Using Natural Language Processing and User Features to Detect and Rank Twitter Accounts.

Darren Simpson

40284701

Authorship Declaration

I, Darren Simpson, confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed;

Where I have quoted from the work of others the source is always given. With the exception of such quotations this dissertation is entirely my own work;

I have acknowledged all main sources of help;

If my research follows on from previous work or is part of a larger collaborative research project I have made clear exactly what was done by others and what I have contributed myself;

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have obtained **informed consent** from all people I have involved in the work in this dissertation following the School's ethical guidelines

Signed:

Date: 10/04/20

Matriculation no: 40284701

General Data Protection Regulation Declaration

Under the General Data Protection Regulation (GDPR) (EU) 2016/679, the University cannot disclose your grade to an unauthorised person. However, other students benefit from studying dissertations that have their grades attached.

Please sign your name below *one* of the options below to state your preference.

~~The University may make this dissertation, with indicative grade, available to others.~~

The University may make this dissertation available to others, but the grade may not be disclosed.

~~The University may not make this dissertation available to others.~~

Contents

Abstract 5

1. Introduction 6

2.0 Research 6
2.1 Introduction
2.2 Defining Fake News
2.3 Opinion on Fake News
2.4 Purpose of Fake News on Social Media
2.5 Identifying Fake News and Malicious Users
2.6 Research Questions

3.0 Problem Identification 19
3.1 Aims
3.1 Twitter Crawler
3.2. Database
3.3 Natural Language Processing
3.4 User Data Algorithm
3.5 Scoring

4.0 Design 21
4.1 Methodology
4.2 Twitter Crawler
4.3 Database
4.4 Natural Language Processing

5.0 Implementation 25
5.1 Design to Implementation
5.2 Architecture
5.3 Problems and Difficulties

6.0 Results/Evaluation 31
6.1 Results
6.2 Evaluation

7.0 Appraisal 34
7.1 Critical Evaluation
7.2 Future Work
7.3 Knowledge/Skills Gained
7.4 Acknowledgements

8.0 Appendix 40
9.0 References 48

Abstract

Using social media has become a part of daily life for millions of users across several sites. One such site, Twitter, is a microblogging website which is used by individuals and companies in order to share information, news and opinion. With social media becoming so widely used, many now supplement the news or rely entirely on websites such as Twitter to receive up to date information on news stories. However, due to the ability to reach a large audience at little or no cost in real time on Twitter, it has become increasingly easy for false or misleading news to be published without the fact checking one would expect from a professional news outlet. What will now be referred to as *Fake News*, can have damaging effects both to individuals and on a societal level, in many cases resulting in echo chambers forming which can further encourage dangerous ideas and opinions and cement them in the mind of the user as fact. It is due to this that it is more important than ever to attempt to find a means of automating the detection of fake news in order to attempt to give users the tools required to make informed decisions regarding what is true and false. As a contribution to the research, this literature review hopes to look at the studies that have been carried out in recent years in order to help detect fake news and misinformation and to determine which methods appear to be valid.

1.0 Introduction

As the internet grows and matures over time, so too has the way that people view information and consume news. There are up to 152 million active users per day on Twitter with 500 million posts being made each day (Omnicoreagency.com, 2020). While this figure covers the entirety of Twitter, it's reported that more than 2/3's of Americans use social media as a primary news source. (Atodiresei, Tănăselea and Iftene, 2018). With users either supplementing or, in some cases, using social media entirely for their source of news it raises the issue of whether the information they are accessing is credible or not. Castillo, Mendoza and Poblete (2010) describe social media credibility as; "…the aspect of information credibility that can be assessed using only information available in a social media platform." If users rely only on what they see within the environment of the site this can lead to false or misleading information being accepted as fact. As a result, the unregulated nature of social media has left many users who are either less aware of this issue or simply choose not to do further research, to be tricked and misled by articles that, whether purposefully or not, are misrepresenting the truth. Twitter is highly influential however despite this there is nothing to determine whether the user posting information (or the information itself) is legitimate or not other than the 'verified' status.

Various approaches have been taken to achieve the purpose of automated hoax detection since even the early days of the internet. Spam and hoax emails have always been an issue and there have been several different attempts to combat them. In more recent years, Social Media Networks have become vastly more popular, and as such, so too has the rise of misinformation spread throughout them. It is unsurprising therefore that there have been attempts in order to develop techniques in order to combat hoaxes and the spread of misinformation.

2.0 Research

2.1. Introduction

The purpose of this chapter is to provide an appraisal of several of the studies that aim to tackle the issue of Fake News on social media across several platforms including but not limited toTwitter. While the main sources of literature used in this research relate to Twitter, as it is the platform that I have chosen to develop my methodology for, it is equally important to assess what other pre-existing research has been made into alternative platforms such as Facebook. The research that I have undertaken looks not only at methodologies for detecting Fake News But also examines research related to both the credibility of social media posts and the user posting them. This chapter will be broken down into four sections which will expand upon the problem of Fake News, the purpose and uses of Fake News, how to identify both Fake News and the users spreading it and finally my overall findings and conclusions regarding the subject based upon the current literature surrounding the topic.

2.2. Defining Fake News

Using the term: "Fake News", is perhaps too broad to truly encompass the several different definitions that can be attributed to it. A study carried out by *European Commission, Brussels (2018): Flash Eurobarometer 464 (Fake News and Disinformation Online),* gives a concise description:

> *"the spread of news that intentionally misleads readers… affecting people's understanding of reality."*

I believe that this is a clear description of the issue. Several studies all agree with the same fact, that one of the main goals of fake news is its intention to mislead users. Alrubaian, Al-Qurishi, Al-Rakhami, Hassan & Alamri (2016) I feel describe it best in their study: *Reputation-based credibility analysis of Twitter Social Network Users:*

> *"Despite the immense potential to use [online social networks] for positive purposes, there is also immense potential for the misuse of this technology to perform various malicious activities, such as spreading rumors or false messages on an inflammatory topic, creating accounts for false identities, and other; users engage in these activities to achieve high influence, instigate chaos, or even sabotage national security"*

As such, I will define fake news for the purpose of this dissertation as: "the misuse of technology to maliciously and intentionally spread fake news and information to mislead users".

2.3. Opinion on Fake News

While looking into the issue of Fake News, a main source regarding opinion is the aforementioned Commission, Brussels (2018): Flash Eurobarometer 464 (Fake News and Disinformation Online). This study was found to be of interest due to several factors which make it incredibly viable when exploring the subject. This study was commissioned by the European Union and as such the information is trustworthy and carried out in a professional manner. The study is also very relevant in terms of recency, having been published in February 2018. Although, while recent, this may not reflect the current opinion of Fake News and disinformation as the topic has become increasingly prevalent since the study was published.

The paper is "designed to explore EU citizens' awareness of and attitudes towards the existence of Fake News and disinformation online" (Commission, Brussels 2018) and continues using the information examined in a Special Eurobarometer survey published in November 2016, a survey which examined public trust in, and attitude towards, the independence of the media. The goal of this survey was to gauge the opinions of 26,576 respondents from different social and demographic groups across all member states of the EU. In regard to this group, there is a socio-demographic analysis, allowing for a breakdown based on age, gender, employment and daily use of social media. Respondents were asked to give their opinion based on several subjects.

For example:

- Levels of trust in news and information accessed through different channels (e.g. Traditional media: newspaper, radio, television & Non-Traditional media: blogs, social media, etc.) – For the purpose of this research I will be focusing on opinions based on non-traditional media, especially social media.

- Perceptions of how often respondents encounter news or information which is misleading or false.

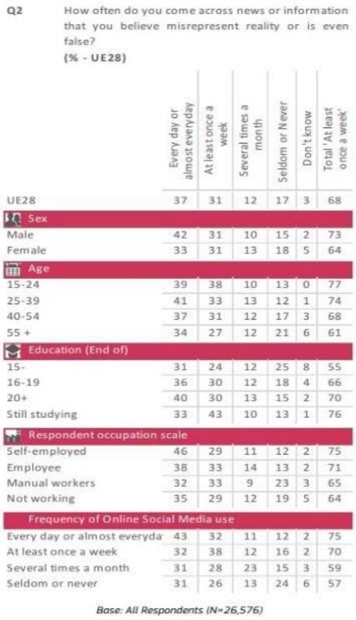
- Confidence in identifying news or information that is misleading or false
- Views on the extent of the problem, in their own country and in terms of democracy in general – in terms of this research there will be a focus on statistics from the United Kingdom.
- Views on who is responsible and how to stop Fake News.

This set of subjects allows to gain a good understanding of not only how a large set of respondents think about certain aspects of false information and Fake News, but also allows to see how factors such as daily social media use can affect how we view news in general. When examining the following statistics, it should be noted that although the study attempts to provide a wide scope the selected group only includes member states of the EU. This Eurocentric data pool potentially excludes data from a large population who, perhaps, have been affected as much, if not more by the prevalence of Fake News; the United States of America. It must also be noted that while this dissertation focuses on the UK, the entire user set will be analysed to gather as full a scope of opinion as possible to better understand how Fake News is influencing users.

Based on the findings of the study, the initial point to take into consideration is that traditional media sources tend to be trusted more than online sources across all of the studied countries, with only 26% of people saying they trust online social networks and messaging apps. Of this 26%, only 2% (1% in the UK) state that they ‘totally trust’ this group and 16%-38% stating they ‘tend to trust’. This demonstrates that even amongst those who do trust social networks that few fully trust what they see.

Most respondents in the study say that they encounter Fake News at least once a week, with more than a third of respondents (37%) report that they come across Fake News either every day or almost every day with a further 31% reporting that this happens at least once a week (Table 1). This figure is consistent across all countries studied, however seven in ten respondents (71%) are totally, or somewhat confident that they are able to identify news or information that misrepresents reality or is false. However, those who regularly use online social networks and those who come across Fake News more frequently, are more confident in their ability to identify it (75% compared to 57% of those who seldom or never use online social networks). In every country studied, more than half of the respondents claim that they are at least somewhat confident that they can identify Fake News.

The United Kingdom, however, appears as one of the most assured with 79% of respondents claiming they are somewhat confident.



Q2 How often do you come across news or information that you believe misrepresent reality or is even false?
(% - UE28)

	Every day or almost everyday	At least once a week	Several times a month	Seldom or Never	Don't know	Total 'At least once a week'
UE28	37	31	12	17	3	68
Sex						
Male	42	31	10	15	2	73
Female	33	31	13	18	5	64
Age						
15-24	39	38	10	13	0	77
25-39	41	33	13	12	1	74
40-54	37	31	12	17	3	68
55 +	34	27	12	21	6	61
Education (End of)						
15-	31	24	12	25	8	55
16-19	36	30	12	18	4	66
20+	40	30	13	15	2	70
Still studying	33	43	10	13	1	76
Respondent occupation scale						
Self-employed	46	29	11	12	2	75
Employee	38	33	14	13	2	71
Manual workers	32	33	9	23	3	65
Not working	35	29	12	19	5	64
Frequency of Online Social Media use						
Every day or almost everyda	43	32	11	12	2	75
At least once a week	32	38	12	16	2	70
Several times a month	31	28	23	15	3	59
Seldom or never	31	26	13	24	6	57

Base: All Respondents (N=26,576)

Table 1.
Frequency users claim to encounter Fake News - 2018

Note: reprinted from

As Fake News becomes more prevalent across all countries and platforms, the opinions regarding the severity of the issue also seem to increase. As of this study, a majority of the respondents think that the very existence of Fake News is a problem in their country, at least to some extent (85% of all respondents). Of this majority, 44% believe it to 'definitely' be a problem and 41% see it as a problem 'to some extent'. The more that users had encountered Fake News, the more likely they are to see it as an issue. Among those who have come across Fake News every day or almost every day, 90% of them see it as a problem in their country and 88% see it as an issue for democracy in general. The proportions are lower among users who claim to seldom or never encounter Fake News (71% and 72% respectively).

With more users highlightingFake News as an issue, the question must be asked as to who is responsible for combating the rise of false information. Online social networks are most frequently viewed by users as being most responsible for stopping Fake News in Ireland and the UK (both 38%), with 26% agreeing overall. However, 45% of all respondents view journalists as having the main responsibility. As stated previously, these figures may not be entirely representative of current day opinions due to more information regarding the utilisation of social media to spread information being available in mainstream media. While a larger group believe that it is the responsibility of journalists to combat Fake News, younger respondents are more likely to believe that online social networks should act to stop the spread of Fake News, for example; 39% of 15-24 year olds agree with this, falling to only 18% among those aged 55% or over (source). The same demographic is more likely to say that it is the citizens themselves that should be responsible; 39% among 15-24 year olds compared with 29% of those aged 55+. Regardless of age, those who use online social services are more likely to say that the networks they are using should be responsible for stopping the spread of Fake News. Among those who use these networks almost every day or daily, 31% agree that the networks should be responsible compared to the 18% who use them seldom or never.

