.

A fundamental part of our pipeline is the database. We used MongoDB a fast and flexible NoSQL database whose query language is based on Javascript. The data base has been crucial to store the raw data, clean them and to store the processed data in the various steps.

Python has been used as general purpose programming language as connection between the Twitter API and our database and in the pre-processing and features engineering of the data.

R and Gephi are two powerful open source analytics tools. R functionalities are seemingly infinite especially thanks to the reach community that supports it, while Gephi is designed specifically for networks analysis and visualisation. Gephi took the network (in the form of adjacency matrix) that we constructed querying MongoDB from R.

4.2 Twitter API

As mentioned the Twitter API is designed for developers that want to get Twitter contents or post to it through their applications but it can be used also for research purpose (at the end what we do is very similar to an automatic application).

The Twitter API is divided in two main parts: the search API and the streaming API. The search API includes a variety of methods that, once created an account as Twitter developer, can be called via web queries (for test purpose there is a console available online that simply works with a normal web browser[3]).

The API returns various type of objects: an “user” object includes characteristics of a Twitter user, a “time line” object contains a set of tweets posted over a time interval by a user, a “tweet” object is a feed posted by an online user that comes with a rich set of meta data including time, users and information related to him/her, list of hashtags contained, meta data included (i.e. links to

[3] https://dev.twitter.com/rest/tools/console

photos or videos), and of course text. If the users agree with it the tweet can contain the location from which it has been published.

Twitter API object are returned in JSON format (JavaScript Object Notation), a nested data structure key-value pairs and lists.

The Twitter API methods have queries limits associated both with the complexity of the query that can be issued and with the number of data that can be retrieved over a time period.

The methods we used are “GET search/tweets” that allows to collect tweets matching a query up to one week in the past and “GET status/sample” from the streaming API that provide a random 1% of the live tweets posted over the time we are connected.

The principal guide to use this tool is the official documentation[4]. We included useful method and practical considerations in the Python notebooks contained in the archive file attached to this dissertation.

4.3 Python and Jupyter Notebook

Python is a very popular general purpose programming language. In recent years it has become very popular among scientists for a number of useful packages such as ***numpy*** and ***pandas*** for mathematics, statistics and data management.

The main packages used in this project are: ***nltk*** (set of function for natural language processing), ***beautiful soup*** (an HTML parser used to filter the hyper textual structure in the phase of keywords mining) and ***twitter*** (a Python wrapper class for the twitter API that basically turns API methods into Python methods)

4.4 R-System

R-System is an analytical tool where data are organised in vectors and list such as data frames (a table to group data into named variables).

R works with tabular data, to query MongoDB and transform the data in JSON format we use the ***mongolite*** and ***jsonlite*** package. R visualisation have been produced using the ***ggplot2*** package.

[4] https://dev.twitter.com/overview/documentation

4.5 MongoDB

MongoDB is particularly useful when dealing with JSON files because it is exactly the format in which it stores the documents.

Different database instances have been created after each manipulation of the data.

The data base has represented the bridge between Python and R, between the collection and pre-processing phase and the analysis phase.

4.6 Gephi

Gephi is designed for network analysis and visualisations. In our case we provided it the adjacency matrix computed in R and used it to visualise the data and run the clustering and Page Rank algorithms.

The software has not had problems in dealing with our dataset but we noticed that when trying to manipulate the set with 2,000+ nodes it started to suffer.

Chapter 5 | AN EXPLANATORY ANALISIS OF THE H-H SMOKING NETWORK

The following figure represent our network of 535 tracked hashtags organised with a force layout. Force layouts such as this one are very useful to visualise and explore small to medium size networks: the whole network is modelled as a system of masses (nodes) and springs (edges between the nodes, whose constant of strength is proportional to the weight of the edge). In our case the edges weight corresponds to the estimated probability of the two connected hashtags of co-occurring:

$$Weigth = \frac{(Number\ of\ tweets\ where\ the\ two\ connected\ hashtags\ occur\ together)}{(Sum\ of\ the\ total\ tweets\ in\ which\ the\ two\ hashtags\ occur)}$$

thus hashtags that are more likely to occur together are visualised closer to each other as attracted towards one another with a stronger spring.