Regarding this survey, there can be several conclusions:

- The survey indicates a degree of mistrust in the media. Only a small minority of respondents 'totally trust' any media sources, although most people at least 'tend to trust' traditional media sources such as radio, television and printed newspapers and magazines.

- When asked directly about Fake News, two-thirds of respondents say they encounter it at least once a week, and most respondents see it as a problem – both in their own country and for all democracy in general. Although people have a reasonable level of confidence in their ability to identify this Fake News, only 15% feel 'very confident' in doing so.

  - This confirms that the existence of Fake News is acknowledged as a genuine, serious issue by the public. The level of concern is widespread across different countries and across different socio-demographic groups.

- There is no clear consensus regarding who should act to stop the spread of Fake News. Many respondents believe that the press have an important role, both journalists and press and broadcasting management. Respondents also think that national authorities, online social networks and the citizens themselves have a responsibility. This suggests that, at least in the view of the public, co-ordinated efforts are required from a range of different institutions and media actors.

The information that has been gained from this literature was incredibly useful for several of the reasons mentioned previously in this section: It is a relevant poll taken from a large number of respondents across several countries which breaks down opinion across various socio-demographics allowing the researcher to further deconstruct opinions influenced by several factors (age, education, social media use). Furthermore, the literature is only two years old, meaning that the information is up to date compared to previous studies. Considering that the literature was commissioned by the European Union, it is understood that the information presented is truthful and relevant for its purpose.

The main point worth noting from this study is that misinformation and Fake News is clearly beginning to be viewed as a serious problem, at least across Europe. This is the main weakness of this study, as although it covers a large group of respondents it was commissioned with Europe in mind therefore it does not help to give a full scope of the issue worldwide

2.4. Purpose of Fake News on Social Media

Using the definitions from the literature above, it becomes clear as to why people or groups would choose to utilise Fake News. There can be many reasons, however, as the definitions show, it is predominantly in order to mislead or deceive users for gain. This can range from personal, financial or political gain. However, for users, it can potentially be a case that they have been misled and have unwittingly helped spread this misinformation. This section will look at studies that research the use of Fake News and will contain sub-sections regarding studies which broadly look at Fake News and then a more focused examination of a study regarding the use of social media and Fake News in the 2016 American Presidential Election.

2.5. Identifying Fake News and Malicious Users

There have been several studies carried out into identifying both Fake News and the users who spread it. These studies attempt to take alternative approaches in discovering what makes a fake story or post and what attributes you can discern to identify a potential bot or malicious user. This section will be broken into two subsections and will look at: identifying a Fake News post and identifying malicious users.

2.5. a) Main Sources of Literature (Malicious Users)

An alternative method of detecting false information is to consider the aspects of the user making the post, as opposed to the content of the post itself. There are several studies which cover this aspect and have various methods which they use in order to detect when a user may be furthering the spread of misinformation. In this section I will firstly evaluate my main sources when considering detection via user details. This information will then be used to outline the potential steps that could be taken and the differences between the methods of each paper. I will draw conclusions throughout in regard to the most important features and discuss the choices that will be made when shaping this dissertation.

Some Like it Hoax: Automated Fake News Detection in Social Networks (2017)

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro

This was published as a technical paper by the School of Engineering at the University of California in 2017; having been conducted by an engineering school, the content of the paper meets a high standard and has been peer reviewed prior to publishing making it a credible source. As it was published in 2017, the information is slightly less recent in relation to others that will be referenced. However, the information which is important relates to the methodologies and classifications which are not overly affected by time, While the topic of the paper is to automate Fake News detection, it focuses entirely on Facebook and how to detect malicious users via their likes on certain pages. While this may not seem entirely beneficial while attempting automated Fake News detection on Twitter, I believe that elements of the study can be used regardless of the social network that is chosen. This particular study aims to examine several methods of automating the detection of Fake News, one such method is by assessing the users who interact with posts. The authors pose the question:

> ...hoaxes can be identified with great accuracy on the basis of the users that interact with them…focusing on Facebook, we answer the following research question: Can a hoax be identified based on the users who “liked” it?

Initially this confirms that ‘hoaxes’, referencing Fake News and misinformation, can reliably be identified based on the users who are involved with certain posts. As stated before, the focus of Some Like It Hoax is Facebook. However, the findings that are discovered can be related to Twitter. Both Twitter and Facebook share a ‘like’ feature and as such, the methods related to how this information is used are incredibly relevant to this dissertation. A noted difference between these networks is that while they share similarities, Twitter also has a ‘retweet’ function which allows users to post a Tweet on their own timeline, either with or without a comment. It would stand to reason that the ‘retweet’ function holds more weight than a ‘like’ due to choosing to display it through your own profile rather than simply leaving a like.

Identifying Fake News from Twitter Sharing Data: A Large-Scale Study (2019)

Rakshit Agrawa, Luca de Alfaro, Gabriele Ballarin, Stefano Moret, Massimo Di Pierro, Eugenio Tacchini, Marco L. Della Vedova

Published in 2019, this study was written within a year of this dissertation, and as such, the information within is as recent as one could hope for. It builds on the previous work done by de Alfaro, Polychronopoulos and Shavlovsky (2015), but applies the reputation of the sites and users posting news as an identifier toward how trustworthy the information is. Unlike de Alfaro, et al (2015) this study also aims for the inclusion of Twitter data, rather than only using Facebook.

Additionally, this study claims that it is ,"the first baseline study of how a Fake News detection method fares when applied to the full breath of news being shared on Twitter in a period spanning several months" (Agrawa, de Alfaro, et al, 2019). The information provided is one of the most up to date implementations of these methods. The study proposes and characterises adaptations made to the algorithm of the previous work in order to make the large scale of information usable in real time. It utilises two algorithms which analyse user data and a reputation-based method of the websites which have published news, stating that:

> ...it can be simpler at least in first approximation to rely on social signals: on the reputation of the sites who published the news, and on the identities of the people who spread them on social media.

In the context of this dissertation, the focus will be on user related methods rather than the reputation of news sites to determine how likely a user is to use misleading or malicious means to spread Fake News.

Reputation-based credibility analysis of Twitter social network users (2016)

Majed Alrubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Mohammad Mehedi Hassan*,† and Atif Alamri

Carried out in 2016, this study was written in accordance with the Department of Information Systems at the College of Computer and Information Sciences, King Saud University, Saudi Arabia. As such, it has been peer reviewed and the information within it is relevant and of a quality standard. This study, as the name would suggest, uses few approaches relating to user sentimentality compared to others, however, what sentimentality features it does employ will be discussed in section 2.5.b). The study asks;

"can we predict the credibility of Twitter users given their account and activity information?" (Alrubaian et al, 2016).

As such, this study focuses on a user's social popularity score, which can be assessed quantitatively using an algorithm that defines this based on relevant social network features. This consists of information such as the number of tweets by a user or number of favourites of the user. Due to the aims of the study, gauging a user's credibility in relation to a given topic, there are several features which will not be relevant toward this dissertation and as such, will be disregarded; the activity of the user in relation to a specific post for example. While the study differs in its aims to this dissertation, the literature is relevant due to several claims regarding the value of user features: "…the most critical indicators are qualitative such as…the tweeter of the tweet" (Alrubian et. al 2016). In comparison to other papers discussed in this section, there is less information regarding the implementation of the methods being carried out, but instead there is a focus on determining what separates a credible and non-credible user.

Automatically Identifying Fake News in Popular Twitter Threads (2018)

Cody Buntain, Jennifer Golbeck

This recent study written in association with the University of Maryland focuses on the detection of Fake News within popular threads on Twitter. While the aim on this dissertation is to detect users based on their details and the language used by them, I believe that the research carried out by Buntain and Golbeck in reference to threads can be applied generally to all users. This study claims to correctly classify ⅔ of Fake News stories using their accuracy prediction model, outperforming any prior work in this area, making it an incredibly useful source in comparison to others. The central research question of this study is: "whether we can automatically classify popular Twitter stories as either accurate or inaccurate (i.e., true or Fake News)" (Buntain and Goldbeck 2018).

This is achieved by using 14 of the most important features defined by Castillo, Mendoza & Poblete, (2013). These are broken down into 4 different sections: Structural Features, Content Features, User Features and Temporal Features. For the purpose of this dissertation, I will be examining elements from the structural features (Twitter specific properties, such as proportions of retweets or media sharing) and user features (properties such as friend/follower count and verified status). Content features relating to the textual aspects will be discussed in section 2.5.b) This study achieves its goals by predicting the accuracy of stories based on training carried out on two credibility focused datasets:

- PHEME journalistic-labelled dataset, which is a curated dataset of threads relating to rumours on Twitter
- CREDBANK crowdsource dataset, a large-scale set of conversations about events on Twitter.

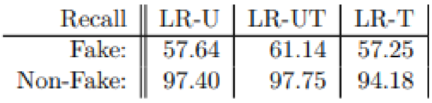
These datasets have been compared to Twitter content from Buzzfeed's Fake News dataset. An analysis is then carried out to determine the features which are most useful for predicting Fake News.

2.5 b) Identifying Malicious Users via User Features

From the related literature, there are several conclusions which can be drawn when considering how to identify malicious users based on their user features alone. Within this section I will discuss my findings with reference to literature and discern what the most important features are in relation to Twitter and the best methods to implement when attempting to identify a user. I will then conclude my findings and outline the key steps that I will take when furthering my own research.

It is essential to first understand why it is important to determine features of a user when attempting to detect Fake News. Several studies point toward this being a useful technique when used in conjunction with other methods when considering detection. *Atodiresei, et al.,* state that, "the users who spread a news item can be a more significant feature than the words appearing in the title or description of the news item itself", this highlights the need to examine the user and their interactions, as well as the content of their posts. The table below shows the result of their study in relation to determining fake and non-fake sources by assessing the features of the user, the features of the text and both features combined. This demonstrates that the user features have a higher percentage of detection, in both fake and non-Fake News sources than using text as a singular identifier. Most importantly, it shows that when combined, there is a noticeable increase in both percentages for identification.

Table 1 Fake and Non-fake recall expressed in percentages
(LR-U user only, LR-UT user & text, LR-T text only)



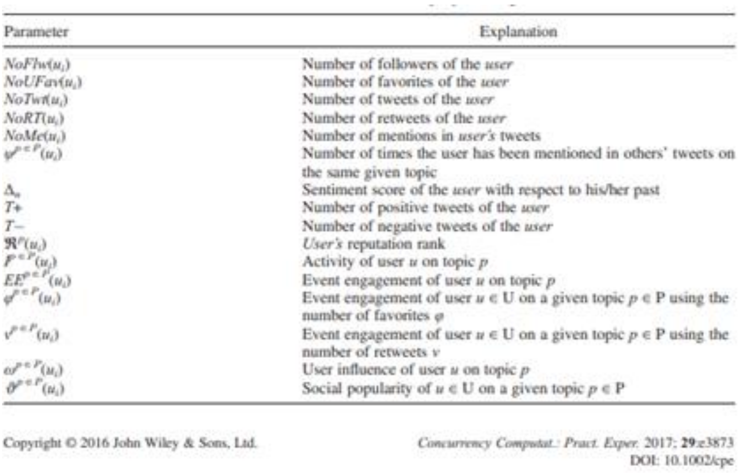
Recall	LR-U	LR-UT	LR-T
Fake:	57.64	61.14	57.25
Non-Fake:	97.40	97.75	94.18

When considering how to determine the credibility of a user, first there must be a decision made based on which features are worth examining. Twitter offers a wide range of information based on user tweets via their API so it is essential to extract only the most relevant features. Of these features, the first which seem to be impactful are those which Twitter users employ to interact with each other. These include Favourites (or likes depending on the study), Retweets, Mentions, Follows and Replies. According to Alrubaian, et al (2016), retweets, mentions and favourites appear to be the most important indicators, "retweets are considered to be one of the best indicators of user popularity from the quantitative perspective".

This conclusion is useful when considering other works. For example, Tacchini, et al.'s study found that, "the majority of the posts have few likes. Hoax posts have, on average, more likes than non-hoax posts" (2017). This study uses Facebook as its chosen social network however this platform uses likes only, rather than the several options Twitter utilises for interaction. Initially I made the hypothesis that Facebook likes / Twitter favourites would have equal weighting when considering users and that Retweeting would have a greater weight, which this study confirms. As such, the findings of Tacchini, et al., will be considered when examining users and the number of favourites on posts by users.

In correspondence to this the table below, included in Alrubaian, et al (2016) illustrates the parameters deemed most important when identifying the reputation of a user. As stated in section 2.5.a), it is worth noting that in regard to users spreading Fake News within Twitter threads and the reputation of users relating to their credibility on a given topic the table contains several parameters which are not necessary for the purpose of this dissertation.

Table 2: Most important features in regard to user reputation on Twitter



Parameter	Explanation
$NoFlw(u_i)$	Number of followers of the *user*
$NoUFav(u_i)$	Number of favorites of the *user*
$NoTwt(u_i)$	Number of tweets of the *user*
$NoRT(u_i)$	Number of retweets of the *user*
$NoMe(u_i)$	Number of mentions in *user's* tweets
$\psi^{p \in P}(u_i)$	Number of times the user has been mentioned in others' tweets on the same given topic
Δ_u	Sentiment score of the *user* with respect to his/her past
$T+$	Number of positive tweets of the *user*
$T-$	Number of negative tweets of the *user*
$\Re^p(u_i)$	*User's* reputation rank
$F^{p \in P}(u_i)$	Activity of user *u* on topic *p*
$EE^{p \in P}(u_i)$	Event engagement of user *u* on topic *p*
$\varphi^{p \in P}(u_i)$	Event engagement of user $u \in$ U on a given topic $p \in$ P using the number of favorites φ
$v^{p \in P}(u_i)$	Event engagement of user $u \in$ U on a given topic $p \in$ P using the number of retweets *v*
$\omega^{p \in P}(u_i)$	User influence of user *u* on topic *p*
$\vartheta^{p \in P}(u_i)$	Social popularity of $u \in$ U on a given topic $p \in$ P

Copyright © 2016 John Wiley & Sons, Ltd. *Concurrency Computat.: Pract. Exper.* 2017; **29**:e3873
DOI: 10.1002/cpe

Another feature worth noting when considering users, is the number of followers and friends which the user has. On Twitter, **followers** are described as the users following the specific user and **friends** are the users a specific user is following. Buntain and Golbeck (2018) include author follower count as one of the 10 most relevant features used when identifying users when comparing it to one of their datasets, although they do not state exactly the number or percentage of followers. However, Alrubaian, et al. states, **"**…users who propagate non-credible tweets seem to have fewer followers and more friends than credible users" (2016). This allows for a rule to be made stating that if a user is following more users than are following them, they are more likely to be less credible, though obviously this could apply to many users.

Furthermore, when gathering user information, it appears that examining the date the account was created can give further indication of users based on previous study:

> "…account age, which can be calculated by using the collection date minus the account creation date…there are no significant differences between the ages of credible and non-credible accounts except that the age of credible accounts tends to be older than 50 days whereas 14% of the non-credible accounts are younger than 50 days" (Alrubaian, et al., 2016)

While 14% seems relatively low, when used in accordance with the other methods, this could lead to a conclusive result. Due to the limited amount of data regarding how to identify Fake News, any additional information that can be gained is imperative to reaching a conclusion.

In addition to previous techniques, it can be worthwhile investigating the amount of posts that are made by a user in total. Evidently certain accounts are likely to have a large quantity of posts depending on factors such as whether they are promotional accounts or if they have been active for a considerable amount of time. As stated by Alrubaina, et al. (2016); "...it follows common sense that, as their objective is to spread misinformation, non-credible users send more tweets compared to credible users" and as such, it stands to reason that accounts which have a large amount of posts collected over a short period of time or that perhaps post a sizeable amount of times per day could be considered as potentially malicious.

Lastly, alternatively to determining whether a user is likely to match the behaviours of a hoax or misleading account, the Verified status can be used to make certain that a user is trustworthy, at least in terms of what Twitter allows. When a user is verified, a blue checkmark badge is applied to their username which allows other users to deem that the user is authentic and is a person of public interest. Typically, this refers to accounts which are maintained by users in music, acting, government, politics, journalism, media, business and other key interest areas. However, it should be noted that a verified badge does not imply that the user is endorsed by Twitter ("About verified accounts", Twitter,2020). As such, any users who are verified will be likely to be credible and as such, it will be necessary to gain information about whether a user is verified in order to adjust how the algorithm views them. In the study carried out by Alrubaina, et al., they state: "we removed users who could be verified as credible because of their popularity as well-known credible sources.", as such, the same steps may have to be carried out in terms of this dissertation.

2.5. c). Main Sources of Literature (Linguistic Approaches)

While examining the user features can help gain the beginnings of an understanding of whether a user is credible or not, it is important to analyse all aspects of what is being posted. There are several factors that can be considered when viewing the content of a user's post. This can range from the overall content of their tweets to the number of hashtags they include within a tweet. In this section I will give a brief review of the literature which has not been previously addressed in section 2.5.a) and elaborate on the studies which have been mentioned prior regarding the linguistic approaches they employ. Throughout this section I will also assess which methods would be best suited when considering this dissertation and the aims which it hopes to achieve.

Identifying Fake News and Fake Users on Twitter (2018)

Costel-Sergiu Atodiresei*, Alexandru Tănăselea, Adrian Iftene

Published in 2018 in association with the Faculty of Computer Science at Alexandru Ioan Cuza University, Romania, this study focuses on determining a score based on the credibility of a tweet. A first version of this project was originally created by masters students and then further developed by Atodiresei, et al. It achieves its goals by collecting Twitter data via a Twitter crawler and putting this into a database in order to apply natural language programming and analysis tools to build the final score. This score is calculated by comparing a specified tweet to a database of tweets confirmed to be true. The more tweets a user makes that contain similar language and sentiment to truthful store tweets, the higher the user score. This alternative method is the reason why this study was not included within section 2.5. a).

This study makes use of several natural language processing tools in order to achieve its goals, using APIs such as OpenCalais and Sentiment140. These APIs make up the most important module within the application as it is responsible for the analysation of tweets. Using these tools, the application extracts entities (such as companies, people, places, products, etc.) in order to compare them to true stories which contain the same, or similar, entities via Named Entity Recognition. It achieves this by splitting text into parts and can classify them as topics, social tags and generates an overall sentiment for the text and hashtags used. Further details regarding the methods and the reasons for using them will be addressed in section 2.5.d).

Automatically Identifying Fake News in Popular Twitter Threads (2018)

Cody Buntain, Jennifer Golbeck

This study is previously referenced due to the useful information within it regarding which features have been determined to be most useful when identifying non-credible users in section 2.5.a). As such, there will be no literature review in this section and the relevant information regarding the linguistic approaches used will be discussed in section 2.5.d).Additionally, building upon the literature on which this dissertation is based , the study also examines the work of Castillo, et al.(2016) and examines the use, frequency, and proportion of question marks, exclamation points and emoticons.

Within the PHEME dataset, 7 of the 45 important features had the highest performance feature set. Of these seven, three which relate to the content of tweets are the proportion of tweets sharing media, the proportion of tweets sharing hashtags and the proportion of smile emoticons. The CREDBANK dataset contains twelve features, including the three previously mentioned, tweets with multiple exclamation marks, tweets with one or more question marks and the average tweet length. Due to the overlap, it can be assumed that the three most likely indicators to be analysed would be tweets which contain media, hashtags and emoticons.

Reputation-based credibility analysis of Twitter social network users (2016)

Majed Alrubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Mohammad Mehedi Hassan*,† and Atif Alamri

This study is referenced in section 2.5.a) in regard to the important features needed to identify users. While the main focus of the study is based on users and building a reputation score, there are linguistic approaches used throughout which will be referenced in section 2.5.d). It should be noted when considering textual features in this study that at the time it was written, Twitter had not increased the character limit of tweets, and as such, some presumptions must be made when discussing this.

2.5. d) Identifying Fake News via Linguistic Approaches

Unlike the literature for user features, linguistic approaches are much less nebulous in terms of what features should be analysed. There are several tools created for the purpose of drawing conclusions from language. This simplifies the process of analysing as it is not the responsibility of this researcher to create the necessary algorithms. The focus instead should be on ensuring that the correct tool is chosen to interact with the data and making sure that the correct features are analysed in order to detect Fake News being spread by users. From the studied literature, there are several features which should be examined, however only a select few seem to infer that text posted by a user can be associated with acting in a malicious manner. In the following section I will describe each of the features accessed by previous studies and assess whether they are relevant for the purpose of this dissertation with reference to the literature.

Sentiment Analysis

"Sentiment defines the factors that affect social relationships, psychological states of users, and their orientation with respect to a given topic", (Alrubian et al. 2016). Sentiment can refer to several aspects in terms of analysis. Two of the main aspects that are referred to within the literature are:

- **Polarity -** referring to the average positive or negative feelings expressed in text
- **Subjectivity** – A score of whether text is subjective or objective.

Buntain and Golbeck (2018) use both of these features, along with disagreement when attempting to detect users spreading Fake News within Twitter threads. By assessing the sentiment of a tweet, it is possible to immediately gauge a tweet without needing to assess the topic. By knowing how biased the text is and whether it contains negative emotion, it is possible to compare the findings of previous literature and make assumptions. Alrubian et al. make reference to, Morozov E, Sen M., ‘Analysing the Twitter Social Graph: whom can we trust?’,(2014),when considering how sentiment can aid determining features of users, stating that; “researchers found that the least credible messages are associated with negative social events and contain strong negative sentiment words and opinions”. Furthermore, considering this they go on to say, **“**non-credible users are more likely to be negative, whereas credible users are more likely to be positive. Our proposed method demonstrates that the sentiment analysis is essential to determining the credibility of the user.” This confirms what was already known considering a large majority of studies which look to achieve similar goals all use sentiment analysis. In the case of this dissertation, the primary focus will be on determining the polarity of a user Tweet as this appears to be one of the most effective measures of how likely a user is to be credible.

Sentiment analysis can also include the features and aspects of a post. In the case of Costel-Sergiu Atodiresei, Alexandru Tănăselea, Adrian Iftene (2018), Named Entity Recognition is used in order to determine the social tags and topics of a post as well as the names, places, products, people, etc. which are referenced. This can be an effective tool when attempting to compare new tweets with stored database tweets that are known to be true or false. Atodiresei et al. uses this to great effect by breaking down tweets in order to compare the similarity of tweets that are known truths. While this seems to be a useful tool, the uncertainty of what is an actual Fake News post makes it difficult to say with confidence that if a post is similar to a fake one, that it is, in fact, not truthful.

Lastly, machine learning can be implemented in sentiment analysis. In particular, the Naïve Bayes classifier is often used in studies as a means of training the AI what patterns it should be looking for in terms of positive/negative sentiment in text. The equation for the Naïve Bayes Classifier is below:

Equation 1: Naïve Bayes Classifier

$$P(c\ |X)=P(xPc|x=Px|cPcPx$$

$$Pc|X=Px1|c\times Px2|c\times\ldots\times Pxn|c\times Pc$$

Source: www.saedsayad.com/naive_bayesian.htm

- P(c|X) is the posterior probability of *class* (target) given *predictor* (attribute).
- Pc is the prior probability of *class*
- Px|c is the likelihood which is the probability of *predictor* given *class*
- P(x) is the prior probability of *predictor*.

A benefit of using this classifier is that it is simple in comparison to others to build and is particularly useful when it is applied to large datasets, due to the lack of iterative parameter estimation. It is because of this that the Naïve Bayer classifier is so widely used on issues such as working through large amounts of text. It also performs well given the correct dataset and can often outperform more complex methods of classification. By applying this classification to the selected dataset of tweets, it is possible after training the classifier on examples of positive and negative tweets that the polarity and subjectivity can be determined.

Alternative Textual Features

There are other features of a tweet which can be used to make assumptions other than simply determining the sentiment of a tweet. In the work of Buntain and Golbeck (2018), they determine several important features based on the work of Castillo et al.(2010), such as an increase in the success achieved when testing against the frequency and proportion of tweets that contain exclamation points and/or question marks. This also proved to be true when considering smiling emoticons which noticeably improved sentiment detection. When testing their algorithms against the PHEME dataset, they found that 7 of the top 45 notable features performed best. Regarding text, these included tweets containing third and first-person pronouns, tweets sharing media and tweets sharing hashtags. The CREDBANK databank returned many of the same features, including tweets with mentions, tweets with multiple exclamation or question marks and the average tweet length. The three features shared by both models were the frequency of smile emoticons and the proportion of tweets with media or hashtags.