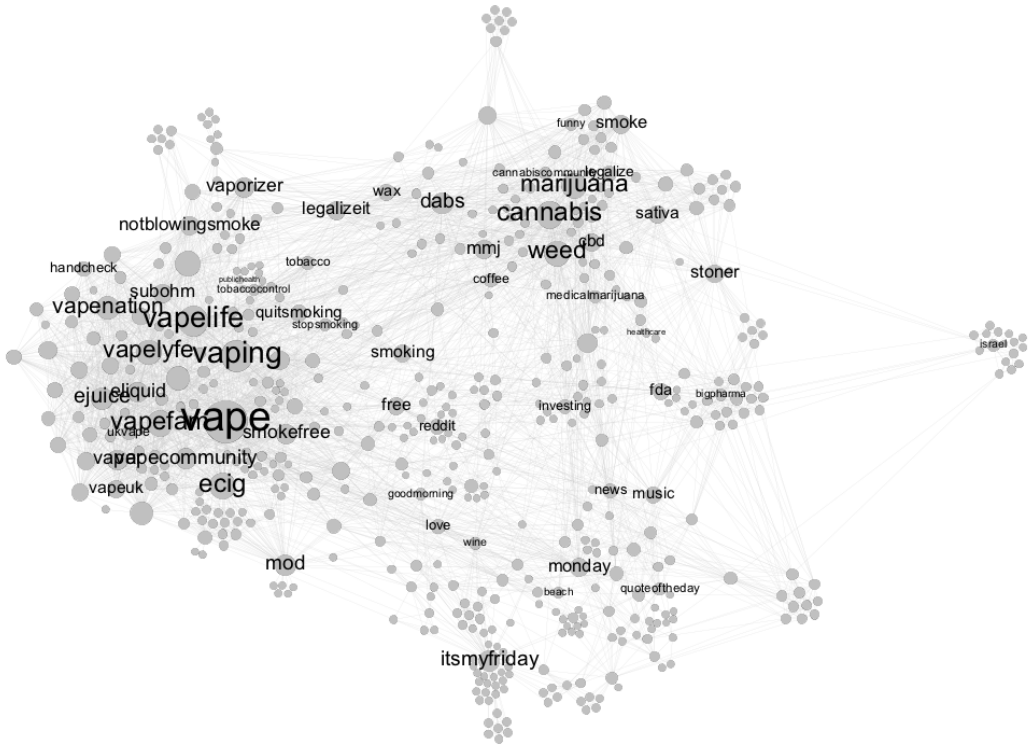


Figure 5: Force layout and nodes dimension proportional to their degree.

5.1.1 Network Characteristics

Our network presents the characteristics of a scale-free network such as the internet network of links between web pages or the network of academic citations.

The average path length of our network is 3.073. This measure represents between the average graph-distance between all pairs of nodes (connected nodes have graph distance of 1). This value is characteristic of scale-free networks.

Another characteristic of scale-free networks is a high clustering coefficient. High value of the clustering coefficient, which indicated how much nodes are embedded in their hub. The average clustering coefficient of our network is 0.686. A value of this coefficient grater than 0.5 indicated that the majority of the nodes is part of a hub.

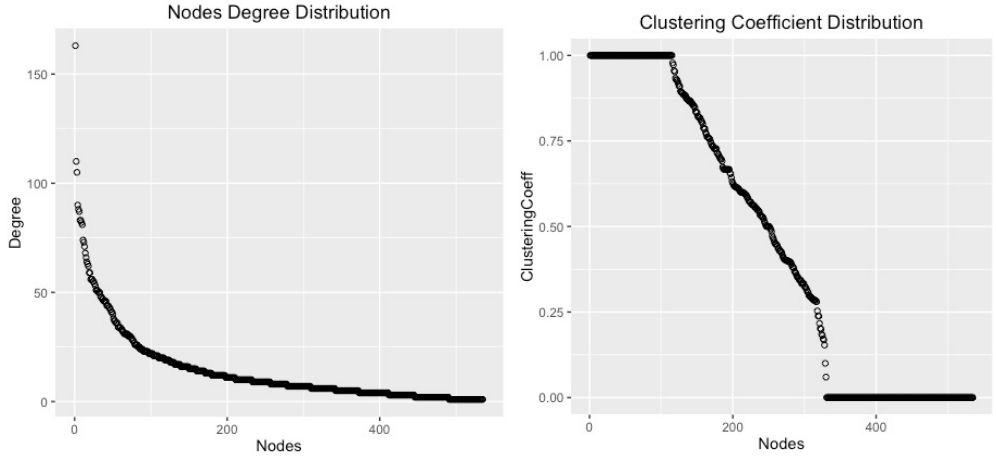
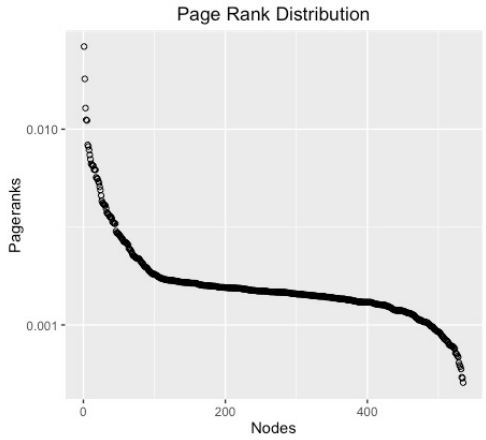


Figure 6: Distribution of network measures.



5.1.2 Smoking Discussion discovery

Clustering our nodes, we find the existence of seven major communities that together includes around two thirds of the total number of our hashtags. The communities have been labelled after an exploration of the major topic they represent and assigned the name of the node with the highest number of connection in them. They contain between the labelled communities contain

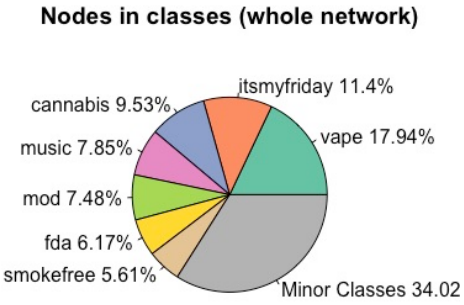
between 22 (smokefree community) and 90 (vape community) nodes. A summary of the classes is presented in the following pie chart while Figure 6 presents the network organised with the force layout where the classes have been highlighted with different colours and some of the main hashtags within them are labelled.

We observed the presence of three main clusters associated with three different smoking products: electronic cigarettes, tobacco and cannabis. Another interesting cluster is represented by healthcare politics related hashtags, in yellow.

After a first qualitative exploration we tried to characterised the clusters found through the features of the hashtags contained in them.

We found that the average lexical diversity of the hashtags contained in the vape class is usually lower (vape lexical diversity: 0.180) compared to the other classes (tobacco lexical diversity: 0.232, weed: 0.204). We think this is due to the high number of repeating tweets containing vaping related hashtags posted usually with link to promote electronic smoking products. As mentioned before this data suggest the same phenomenon mentioned by Freeman (Freeman, 2011): the online space seems to appeal to electronic smoke producers that promotes their products often associating a link to their website with popular hashtags.

Figure 7: Nodes per module ratio with respect to all the whole network.