When comparing these findings with other works, it becomes apparent which features to focus on. Fortunately, the work of Alrubaian et al. contains findings which coincide with the work of Buntain and Golbeck relating to mentions, hashtags and the length of user tweets. They state that, "Almost 96% of credible users have no mentions in their tweets…whereas 46% of non-credible users have at least two mentions in their tweets" (2018). This allows for a rule to be determined in later work when attempting to detect non-credible users. They also note that: "…non-credible tweets tend to have at least one hashtag." (Buntain and Goldbeck, 2018). While this does not bring a huge amount of information, it certainly does fit with earlier works and allows for less scrutiny to be placed on tweets without hashtags. Finally, they discovered that, "…non-credible users tend to have fewer characters per tweet than credible ones…the number of words used by the credible user are more than those used by non-credible users." (2018). This would suggest that users who are closer to the maximum character count of a Tweet are more likely to be credible. Sadly, there are no figures regarding the averages of characters per tweets in this study. As stated previously, this study was written in 2016 and the character count of a tweet was increased in 2017, so it will have to be presumed that this is still the rule, even with a larger character limit.

2.5. e). Conclusion

Based on the studied literature relating to using both user and textual features, there are several rules which can be determined as a result of their findings. By taking these previously discovered connections that have been proven to aid detection, this dissertation will aim to apply the findings in order to achieve automated recognition of users spreading Fake News. I will compile a list of identifiers below that will be considered as I attempt to construct my own algorithm.

User:

- The number of friends and followers a user has. Specifically, if a user has fewer follows than friends
- The quantity of Tweets a user has made
- The date an account was created in comparison to its first post
- The verified status of an author

Tweet Content:

- The Sentiment (polarity and subjectivity) of the content of a Tweet.
- The number of favourites and retweets of a post.
- The amount of replies made on a Tweet.
- The quantity of mentions in a post, specifically if there are 2 or more.
- The quantity of hashtags in a post.
- The quantity of URLS or Media shared in a Tweet.
- The length of Tweets, specifically users with fewer than the average amount of characters used.

2.6. Research Questions

As a result of the research carried out from the literature, I have decided upon three questions which will be used to shape my project in an attempt to answer what I believe to be the most relevant points in relation to the spread of misinformation and Fake News on Twitter:

1. Can the combination of linguistic and user features accurately detect the likelihood of a user being untrustworthy?

2. Is it possible to create a scoring system based on previous research which will effectively categorise users?

3. Which data points need to be used in order to determine the features of a user?

3.0 Problem Identification

3.1. Aims

The aim of this dissertation is concise in terms of what the final deliverable should be: an application which can automatically determine a score for a user based on how likely they are to be acting in good faith.

This can be achieved based on the research questions which have been raised as a result of previous studies, and a set of steps can be decided for this dissertation. Firstly, information must be accessed and stored from Twitter in order to be analysed. This will require a crawler that will interact with the Twitter API in order to gain user and tweet details. These details then must be stored in a database in which users can be accessed and their corresponding tweets.

Following this, Natural Language Processing Tools must be implemented to access the content of Tweets within the database in order to determine the sentiment of each. A score must be decided based on these characteristics. An algorithm must then be written which will use the features concluded from the research based on users. This algorithm will assess these elements and create a score for each user, which will then be used in conjunction with the Tweet Score in order to reach a final user score. The hopes of the dissertation are that this final score will accurately reflect the nature of the user.

In the section below I will outline the specific aims of each component of the proposed application:

3.1. Twitter Crawler

To capture the data that will be used in the further elements of the project, my first goal is to create a twitter crawler which will allow for information regarding both the tweet itself and the user who posts it to be gathered. Once the selection of users to study has been made, the crawler will be given a list of usernames which it will gather information from. By accessing the Twitter API, the crawler will be able to pull out details which will be vital in all other elements of the project. By extracting this data, it will be possible to analyse which details are available from a tweet and write methods which will remove only the important features to the database. The crawler will also contain additional code to quantify the number of elements such as hashtags, emojis and mentions and to determine whether links or media are being shared within the Tweet.

3.2. Database

A database will utilised in order to normalise the information which is being pulled by the twitter crawler. This database should be separated into two sections: Users and Tweets. Each user may have several tweets attributed to their id which will allow for easier analysis of specific users when searching by said id. This will allow for several queries to be run on the database throughout the rest of the application. By having the relevant features of users and tweets as columns in the database, it will become trivial to find out large-scale details. For example, if a query is used to sort all tweets which contain two or more mentions, it could then be compared to the sentiment field. This would then allow for validation of the assertions made in relation to these two features.

3.3. Natural Language Processing

When considering Natural Language Processing, there will be two main elements which will have to be determined based on the research carried out and the aims of the dissertation. Firstly, the polarity of the Tweet content. This will be achieved by using a natural language tool which will allow for the use of a Naïve Bayes Classifier to ascertain the polarity. Ideally a rating of positive or negative will be placed into the database for the corresponding tweet depending on the outcome. This can either be a numerical value or text stating the polarity depending on what proves most valuable in testing the text scoring algorithm.

Secondly, the subjectivity of a tweet should be gathered from the content in order to decide whether a tweet is subjective or objective. This will add a second layer to the processing in order to determine a more accurate score based on the textual features.

3.4. User Data Algorithm

The aim of the user data algorithm will be to analyse the features in the user table in the database against the findings of the research. This will result in a user id being called and each feature being evaluated compared to what we know can identify users as untrustworthy. A score will be initially attributed to the user and as each feature is queried, the score will reflect the findings. For example, if the user has fewer followers than friends, they are more likely to be untrustworthy and as such, the score will be altered. Each feature will have a weighting depending on how relevant it is relating to the literature to better reflect this likelihood. This will be developed further in the next section.

3.5. Scoring

When considering the factors of the user and the tweet, it is essential to have a way to quantify the findings. By combining the scores based on the findings of the previous sections, a final score will be generated which will be the user's rating. Based on the literature of Atodiresei, et al., this score will be based out of a rating of 100, with scores above 50 being likely to be true and scores below 50 more likely to be false. Obviously, this allows for a degree of uncertainty as if a score is, for example, 45, it's far from conclusive that they are a user who is acting maliciously. It is the primary goal of the scoring functions to return a representative score which can deliver an accurate result based on all other components of the application. This score should give an impression of a user and the content that they post at a glance.

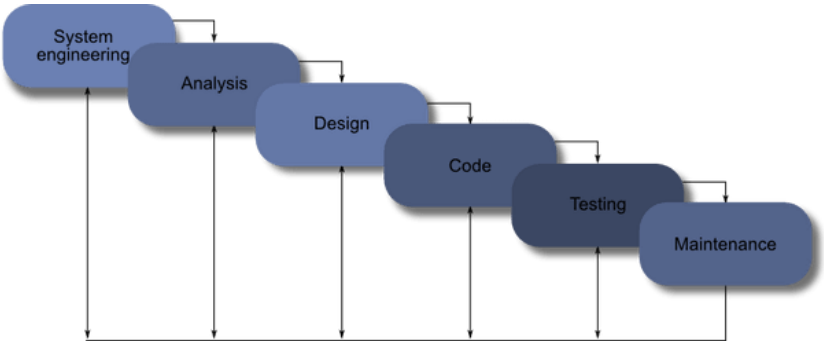
4. Design

Within this section, I will discuss the various steps taken when creating the design for the application proposed in this dissertation. This will cover the design methodology employed for the project overall and include sections on the design choices made for the individual components.

4.1. Methodology

When considering the format of this dissertation, the most appropriate choice of design methodology was to employ a strategy which used a bottom-up view of the system. Due to this project being founded and led by the process of carrying out research of other work in a similar field, it meant that initially, the components which would then build toward a final product had to be decided first. As such, the choice was made to use a waterfall model. The waterfall model implements six steps when designing an application: requirements, analysis, design, coding, testing and deployment/maintenance. **Sections 2 & 3** cover the requirements and analysis of the project and design, coding (implementation) and deployment (results) are in **Sections 4, 5 & 6** respectively. This suits the flow of the dissertation well as it involves following the steps in order and not progressing until a step is completed. An advantage of using this methodology also means that, much like this dissertation, it is possible to determine an end goal early in development. While in a formal setting this would allow for very few revisions to be made, as an individual it provides an ideal framework for developing an application.

Fig 1 Waterfall Method Flowchart



source: airbrake.io/blog/sdlc/waterfall-model

4.2. Twitter Crawler

Designing the Twitter Crawler should be straightforward due to the large amount of documentation online regarding both the Twitter API and existing applications carrying out similar tasks. Having determined the elements which I wished to extract from a Tweet and User, I was able to create pseudocode which accurately defines the steps taken:

Pseudocode 1: Twitter Crawler

```
Authenticate Twitter App
Get All Tweets From Username
Get User Details From Username

For User in Users:
    Get User_Details, Followers, Friends, Favourites,
        Account_Creation_Date, Verified_Status
    Write to Database Table User

For Tweet in Tweets:
    Get Tweet_id, Likes, Retweets, Replies, Timestamp,
    Hashtags, Mentions, Emojis, URLs, Media

    Write to Database Table Tweet
```

These steps would allow for a list of users to be specified and for each user, the necessary elements to be extracted from the API in order to add them to the database, which would allow for analysis to be carried out on each field.

While this design may seem simplistic, the hope of the application was to achieve high cohesion so that each element was acting as close to independently of the others as possible. This meant that when designing the Twitter Crawler, the only other modules of the application that it was reliant on was the database, to insert information. This information would then be used by other independent modules to act on the data. As the following sections will show, this allows for the database to the central point of the program.

4.3. Database

As stated in the previous section, the database will be designed as the main point of interaction between the other components of the application. As such, when designing the database, it is important that the data is normalised, in order to allow the different elements of the program to user the data in their own discrete ways. As shown in the above pseudocode, there will be two tables in the database, User and Tweets. These tables will operate on a one-to-many relationship, and as such, a link table will not be required between the two tables.

4.4. Natural Language Processing

Designing the natural language fully is dependent on the choice of NLP tool used to carry out the task. While there are many choices, the tasks which will be carried out can be designed prior to this choice. These will be demonstrated in pseudocode below:

Pseudocode 2: Natural Language Processing

```
Train Classifier on Training Set

Test Accuracy

Get All_Tweets

For Tweet in Tweets:

     Calculate Polarity on Tweet_Content

     Calculate Subjectivity on Tweet_Content

     Update Tweet_id Polarity, Subjectivity
```

Another feature of the natural language processing that must be considered is the choice of training set for the Naïve Bayes Classifier. When initially testing, I believe that a simple set will be beneficial for the purpose of making sure it is performing well. However, depending on the results when testing against the content of tweets it may be necessary to include a more advanced training set. I will attempt to make use of previously constructed data from online resources which have proven to provide high accuracy, and failing this, will use a dataset of tweets which have been confirmed truthful from the research.

4.5. Scoring

The scoring algorithm is perhaps the most difficult to design, due to the lack of a pre-existing formulas existing with proven success. As such, for each element of the scoring (tweet & user) there will have to be testing carried out to determine the weightings for each method in the algorithm. While the scoring will be based on previous works to some degree, many choices must be made by myself with the information that has been gathered in the research. For example, several studies agree that untrustworthy posts tend to use negative language, therefore it makes sense to reason that this is a prominent identifier. As such, should a post have a negative sentiment, it will decrease the score of the tweet.

Unlike the study carried out by Atodiresei et al., I will be initially be setting all users and tweets at a score of 100, rather than 0. I will begin by assuming a user is trustworthy as there are undoubtably more reliable users on Twitter than there are users acting maliciously. When each element of the user and the tweet is inspected, if it matches information which is known about untrustworthy users, then points will be deducted from the initial 100. Proven factors such as the account age being younger than 50 days will have a higher impact on the reduction than factors such as a tweet containing media. While this has been proven to be an identifier of untrustworthy users, it does not reason to deduct a large amount of points from posts simply for posting media or a URL. On the opposing side, should a user have a score which exceeds 100, the score will be reset to 100 as the is the maximum. The Pseudocode below will demonstrate the scoring system design, with placeholder scores:

Pseudocode 3: Scoring

```
Calculate User_score:
      Users = database.all_users
      For user in users:
            User_score = 100

            If friends < followers
                  User_score – 5

            If current_time – account_created < 50
                  User_score – 15

            If verified == True
                  User_score = 100

            If User_score > 100
                  User_score = 100

            Update User Database

Calculate Tweet Score:
      Tweets = database.all_tweets
      For tweet in tweets:
            Tweet_score = 100

            If likes > average
                  Tweet_score – 10

            If retweets > average
                  Tweet_score – 10

            If contains_url == True
```

```
                        Tweet_score - 5

                  If contains_mendia == True
                        Tweet_score - 5

                  If hashtags >= 2
                        Tweet_score - 15

                  If mentions >= 2
                        Tweet_score - 15

                  If sentiment == 'negative'
                        Tweet_score - 30

                  If tweet_length < average
                        Tweet_score - 20

                  Update Tweet Database

Calculate Total Score:

            (User Score + average_tweet_score) / 2
            Update User Total_score
```

5. Implementation

5.1. Design to Implementation

With the design phase concluded, the next task is to take what was decided upon and transfer the choices to code. The initial choice that had to be made was which programming language would be the most appropriate for the purpose of this dissertation. After discussing options with my supervisor, I was guided toward looking for projects which attempted to achieve similar goals. In carrying out the research into methods for accessing data through the twitter API and into available NLP tools, it became apparent that Python had many libraries which could be of use. In the past I had completed projects which involve interacting with an API in Python, so it became the obvious choice when deciding on the programming language used to write the application. As it was a programming language I was confident in, I was able to set about writing the Twitter Crawler which would be necessary in order to populate a database with the elements needed. In the following sections I will detail the process of transforming the design of each of the components into code and implementing them in the application.