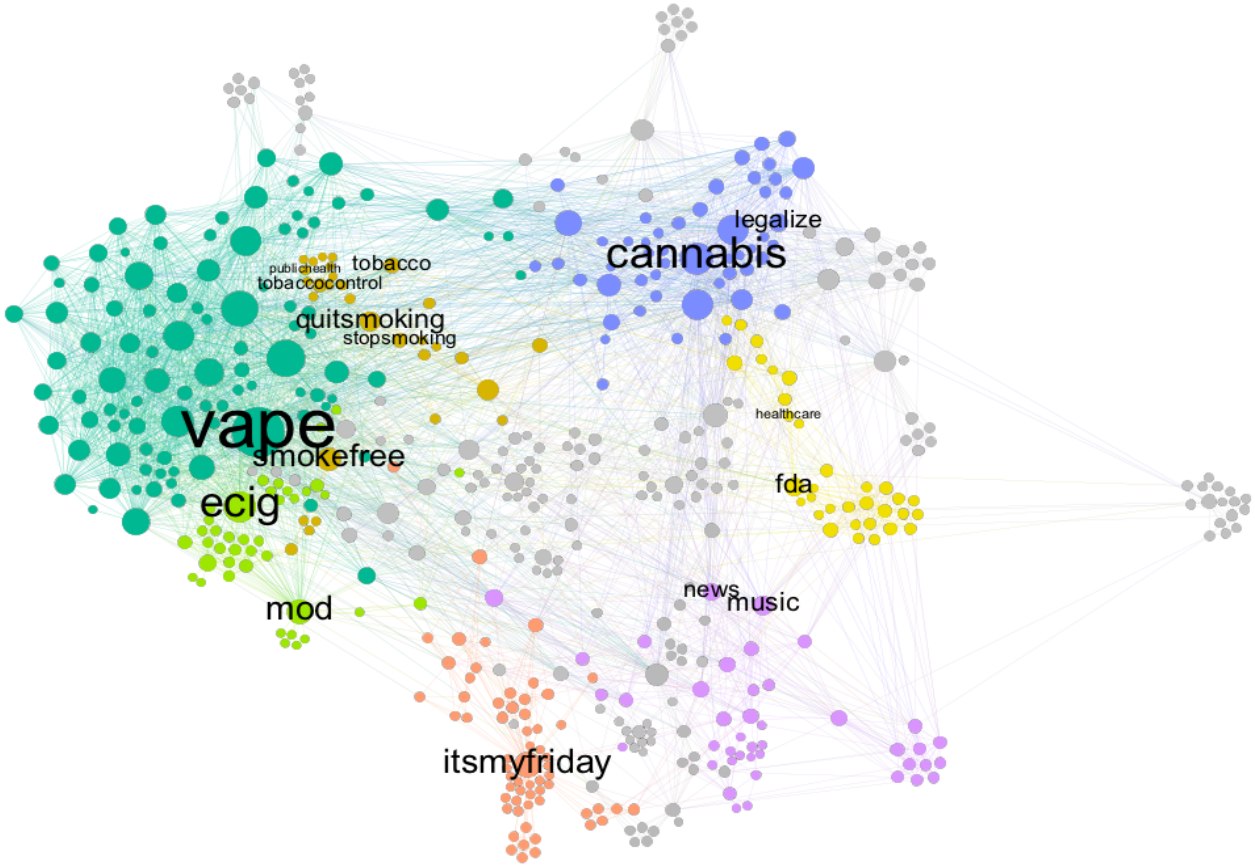


Figure 8: Coloured modules.

In Figure 9 the distribution of the sentiment for the top hashtags within six of the main classes. Hashtag importance has been determined trough the Page Rank algorithm.