5.1.1 Twitter Crawler

Initially, I had attempted to write code which would interact directly with Twitter API. After applying for developer access and using the appropriate application keys, I was able to begin extracting data. This process proved to be difficult due to restrictions placed on what information is available through the Twitter API and in certain cases when analysing data there were missing fields. As stated in the previous section, there are pre-existing APIs written in python to carry about these tasks more effectively.

I chose to use Tweepy, a python open-source library for accessing the Twitter API. Tweepy supports accessing Twitter via OAuth (An authentication method which requires consumer and access keys), in order to access objects and use methods that the official Twitter API offers. Implementing this involved creating classes for the Twitter Authenticator itself, a Twitter Streamer used to connect to the twitter streaming api and a Twitter Listener to print received tweets.

After reading the documentation and experimenting, the twitter crawler was at a point in which it could collect a singular tweet based solely on the *user.screen_name*. I then modified the code in order to collect all tweets from a user timeline, however, this did not always collect a complete list. This will be discussed in **section 5.3**. Once it was possible to access the status of a user, I could then extract the relevant elements from each user. I initially tested this using my own twitter account and Donald Trump's account as it allowed for one account which I could user to post and test and another which had a wealth of data, such as a high quantity of likes, retweets, etc. Using an account with a high level of activity also made testing elements such as Tweets_Per_Day easier to track.

Other than the initial set up and researching how to pull in a singular tweet, the implementation of the Twitter Crawler was fairly intuitive to us. Thanks to documentation on Tweepy, it was possible to use the Status objects being returned to determine what to be pulling from the API. The Status would be returned in a large block of code, so to better analyse it I converted it to JSON and saved it at its own file. This allowed for the file to be entered into a tool for formatting JSON by level which allowed me to see the various sections of the code and what I needed to enter to return the features I needed. There were a few challenges which I faced while creating this component of the application, however they will be discussed in further detail in **section 5.3**.

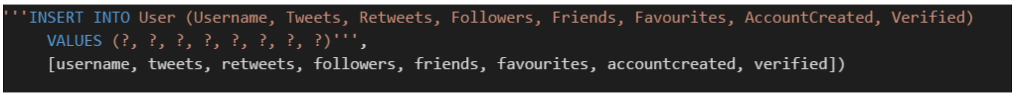
Lastly, in regard to the dataset which was used, the Twitter Crawler was used to access the TweetId's of 550 posts containing fake news and 550 posts from confirmed true stories which were gathered from Politifact. This returned the username for the user posting the Tweet and the username was then entered as the array of users to gather the relevant data from to enter into the database and then analyse.

5.1.2 Database

I chose SQLite 3 for creating my database as I had used it previously on a project and had some confidence in using it for tasks such as these. One of the main reasons was that it did not require an actual database to be set up and instead it could be accessed solely through a file. SQLite is well suited toward smaller scale projects; however, I am aware if I wanted to scale up the project, an alternative database would have to be used.

The implantation of the database was straightforward due to the planning carried out in the design phase of the project. I did, however, neglect to construct the SQL statements which would be used during the design. This did not cause too many issues during the implementation though, due to knowing which features I was pulling from the API beforehand. This allowed for me to create the two tables necessary and their associated fields easily (fig 2)

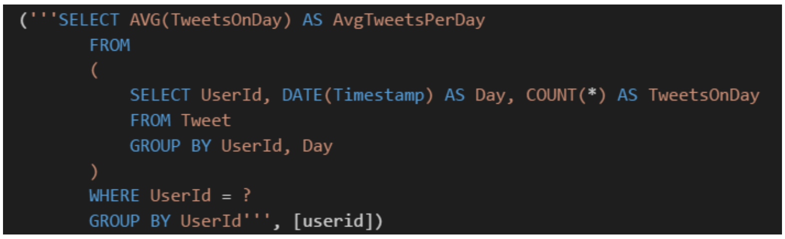
Figure 2 SQL Statement to Create User Table



```
'''INSERT INTO User (Username, Tweets, Retweets, Followers, Friends, Favourites, AccountCreated, Verified)
    VALUES (?, ?, ?, ?, ?, ?, ?, ?)''',
    [username, tweets, retweets, followers, friends, favourites, accountcreated, verified])
```

Once the tables were created, I was then able to create the methods which would be used when interacting with the data. This ranged from simply getting all tweets and all users, to more complicated statements such as determining the number of tweets a user made on average each day (fig .3). The database developed over the course of the project as I discovered more ways in which I needed to analyse and manipulate the data. In this regard, I feel that further planning the design phase could have led to this being more structured. However, thankfully the elements of design which had been carried out, such as normalisation and the decision to use only two tables meant that while new statements and methods were always being added, things never became overcomplicated.

Figure 3 SQL Statement to Determine User's Average Tweets Per Day



```
('''SELECT AVG(TweetsOnDay) AS AvgTweetsPerDay
        FROM
        (
            SELECT UserId, DATE(Timestamp) AS Day, COUNT(*) AS TweetsOnDay
            FROM Tweet
            GROUP BY UserId, Day
        )
        WHERE UserId = ?
        GROUP BY UserId''', [userid])
```

5.1.3 Natural Language Processing

While I had never used any form of natural language processing in the past, I had done some research during the design phase and read about the possible tools available to me in several of the research studies. As stated previously, one of the main reasons which I choose to use Python as my programming language of choice for this application was the large number of available libraries for tasks such as these. Fortunately, due to the amount of options and the availability of documentation for them, I was able to create Pseudocode which captured the steps before having to write any code. As such, when it came time for implementation, I was already aware of the tasks I would have to carry out.

I choose to use TextBlob, a Python library which is used for processing textual data. There were many other available choices, though TextBlob seemed to be intuitive to use and, while a powerful tool for the job, did not contain too many features which would go unused in my project. I saw no reason to use an NLP tool which was overly powerful as it seemed to be redundant for my purpose. The TextBlob API allowed for sentiment analysis and classification and contained easy to follow documentation for each, which made a reasonable choice for the tasks I was hoping to achieve.

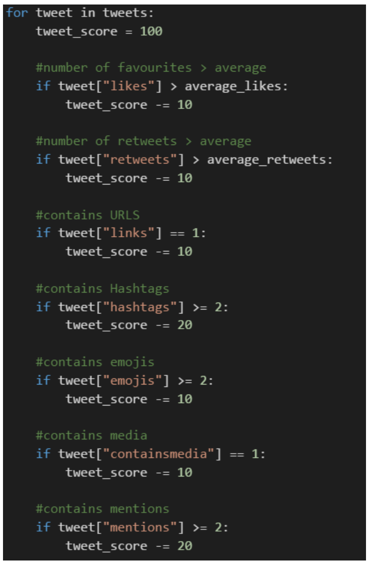
Implementing this proved to be reasonably simple, though it should be noted that the methods which I wrote utilising it and the training and testing of the classifier were far from exhaustive. First, I designed the methods for accessing the content of each tweet in the database and applying sentiment analysis to the text. Sentiment included both the polarity and the subjectivity of the text, therefore I decided to extract these two features individually in order for them to be stored in their own field in the database. Once I had tested this on some examples however, I found that the results far from accurate and several tweets which I personally would have considered negative seemed to return as positive. This further influenced my choice to include a Naïve Bayes Classifier in order to better train the NLP portion of the application.

When implementing classification, a training set of sentences and their associated polarity (positive/negative) are entered into an array which is then tested against a second list of testing data which follows the same format. A method can then be run through the API in order to gauge the accuracy of classifier. My initial training set was made up of six positive and six negative statements, with the testing set made of a further three positive and three negative statements. This resulted in an accuracy of 0.833, with 1.0 being the highest, I felt this was a reasonable result and as such did not feel that further training data was require at that time. Once these steps were complete, all that was required was to run the classification method on all tweet content in the database and assign the appropriate polarity and subjectivity to the tweet which it came from. While I don't believe the natural language processing performed poorly, as I will go on to discuss in **section 6.2**, I do feel that there could have been improvements made and will elaborate on them in **section 7.2**.

5.1.4 Text & User Details Algorithm

One of the biggest questions when implementing the algorithms for scoring users and text features was what weighting to give the calculations based on the information. As stated in **section 4.5**, it stood to reason that features which were agreed upon by multiple studies should hold the highest weighting. This meant that features such as negative sentiment, multiple mentions and an account younger than 50 days old resulted in the subtraction of the most points from the scores. When implementing the scoring for the verified status, I initially chose to have it set the user score to 100, however I feel it is unfair to assume that just because a user is verified, they are completely trustworthy. As stated previously, verified status is not an endorsement from twitter, and as such I set the score to add an additional 50 points to the user if they were verified (fig 4).

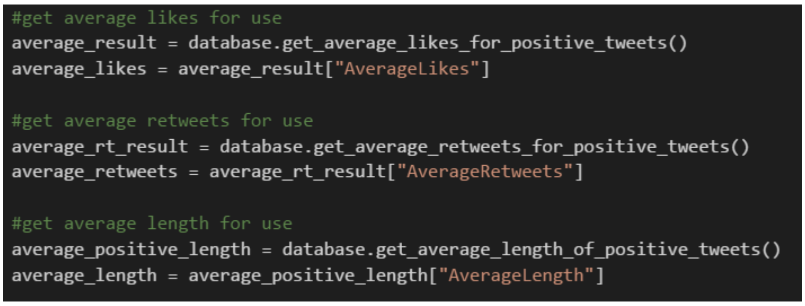
Figure 4 Example of Tweet Scoring Method



```
for tweet in tweets:
    tweet_score = 100

    #number of favourites > average
    if tweet["likes"] > average_likes:
        tweet_score -= 10

    #number of retweets > average
    if tweet["retweets"] > average_retweets:
        tweet_score -= 10

    #contains URLS
    if tweet["links"] == 1:
        tweet_score -= 10

    #contains Hashtags
    if tweet["hashtags"] >= 2:
        tweet_score -= 20

    #contains emojis
    if tweet["emojis"] >= 2:
        tweet_score -= 10

    #contains media
    if tweet["containsmedia"] == 1:
        tweet_score -= 10

    #contains mentions
    if tweet["mentions"] >= 2:
        tweet_score -= 20
```

Several things became apparent when writing the algorithms. Based on the research, when finding statistics such as the average_likes of tweets, it occurred to me that these statistics should be comparing positive and negative users. As such, the averages are made up of only the tweets with positive sentiment. For example, the research states that non-credible accounts on average have a shorter tweet length than credible accounts. Therefore, if a tweet has a shorter length than the average of all positive sentiment tweets, the overall tweet score is deducted from. This change was applied to the average likes, retweets and length (fig 5).

Figure 5 Example of Methods Called for Positive Averages



```
#get average likes for use
average_result = database.get_average_likes_for_positive_tweets()
average_likes = average_result["AverageLikes"]

#get average retweets for use
average_rt_result = database.get_average_retweets_for_positive_tweets()
average_retweets = average_rt_result["AverageRetweets"]

#get average length for use
average_positive_length = database.get_average_length_of_positive_tweets()
average_length = average_positive_length["AverageLength"]
```

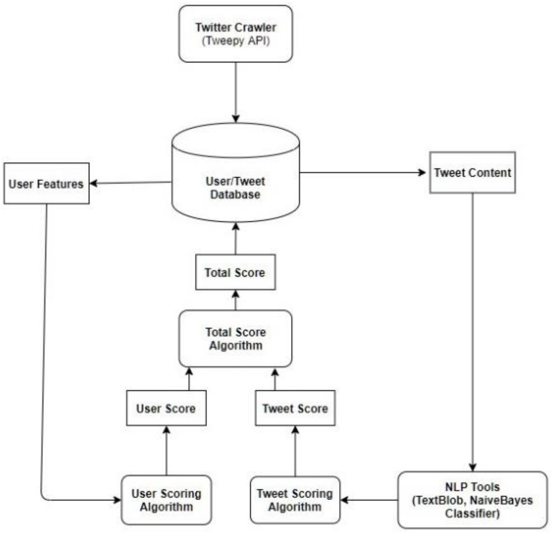
While there were a few additions which I had not considered during the design phase, such as the average amount of tweets per day, the process of implementing the scoring system was relatively straightforward. As this was a set of methods which I had written myself, the design translated to code with ease. Due to the scoring simply adjusting three figures which I had declared myself, all that was required was to check the information in the database against these algorithms. In the case of the total scoring algorithm, what I had created in the design phase turned out to be exactly what was implemented into code. As both scores were out of a maximum of 100, I could take the average of all Tweet scores and add that to the User score. I was able to then half that score to create a final score out of 100. The choice made for scoring could, however, use further alteration in order to provide more accurate feedback, this will be discussed further in **section 7.2**.