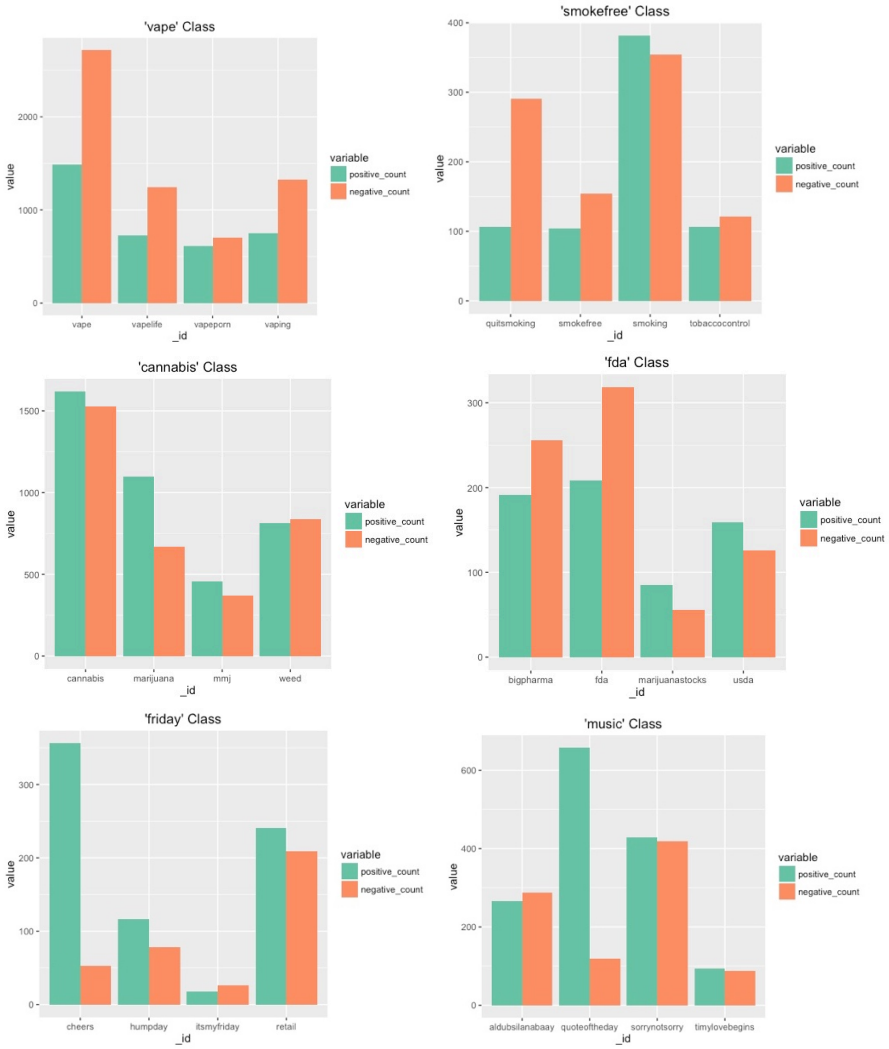


Figure 9: Set o bar charts indicating number of positive and negative tweets containing hashtags on the x axis.

We think that the sentiments related to the vape class may be more biased for the tendency of the classifier to classify non English tweets as negative. It has been observed in fact that even if we queried only for English tweets this class tended to have a lot of non English tweets where electronic smoking products from all the world are promoted through the same English hashtags. Finally, in the following set of table we report the three main hashtags we report the most connected hashtags for the three classes representing the three main types of smoking products that emerge from our network along with their mean retweet count their lexical diversity and the top ten words their occurred with.

#vaping	
Mean Retweet Count: 5.789	
Lexical Diversity of tweets corpus: 0.188	
Word	**Occurrence**
vaping	2,115
vape	1,396
vapelife	434
ecigs	379
ecig	256
vapefam	232
eliquid	229
vapor	215
coil	171
vapeporn	162

#smokefree	
Mean Retweet Count: 3.373	
Lexical Diversity of tweets corpus: 0.320	
Word	**Occurrence**
quitsmoking	356
smoking	120
vaping	109
vape	102
vapelife	84
cloudchasing	82
thetikihutvapes	82
tikilife	82
quit	37
notblowingsmoke	33

#cannabis	
Mean Retweet Count: 7.270	
Lexical Diversity: 0.174	
Word	**Occurrence**
cannabis	3,252
marijuana	1,428
weed	971
legalize	442
mmj	324
high	315
mme	297
smoke	284
legalise	280
bong	271

Table 9: Set of tables containig mean retweet count and top 10 occurring words for hashtags #vape, #quitsmoking and #cannabis.