5.2. Architecture

Implementing the architecture for this application also proved to be uncomplicated as a result of a design which boasts high cohesion. By making each component mostly independent, with only the database as their main connection to the application, systems could be developed without interfering with others. The Twitter Crawler would feed information of the users and their tweets into the database. The database would then be accessed by the NLP component, which would assess the content in the Tweets table and assign a sentiment in the database. Once that was completed, the two scoring algorithms could analyse their respective table and return pass their scores into the total score algorithm which would then update the final score for all users in the database.

This architecture may seem simplistic in nature but considering the focus of this dissertation is on assessing data, it made sense to make the database the focal point of the application (fig 6).

Figure 6 Architecture Diagram for Application



5.3. Problems and Difficulties

5.3.1 Tweepy API

One of the initial issues that was discovered while accessing twitter data through the Tweepy API was that it was unable to access the full scope of information that a tweet contains. For example, it was unable to retrieve the information on the amount of replies to a tweet, which was one of my initial identifiers from the literature. This was included in the design phase but due to being unable to access it through the API, it unfortunately had to be removed.

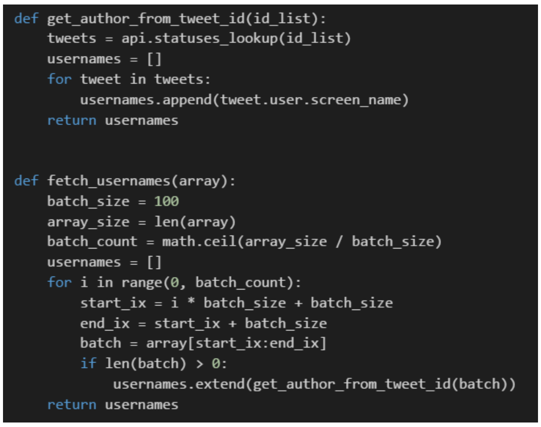
Another issue was that due to the nature of the Twitter and the Tweepy API, posts which a user has chosen to retweet will be returned with the number of retweets, but not the number of likes. If a user has made a comment on this retweet, the number of likes on that comment will be counted, but not the original likes. Therefore, when accessing a user’s tweets, any time that “RT” appears at the start of a tweet, the count “retweet_amount” is incremented and the tweet is not considered in determining the score of the user as this would affect the mean number of likes.

An additional issue with implementing the Tweepy API where that the number of tweets which could be accessed were rate limited, therefore determining the quantity of tweets a user had made was not as accurate as hoped.

While not necessarily the fault of the tweepy API, the dataset containing confirmed credible and non-credible accounts only had the TweetId associated with the fake or true posts. As such, I had to write a method which would access the API and get the author of a tweet based on the id rather than simply entering the usernames to the earlier user search method. However, when accessing the API to find the username, I discovered that Statuses_Lookup only accepts 100 IDs at a time. This meant that it was necessary to enter the selection of IDs into the API in batches. This caused some complications as it was then necessary to write a method which would take a file containing an array of IDs and then divide the list into batches to be entered depending on the specified amount of IDs that I requested to be entered (Any amount up to 100). This proved to be a harder task than expected and took quite a while to resolve. The code written to resolve this issue is shown below (fig.7).

In addition to this, running such a large quantity of data meant that it took several hours to create the users and tweets in the database. In one case the application had run for some time before encountering a duplicate user which caused a crash. This was remedied by checking the user list for duplicates in the batch before returning the list.

Figure 7 Code solution to entering large numbers of users



```
def get_author_from_tweet_id(id_list):
    tweets = api.statuses_lookup(id_list)
    usernames = []
    for tweet in tweets:
        usernames.append(tweet.user.screen_name)
    return usernames


def fetch_usernames(array):
    batch_size = 100
    array_size = len(array)
    batch_count = math.ceil(array_size / batch_size)
    usernames = []
    for i in range(0, batch_count):
        start_ix = i * batch_size + batch_size
        end_ix = start_ix + batch_size
        batch = array[start_ix:end_ix]
        if len(batch) > 0:
            usernames.extend(get_author_from_tweet_id(batch))
    return usernames
```

5.3.2 SQLite 3 Database

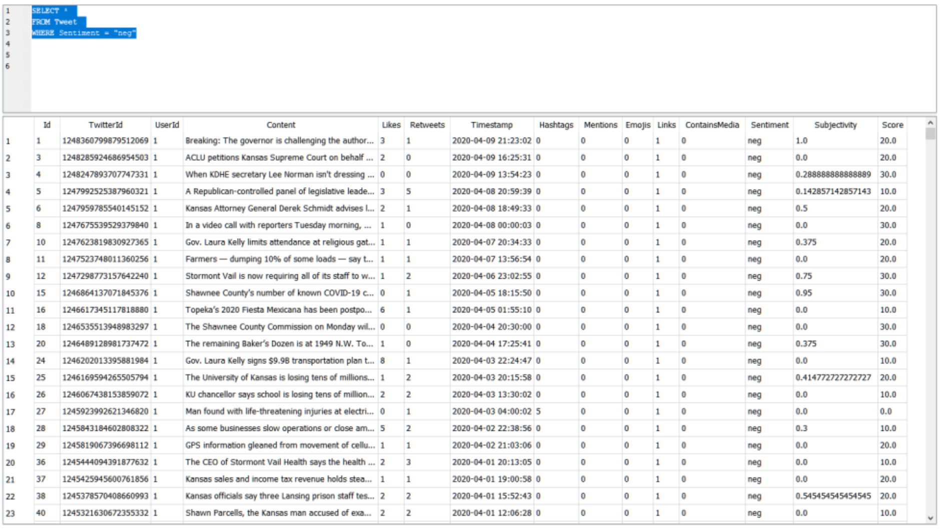
An issue with SQLite database which I choose to use was that it did not scale particularly well to a large set of data. As mentioned in the previous section, the database could take an incredibly long time to enter the data being collected by the API. This resulted in a substantial amount of downtime between runs of the application. Due to this, I had to minimise the amount of data being entered in order to have more efficiency when testing for results.

6. Results/Evaluation

6.1. Results

Due to the time required to run the application, the final dataset was made smaller to accommodate for the runtime. As such, the results are based on 50 confirmed fake news posters and 50 confirmed truthful posts.

After the initial run of the application, I was incredibly pleased to see the results. When sorting users by negative sentiment, it was clear that tweets were not simply losing points based on the sentiment of the post. Several users had a score ranging from 10.0 – 30.0, which implies that there is a correlation between the identified features which were hypothesised to identify a malicious user (fig. 8). In several cases users had a tweet score of 0.0 which meant that they had the majority of identifiers and in a few cases had entered into a negative before being reset to 0. In comparison, the majority of users who had positive sentiment were, as you would expect, in the positive range. However, it is still worth noting that even fake news tweets with a positive sentiment were still scoring around 50.0, and as such no real determination can be made regarding whether were acting maliciously or not. Although this doesn't sound necessarily positive, if the application were to return a score based on a tweet as 50/100, I would assume that it would give a user an indication that there are elements of it which should not be trusted.

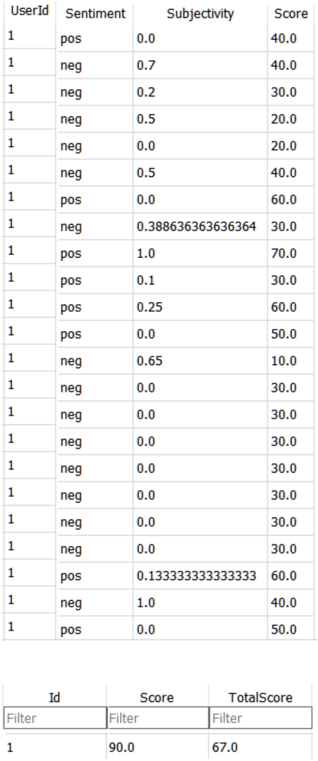


```
SELECT *
FROM Tweet
WHERE Sentiment = "neg"
```

	Id	TwitterId	UserId	Content	Likes	Retweets	Timestamp	Hashtags	Mentions	Emojis	Links	ContainsMedia	Sentiment	Subjectivity	Score
1	1	1248360799879512069	1	Breaking: The governor is challenging the author...	3	1	2020-04-09 21:23:02	0	0	0	1	0	neg	1.0	20.0
2	3	1248285924686954503	1	ACLU petitions Kansas Supreme Court on behalf ...	2	0	2020-04-09 16:25:31	0	0	0	1	0	neg	0.0	20.0
3	4	1248247893707747331	1	When KDHE secretary Lee Norman isn't dressing ...	0	0	2020-04-09 13:54:23	0	0	0	1	0	neg	0.288888888888889	30.0
4	5	1247992525387960321	1	A Republican-controlled panel of legislative leade...	3	5	2020-04-08 20:59:39	0	0	0	1	0	neg	0.142857142857143	10.0
5	6	1247959785540145152	1	Kansas Attorney General Derek Schmidt advises l...	2	1	2020-04-08 18:49:33	0	0	0	1	0	neg	0.5	20.0
6	8	1247675539529379840	1	In a video call with reporters Tuesday morning, ...	1	0	2020-04-08 00:00:03	0	0	0	1	0	neg	0.0	30.0
7	10	1247623819830927365	1	Gov. Laura Kelly limits attendance at religious gat...	1	1	2020-04-07 20:34:33	0	0	0	1	0	neg	0.375	20.0
8	11	1247523748011360256	1	Farmers — dumping 10% of some loads — say t...	1	1	2020-04-07 13:56:54	0	0	0	1	0	neg	0.0	20.0
9	12	1247298773157642240	1	Stormont Vail is now requiring all of its staff to w...	1	2	2020-04-06 23:02:55	0	0	0	1	0	neg	0.75	30.0
10	15	1246864137071845376	1	Shawnee County's number of known COVID-19 c...	0	1	2020-04-05 18:15:50	0	0	0	1	0	neg	0.95	30.0
11	16	1246617345117818880	1	Topeka's 2020 Fiesta Mexicana has been postpo...	6	1	2020-04-05 01:55:10	0	0	0	1	0	neg	0.0	10.0
12	18	1246535513948983297	1	The Shawnee County Commission on Monday wil...	0	0	2020-04-04 20:30:00	0	0	0	1	0	neg	0.0	30.0
13	20	1246489128981737472	1	The remaining Baker's Dozen is at 1949 N.W. To...	1	0	2020-04-04 17:25:41	0	0	0	1	0	neg	0.375	30.0
14	24	1246202013395881984	1	Gov. Laura Kelly signs $9.9B transportation plan t...	8	1	2020-04-03 22:24:47	0	0	0	1	0	neg	0.0	10.0
15	25	1246169594265505794	1	The University of Kansas is losing tens of millions...	1	2	2020-04-03 20:15:58	0	0	0	1	0	neg	0.414772727272727	20.0
16	26	1246067438153859072	1	KU chancellor says school is losing tens of million...	2	2	2020-04-03 13:30:02	0	0	0	1	0	neg	0.0	10.0
17	27	1245923992621346820	1	Man found with life-threatening injuries at electri...	0	1	2020-04-03 04:00:02	5	0	0	1	0	neg	0.0	0.0
18	28	1245843184602808322	1	As some businesses slow operations or close am...	5	2	2020-04-02 22:38:56	0	0	0	1	0	neg	0.3	10.0
19	29	1245819067396698112	1	GPS information gleaned from movement of cellu...	1	1	2020-04-02 21:03:06	0	0	0	1	0	neg	0.0	20.0
20	36	1245444094391877632	1	The CEO of Stormont Vail Health says the health ...	2	3	2020-04-01 20:13:05	0	0	0	1	0	neg	0.0	10.0
21	37	1245425945600761856	1	Kansas sales and income tax revenue holds stea...	1	1	2020-04-01 19:00:58	0	0	0	1	0	neg	0.0	20.0
22	38	1245378570408660993	1	Kansas officials say three Lansing prison staff tes...	2	2	2020-04-01 15:52:43	0	0	0	1	0	neg	0.545454545454545	20.0
23	40	1245321630672355332	1	Shawn Parcells, the Kansas man accused of exa...	2	2	2020-04-01 12:06:28	0	0	0	1	0	neg	0.0	10.0

Figure 8 Results when Assessing Negative Tweets

When studying the user scores, impact of the average tweet can result in them being identified as non-credible. As stated previously, with some scores hitting the minimum result of 0.0, it greatly impacts their result compared to those who average within the 50.0 – 70.0 range. As a result, this allows to determine the quality of a user based upon not only possible factors relating to their account, but on the history of the posts which they have made (fig 9.)



UserId	Sentiment	Subjectivity	Score
1	pos	0.0	40.0
1	neg	0.7	40.0
1	neg	0.2	30.0
1	neg	0.5	20.0
1	neg	0.0	20.0
1	neg	0.5	40.0
1	pos	0.0	60.0
1	neg	0.388636363636364	30.0
1	pos	1.0	70.0
1	pos	0.1	30.0
1	pos	0.25	60.0
1	pos	0.0	50.0
1	neg	0.65	10.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	neg	0.0	30.0
1	pos	0.133333333333333	60.0
1	neg	1.0	40.0
1	pos	0.0	50.0

Id	Score	TotalScore
Filter	Filter	Filter
1	90.0	67.0

Figure 9 Effect of Low Tweet Score On Total Score

6.2. Evaluation

In evaluating the results of this dissertation, I believe it would be best to compare what has been achieved to the research questions which have been laid out in **section 2.6**. In this section I will consider each of the questions posed and evaluate whether or not they have been answered and examine the quality of the results:

Can the combination of linguistic and user features accurately detect the likelihood of a user being untrustworthy?

I believe that this particular question had been answered by previous studies long before I had asked it. However, other research involves alternative steps toward achieving this goal. I would suggest that a finding in this research is that in order to accurately detect an untrustworthy user, it is essential to have information about both the user and their post. Had I applied the findings of the research to the user or text alone, I've no doubt that my findings would have been vastly different. For one, of the research previously carried out, there seem to be few discernible features attributed to users which can accurately determine their credibility. If I were to judge accounts based solely on the user, I would surmise the very few users would be flagged as untrustworthy. Secondly, by text analysis alone, it is possible to gather data regarding a tweet which can point toward the alignment of a user. It should be noted that without taking the entire scope of a user's posting history and attributing the average results to them, it is difficult to truly determine their goals. Additionally, while conclusions can be drawn with language analysis and in regard to the content of a Tweet, I believe that to truly achieve an accurate picture of the user, you must use all of the information which is available to you.

As such I would say that this question has been answered, at least when regarding the scoring system which was put in place in reference to the literature. The results demonstrate that with a combined dataset of the user's activity and the content of the posts they make, that proven negative accounts are being ranked lower than those which are confirmed as positive.

Is it possible to create a scoring system based on previous research which will effectively categorise users?

While I do not believe that the scoring system put in place is as accurate as it could be with alterations and more statistics, the results do point toward it reaching the desired outcome. When reflecting on the results, it is clear that the scoring can determine a tweet to have a score of 0.0, which implies that certain posts by proven non-credible users match almost every rule which had been set out with the goal of detection in mind. When looking at the average scores for tweets of non-credible accounts, it is also rare to see a score surpass 50.0, which would imply that even when attempting to emulate a positive post, there are still indicators which stop them from passing into the "positive" range of results.

As a result of this, I would say that this question has been answered. When previous research is applied to creating a scoring system, it is possible to accurately and effectively categorise a user based on these features.

Which datapoints need to be used in order to determine the features of a user?

Without repeating the points of the previous answer, I believe it's clear this dissertation has correctly identified several of the features of a user which could be used for detection. As a result of the research, certain factors contribute more than others, such as an accounts age. In fact, I would say that all the datapoints outlined in **section 2.5.5** have a noticeable impact when attempting to determine features of a user, and as such, I would say this question has been answered.

7. Appraisal

7.1. Critical Evaluation

7.1.1 Overview

The completion of this project is both very satisfying and somewhat frustrating. By the end of this dissertation I have managed to ask and answer the research questions which I set out by creating an application which, I feel, could be developed further to the point where it could potentially have real world use when considering fake news detection on Twitter. However, despite this, I can't help but feel that a series of mistakes and mismanagements resulted in a project which, although I am proud of, could have been outstanding.

As the following sections will go on to discuss, my initial planning, organisation and interaction with my supervisor was a key misstep in the creation of this project. By not appropriately utilising my time, the resources available or the opinion and guidance of those more knowledgeable in this field, I feel I have squandered an opportunity to excel.

While I am pleased with the result, I feel that this dissertation was a project of two halves, both with their own faults and successes. Looking back on the project, I am impressed with the outcome I managed to achieve considering the initial lack of progress which was made. I hope that while reviewing these sections and the attached appendix, that it is clear that what was initially an early mismanagement of the project transformed into a piece of work which I came to be very proud of, despite its weak beginnings.

7.1.2 Planning & Research

As I am sure will become apparent in the following section and the attached Interim Report in the appendix, the first semester of the project was not handled well regarding initial planning and research. While I was considering the end goal of the project and the steps toward it, my choice of reading for research was related more to the subject as a whole rather than an academic approach toward it. Much of the reading was based on websites regarding fake news and other applications which achieved similar aims. While this may seem reasonable, I would describe it as surface level research at best.

As a result, after deciding on an aim for the project, in meetings with my supervisor I was unable to bring much of my own opinion into discussion. Nor was I was able to discuss the works of others in a manner that was constructive, other than simply mentioning that I had read about other applications doing similar things. I will discuss the supervisor meetings further in the following section, though it is necessary to mention during the initial planning and research. I was urged to step away from simply reading from websites and directed toward academic papers as a source of information. While I did take this advice, I don't think I understood the importance of carrying out research on previous work and what that would do for the path my project would take.

Within the first six months of the project, I would describe my deliverables as a vague goal will very little thought behind it other than what I hoped it would achieve. What reading I had completed was useful, but not relevant toward the methods which I would have to employ when creating my own application. While my research had not been ideal, I had written the code for my twitter crawler, though that code was vastly different to the final version. It is difficult to discuss the shift in work ethic without referencing meetings too heavily, but simply put, the interim meeting opened my eyes to just how little work I had done. This will be explained further in **section 7.1.3**. What is worth noting is that after that meeting, my drive to research, plan and create a better project was vastly different to the first semester.

When talking to my supervisor, I was told that when he been working on his most recent research, he had put six months into reviewing the work that had been done before on the topic. I would attribute this as a turning point for me, as a I realised that in order to truly understand the problem and the methods which had been put toward solving it, I had to look to the wealth of academic studies which were available to me. My strive toward salvaging this project began during the break between the first and second semester.

When planning my project, I set aside an initial month exclusively for research. While it may seem counter-intuitive to do so with a short deadline, what I had been told by my supervisor stuck with me. As this was during the break, there was a lot of downtime, regarding other university projects and external work, so it seemed the most productive to use this time to fully explore the associated reading. I began by finding two research projects which seemed to attempt to solve similar issues and achieve similar goals. From there, I took notes on each and compiled a list of referenced literature in each which seemed to be relevant to my dissertation. I continue to follow this format until I felt I had compiled a list of both useful and relevant research, but also the notes on each and the sections which I felt would guide my project. I hope that in reading **section 2** that the amount of research I carried out is apparently, while I am aware there certainly could be more.

Upon completing my initial list of studies, I began to outline the steps which others had taken and created a list of potential methods, as well as downloading potential datasets from studies which provided them for later use. This allowed me to begin to realise the scope of my project and the information with which I would be interacting. I believe this is what allowed for my design to be comparatively quite close to what I implemented, as I was aware of many of the features I'd require before ever writing code other than the crawler.

Had I began researching and planning earlier in the lifecycle of this dissertation, it is frustrating to consider what could have been achieved with the extra time. When considering the entire month spent on research at the start of the year, if that time had been spent developing on something which had been more concrete before then, I feel I could have completed several improvements to the final project. These will be discussed in **section 7.2**.

Lastly, the lack of research prior to my Interim Meeting meant that when meeting with my second supervisor I was ill equipped for the meeting and any questions or discussions which could have been raised. I felt as though I was wasting the time of both of my supervisors. While they were incredibly patient and offered useful advice and insights into how I should progress, I think the realisation of the scope of work which I still had to carry out was what went on to shape how my dissertation was carried out.

7.1.3 Organisation and Meetings

As stated in the previous section, my initial planning and research was only exacerbated by a lack of organisation and failure to attend meetings with adequate discussion topics and progress.

While I entered the project with the intention of keeping relevant minutes for each supervisor meeting, this began to be neglected as the timeline progressed, as did keeping regular weekly meetings (**section 8.2.2)**. While there are only three entries in the appendix for meetings, this only represents the minutes taken, rather than the number of meetings which took place. As is apparent when reviewing the minutes, my supervisor had made a push toward reading academic literature most meetings and, as stated in the previous section, this advice had not been followed.

In reviewing the Interim Report (**section 8.2.4)**, this enforces the statements made in the previous section regarding only having a slight knowledge in the direction of the project and the subject. The main issues which appears across both the supervisor meetings and the Interim Meeting is that the scope of the project was too broad. I believe this relates back to the lack of research carried out before this point and being unaware of possible methods which could be applied to the problem. The Interim Meeting does not reflect well upon this dissertation, though it very much reflects the amount of work which had been carried out to that point. I hope that in reviewing the progress made since that time that it is apparent just how large of an impact that meeting had on me.

Despite this, one of my largest regrets comes after the Interim Meeting. After an encouraging discussion with my supervisor I felt shaken but was aware of the amount of work that had to be carried out and was determined to complete it. In the time between that meeting and the hand in of this report, I have made no attempts to communicate with my supervisor. Initially I had planned to go and put all of my effort into making strides toward progress in this dissertation and to then contact him in order to show that I had taken the advice I had been given onboard. However, with the more time that passed over the break, the more I wanted to hold off in order to show an improvement in the work. This became something a habit and I would often write emails with my attached work and then not send them, convincing myself that I would do it the next day. I'm unsure what led to this being the case, as in hindsight it is an incredibly unprofessional way to act and with no doubt has impacted the overall results of this dissertation. I'm not sure whether I was worried about the work not being of a high enough standard or being told that I had done non-relevant research, but I could not bring myself to send an update. I'm aware that this is foolish as if I had been progressing in the wrong direction, I've no doubt that my supervisor would have guided me to the correct path.

As stated before, the handling of my meetings and contact with my supervisor is easily the biggest oversight during this project. I have often said during the process of working on this dissertation during my second semester that I simply wish I'd put in the time in the beginning and utilised all of the resources available, as they are in place to help me succeed. If I could start the project over, I would certainly do many things differently, my only hope is that the work submitted is of a high enough standard to demonstrate just how determined I was to amend my early, and unfortunately ongoing, mistakes. I would also like to apologise for the stark contrast in the following sections as I discuss what when well.

7.1.4 Design Methodology

I feel that the methodology which I choose proved to be very efficient for this dissertation. As I was working independently, it meant that the requirements in each stage of the methodology were always met before moving on to the next step. It allowed for me to set a timeframe for individual sections in order to better plan how much time I should spend doing each task, which when working to a deadline was crucial. Lastly it enabled me to create an early end goal which could be developed upon as I completed more work. While I knew what I wanted to create in the early stages of the project, being able to build upon it based on research made the designing and implementation a more streamlined process as I could develop of the project upon existing work. As such, I am pleased with my choice in methodology, though I am aware that larger scale projects or projects which involved a team and clients would not be best suited to this style.

7.1.5 Implementation

I feel that a strong emphasis on research and design, enforced by the choice of methodology, meant that the actual implementation of the code proved to be quite straightforward. Obviously, as with all implementation, there were some unforeseen problems and adjustments which had to be made along the way. Thanks to having effectively planned the methods required and identified the information which I needed to locate, I was able to focus more on how to get what I needed rather than spend time looking for what I needed.

While in retrospect there are a few elements of the application I would change, such as the choice in database, I am confident that the program is well realised and implemented to a high standard when comparing it to what I had initially sought out to create. The code includes all methods I had wanted to include and interacts with all features which I could feasibly obtain while carrying out the processes successfully, despite the inevitable long runtimes.