Chapter 6 | CONCLUSIONS AND FUTURE WORK

As already stated, the expansion of the internet and proliferation of social have made communication easier across the world. Social media opened new question (e.g., big data management, sharing huge quantities of a new unstructured information) and for scientists of many, if not all, fields, social media is a mine of raw data, with new findings waiting to be discovered. As already mentioned, data collected from social media has been used in the business and marketing, politics and even the health sciences; the purposes of data from social media, such as Twitter, can provide information on innovative ways to advertise a product, ways to strategize political campaigns, reveal people's attitudes towards certain public figures or even controversial health issues. The challenge in research is how to gather such data, prepare it and analyse, in order to arrive to novel findings.

This work proposes that new methods of analyses, at the macro level, can lead to new findings on conventional research topics by assessing social media data in the field of health sciences. Namely, this research argues that through data mining and network analysis with natural language processing of smoking-related data from social media, we find newer information about people's sentiments and habits toward the conventional health topic of smoking. These methods, data mining and network analysis, are virtually inexistent in the literature of smoking, particularly works on sentiments towards smoking. The vast majority of the literature relies on data gathered from surveys, and some on focus groups. Thus, this particular work proposes an innovative method of analysis.

Through data mining of social media data, we can collect and extract new meanings from virtually live data from individuals located in different geographical points. Furthermore, it provides individual information that is, technically, unbiased by an observer, such as the interviewer during a survey questionnaire; thus, social media data can be more representative of the actual sentiments, opinions, habits and thought process of the individual.

Networks analysis, on the other hand provides an unusual level of analysis, higher than simple aggregates, and an original way of physically visualising that level. Additionally, network analysis allows us to easily derive clusters and apply ranking techniques such as Page Rank an invaluable resource to find results on big amounts of data with a great degree of detail.

To support our argument that such tools and methods of data gathering and analyses can provide novel information, we focused on a conventional topic of health sciences, smoking. We focused

our attention on the hashtags, a device used in many ways to highlight and summarise contents. We then focussed on finding a set of initial 'important' hashtags for the smoking discussion on Twitter by firs mining web pages for important smoke-related keywords used to collect a first dataset of 26,000 tweets from which we extracted our starting set of hashtags. We then applied a snowball sampling procedure to find other hashtags related to the previous ones checking for tweets where the hashtags present in our set co-occur. Based on the probability of co-occurrence we build a undirected weighted network to map the relationship of these hashtags with respect to their co-occurrence. We used natural language processing to further characterize our hashtags not forgetting that they appear within a context, the tweets, and found the number of times they are present in positive or negative tweets, the mean retweet count (as an additional measure of how much this hashtag are shared in the community) the top words occurred with them in the tweets.

Our findings did contribute to the existing literature, which is traditionally focused on tobacco smoking, moreover. We found that hashtags related to electronic smoking represent the biggest class of the hashtags we collected with a share of 17.94% of the whole set. That within this class there is a strong presence of e-cigarettes and vaporisers commercial business using the media to promote their products. We also found that the discussion on smoking on Twitter is mainly related to three type of smoke: electronic smoke, combustible tobacco and cannabis consumption. Regarding the sentiment aspect, we found that for the four most used hashtags within the cannabis class it is slightly more positive with three hashtags present more often in positive tweets than negative ones. In the module containing hashtags and terms related to FDA (Federal Drugs Administration in the US) the two most used hashtags are more likely to be used within negative tweets.

Thus, we propose that in future research, particularly those related to sentiment analyses based on linguistic approaches in any field, opt for the use of social media data collection, data mining and network analyses methods to further knowledge to the bigger picture of whichever field it is being applied. Other uses of data mining and network analysis include: political discussions, the possibility of considering the time dimension to investigate the existence of trends and their implementation for live applications for knowledge discovery.

REFERENCES

Agarwal, A. (2010). *Sentiment Analysis of Twitter Data.* Columbia University New York, NY 10027 USA: Department of Computer Science.

Alec Go, R. B. (2005). *Twitter Sentiment Classification using Distant Supervision.* Stanford, CA 94305: Stanford University.

Bridianne O'Dea, S. W. (2015). *Detecting suicidality on Twitter.* Randwick,NSW2031,Australia: BlackDogInstitute,The University of New South Wales.

Cornelia Pechmann, K. D. (2016). *Randomised controlled trial evaluation of Tweet2Quit: a social network quit-smoking intervention.* Stanford, California, USA: Stanford Prevention Research Center, Stanford University.

Dutta, A. (2016). *A Novel Extension for Automatic Keyword Extraction.* Mesra, Jharkhand, India: International Journal of Advanced Research in Computer Science and Software Engineering.

Enrico Steiger, J. P. (2015). *An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data.* Heidelberg University : GIScience Research Group, Institute of Geography.

Eysenbach, G. (2009). University Health Network, Toronto, Canada: Centre for Global eHealth Innovation.

Freeman, B. (2011). *New media and tobacco control.* Sydney, Australia: Sydney Medical School, The University of Sydney, 226A, Edward Ford Building A27, NSW 2006, Australia.

Galhotra, G. R. (2015). *Final Report: Keyword extraction from text.*

Jidong Huang, R. K. (2014). *A cross-sectional examination of marketing of electronic cigarettes on Twitter .* Chicago, Illinois, USA: Institute for Health Research and Policy, University of Illinois at Chicago.

Johan Bollena, H. M. (2010). *Twitter mood predicts the stock market.* School of Informatics and Computing, Indiana University: Journal of Computational Science.

Kazushi Ikeda, G. H. (2013). *Twitter user profiling based on text and community mining for market analysis.* Ohara, Fujimino, Saitama 356-8502, Japan: KDDI R&D Laboratories, Inc.

Kirsch, S. (1995). *The Incredible Shrinking World? Technology and the Production of Space .* University of Colorado, Boulder, CO 8030-0260, USA: University of Colorado.

Lee, B. P. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Ithaca, NY 14853 USA: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Lehmann, J. e. (2012). *Dynamical Classes of Collective Attention in Twitter.* Barcelona, Spain: Universitat Pompeu Fabra Barcelona.

M. D. Conover, J. R. (2010). *Political Polarization on Twitter.* Indiana University, Bloomington, IN, USA: Center for Complex Networks and Systems Research.

Milgram, J. T. (1969). *An Experimental Study of the Small World Problem.* New York: Published by: American Sociological Association.

Myslín, M., Shu-Hong Zhu, P., Wendy Chapman, P., & Mike Conway, P. (2013). *Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products.* San Diego, La Jolla, CA, United States: University of California.

Newman, M. E. (2003). *The Structure and Function of Complex Networks.* Society for Industrial and Applied Mathematics.

Page, R. (2012). *The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags.* University of Leicester, UK: Discourse & Communication.

Pete Burnap, O. F. (2013). *Detecting tension in online communities with computational Twitter analysis.* Cardiff, UK: Cardiff School of Computer Science & Informatics.

Sergey Brin, L. P. (1998). *The anatomy of a large-scale hypertextual Web search engine.* Stanford. CA 94305, USA : Computer Networks and ISDN Systems .

Sherry L Emery, L. V. (2014). *Wanna know about vaping? Patterns of message exposure, seeking and sharing information about e-cigarettes across media platforms.* Chicago, Illinois, USA: Institute for Health Research and Policy, University of Illinois at Chicago.

Stuart Rose, D. E. (2009). *Automatic keyword extraction from individual documents.*

Tarau, R. M. (2004). *TextRank: Bringing Order into Text.* Department of Computer Science University of North Texas.

Turney, P. D. (1999). *Learning Algorithms for Keyphrase Extraction.* Ottawa, Ontario, Canada, K1A 0R6: Institute for Information Technology, National Research Council of Canada.

Wen Zhang, T. Y. (2010). *Text clustering using frequent itemsets.* Beijing 100190, PR China: Lab for Internet Software Technologies, Institute of Software, Chinese Academy of Sciences.

Weng, L. e. (2013). *Virality Prediction and Community Structure in Social Networks.* Bloomington, IN 47408, USA: School of Informatics and Computing, Indiana University.

Yuyang Gao, W. L. (2013). *Comparison of directed and weighted co-occurrence networks of six languages.* handong 250100, China: School of Computer Science and Technology, Shandong University.