7.1.6 Results

Overall, I am incredibly pleased with the results of this dissertation. Obviously, I regret the handling of certain previously discussed aspects, but this does not detract from the outcome of what I managed to create. Regarding the results of the actual application I am very proud of the work which I have achieved and the progress which I managed to make toward answering my research questions. I feel that I have been able to answer each of my questions and in fact have excelled above what I had first imagined I would be able to complete. While the planning and organisation of the project do detract from this, I sincerely hope that the results speak for themselves in terms of what I set out to achieve.

7.2. Future Work

In this section I will outline the various steps which I believe could improve the application, making for both more useful results and a more intuitive system overall.

The first improvement which I would consider in the implementation of an orchestrator which could be run in order to run the components of the application in order. Currently I am responsible for running each separate component in the command line. By doing this it would reduce downtime and make for a much more streamlined process. Additionally, the inclusion of a user interface which could be accessed in order to interact with the four components of the system would make for a more user-friendly way to access and display the information which is found.

Secondly, the database is dropped and rebuilt each time it is run in order to make sure that it is receiving the most up to date information. Obviously, this is far from ideal. A great improvement which could be made is the use of a real time service that monitors twitter feeds and invokes the crawler when there is a change, such as a new tweet or an additional like on a post. Batch inserting queries rather than running them individual would greatly speed up the run time when entering new users. It would also avoid opening and closing the connection to the database with ever new user.

Additionally, further testing and experimentation with the scores attributed for user features and text could have resulted in more accurate results. Due to the lack of concrete figures from the research I carried out it was difficult to choose how many points to add and subtract from accounts. By further experimenting to see how changes affect the final score and by seeking out exact figures on the percentages of users which match the problem features in a non-credible dataset, it could be possible to create a scoring system which reflects the situation better with a higher accuracy.

Lastly, a more substantial and comprehensive training set for the natural language processing would result in more accurate performance when determining the polarity and subjectivity of user posts. Ideally, a training set constructed from known false and true tweets would lead to overall improved accuracy in the application when given other examples. Another improvement could be achieved by adding more options for how polarity is expressed could also help to when calculating final scores. By being able to classify a tweet as somewhat positive or somewhat negative, it would allow for a more nuanced look at how the choice of language influences this study. Finally, if the natural language processing could determine which emoji is being used and attribute polarity to it, this could equally help in improving the accuracy of both the sentiment and the overall score in the application.

7.3. Knowledge/Skills Gained

This dissertation allowed for me to not only gain insight into a topic which I was already interested in, but also aided me in developing upon previous skills and learning new ones. I will divide this section into two parts: Programming and Fake News:

7.3.1 Programming

In this section I will discuss the new skills and knowledge which I have learned as well as the skills which I have developed on. Although it was not the most challenging implementation, this was my first use of natural language processing. I was fortunate that the documentation was detailed, as it allowed for a much quicker learning period when using TextBlob. This was also my first attempt at creating a crawler to gather information. While this was made easier by using Tweepy, it was still only a tool which I had to utilise to reach my goals.

While I have worked once with an API in the past, this was certainly the most in depth and challenging coding I have done on that topic. Previously I had only used an API to retrieve figures for a website, but this work led to much more manipulation of the data which was being retrieved. I had also used SQLite in the aforementioned project, so was aware of how to use it to a degree. The large-scale nature of the data and using the database to analyse information however was an entirely new experience for me. While in retrospect I wouldn't say that SQLite was the ideal tool for the job, I would say that I feel much more confident in my ability to use SQL for querying a database.

7.3.2 Fake News

While dealing with an issue such as fake news, it can become a little disheartening to see the flood of misinformation and the several studies carried out which, as of yet, have been unable to find a definite solution to the problem. However, in reading *Flash Eurobarometer 464 (Fake News and Disinformation Online)* I did feel optimistic in seeing that, at least across Europe, people are beginning to see the spread of fake news as an issue not only on the small scale, but to democracy as a whole. This study contained a lot of useful information considering opinion, which was surprisingly positive.

Secondly, I learned much about how misleading accounts operate in order to mimic actual news and reach as many users as possible. While the aim of my study was to find the main identifiers of a fake news author and their posts, in reading the many studies, I learned much more than I had first anticipated about the topic. This gave me an enthusiasm toward my research that previously was not as present. While I'm aware that the aim of this dissertation is not to necessarily solve the problem, I was passionate about the research which I could add to the topic.

While not fake news related, I also feel that I have learned both how to use the research of others to build upon and guide a project, as well as the benefits in carrying out adequate research. Prior to this dissertation I often felt that I was capable of doing a small amount of reading and then carrying out my work individually. I now realise that one of the major benefits of the computing community is that work is always evolving and building upon the works of others is fundamental. There is no reason to reinvent the wheel, and I feel that is one of the best pieces of knowledge which I have gained.

7.4. Conclusion

In Conclusion, this dissertation has shown that it is possible to analyse the features of users and their posts in order to determine judgements on how likely they are to be acting in good faith. By comparing confirmed truth with confirmed falsities, it is clear to see that a trend occurs amongst users hoping to propagate fake news among others. While this study only covers a small portion of users comparatively to the millions of Twitter users active, it does succeed in at least indicating users who are worthy of scepticism, which I believe if a success.

While the users creating these posts and attempting to spread misinformation are a major factor to blame, the onus is on us, as readers to consider what we are reading and to make the steps to truly attempt to ascertain the truth. In a modern era where there is no way to stop fake news being spread, each and every one of us should take that extra minute to think: Is this true?

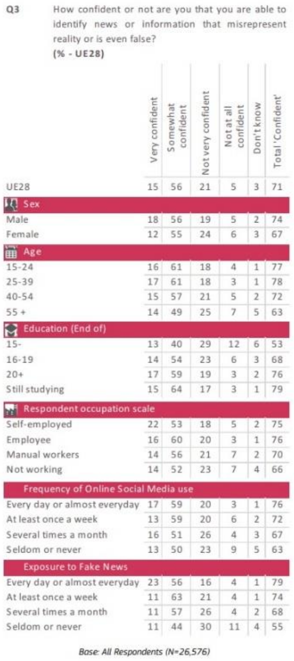
7.5. Acknowledgements

I would first and foremost like to thank my family for their continued support throughout my academic career. I would also like to thank Maegan for being a constant font of inspiration and for keeping me working hard to achieve the best that I can. I would like to say a huge thank you to Sam Hood, for his patience and enthusiasm, which without I’m not sure this dissertation would even exist. Lastly, I would like to give thanks to my supervisors and the staff at Edinburgh Napier for the fantastic work that they continue to do.

8. Appendix

8.1. Research Figures

Figure 1 – Socio-demographic in confidence to identify false information

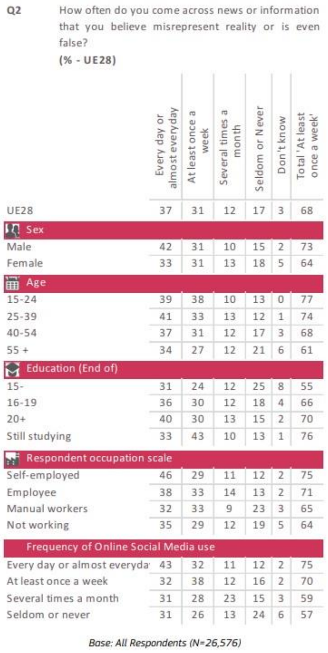


Q3 How confident or not are you that you are able to identify news or information that misrepresent reality or is even false?
(% - UE28)

	Very confident	Somewhat confident	Not very confident	Not at all confident	Don't know	Total 'Confident'
UE28	15	56	21	5	3	71
Sex						
Male	18	56	19	5	2	74
Female	12	55	24	6	3	67
Age						
15-24	16	61	18	4	1	77
25-39	17	61	18	3	1	78
40-54	15	57	21	5	2	72
55 +	14	49	25	7	5	63
Education (End of)						
15-	13	40	29	12	6	53
16-19	14	54	23	6	3	68
20+	17	59	19	3	2	76
Still studying	15	64	17	3	1	79
Respondent occupation scale						
Self-employed	22	53	18	5	2	75
Employee	16	60	20	3	1	76
Manual workers	14	56	21	7	2	70
Not working	14	52	23	7	4	66
Frequency of Online Social Media use						
Every day or almost everyday	17	59	20	3	1	76
At least once a week	13	59	20	6	2	72
Several times a month	16	51	26	4	3	67
Seldom or never	13	50	23	9	5	63
Exposure to Fake News						
Every day or almost everyday	23	56	16	4	1	79
At least once a week	11	63	21	4	1	74
Several times a month	11	57	26	4	2	68
Seldom or never	11	44	30	11	4	55

Base: All Respondents (N=26,576)

figure 2 – Socio-demographic frequency of false information



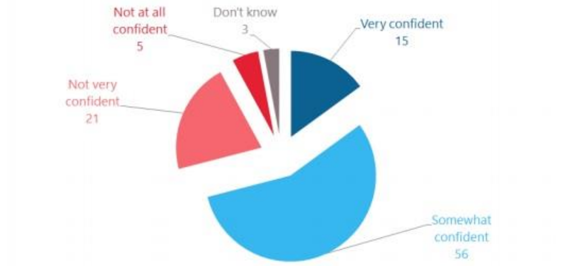
Q2 How often do you come across news or information that you believe misrepresent reality or is even false?
(% - UE28)

	Every day or almost everyday	At least once a week	Several times a month	Seldom or Never	Don't know	Total 'At least once a week'
UE28	37	31	12	17	3	68
Sex						
Male	42	31	10	15	2	73
Female	33	31	13	18	5	64
Age						
15-24	39	38	10	13	0	77
25-39	41	33	13	12	1	74
40-54	37	31	12	17	3	68
55 +	34	27	12	21	6	61
Education (End of)						
15-	31	24	12	25	8	55
16-19	36	30	12	18	4	66
20+	40	30	13	15	2	70
Still studying	33	43	10	13	1	76
Respondent occupation scale						
Self-employed	46	29	11	12	2	75
Employee	38	33	14	13	2	71
Manual workers	32	33	9	23	3	65
Not working	35	29	12	19	5	64
Frequency of Online Social Media use						
Every day or almost everyda	43	32	11	12	2	75
At least once a week	32	38	12	16	2	70
Several times a month	31	28	23	15	3	59
Seldom or never	31	26	13	24	6	57

Base: All Respondents (N=26,576)

Figure 3 – Confidence in identifying false information (graph)

Q3 How confident or not are you that you are able to identify news or information that misrepresent reality or is even false?
(% - EU)



Base: All Respondents (N=26,576)

8.2. Organisation & Meetings

8.2.2. Supervisor Meeting Diary

EDINBURGH NAPIER UNIVERSITY

SCHOOL OF COMPUTING

PROJECT DIARY

Student: Darren Simpson **Supervisor: Nikolaos Pitropakis**

Date: 01/10/19

Objectives:

- Discussed route for project & deliverables that would be expected by the end of the project.
- Discussed ways to begin project, deciding on the routes that were possible
- Discussed problems with AI inability to determine irony or humour.

Progress:

- Decided on using a twitter crawler to collect data
- Arranged meeting time as each Thursday at 11:30 until academic year break (21/12/19)

To do:

- Research how to sort accounts on twitter by popularity and rank. Make sure information is recent and relevant and uses technology that I am familiar with.
- Bring 3 ideas of possible methods regarding how this would be achieved.

Student: Darren Simpson **Supervisor: Nikolaos Pitropakis**

Date: 10/10/19

Objectives:

- Look into literature regarding:
- ranking/ordering twitter profiles
- Malicious twitter accounts
- Complete IPO

Progress:

- Begun working on twitter crawler
- Looked into the steps needed to begin project – Discussion: How to rank twitter accounts? Should they have a benchmark of what is a positive or negative account. Sort by negative language.

To do:

- Continue working on twitter crawler
- Read literature

Student: Darren Simpson **Supervisor: Nikolaos Pitropakis**

Date: 08 /11/19

Objectives:

- Fake news too broad a topic – Decide on a definitive project idea
- Look at how natural language can be used to determine a score for a twitter account based on negativity or negative opinion

Progress:

- Completed Twitter crawler
- Begun reading literature

To do:

- Email Nikolaos with definitive project goal
- Schedule meeting with second advisor
- Find literature based on delivering a scoring system on twitter
- Find Literature based on machine learning
- Find Literature on natural language processing

8.2.3. Initial Project Overview

Initial Project Overview

SOC10101 Honours Project (40 Credits)

Is it Possible to Use Artificial Intelligence to Determine the Validity of Accounts on Twitter to Combat Fake News?

Overview of Project Content and Milestones

The Main Deliverable(s):

An application which uses a data source gathered from twitter in order to identify the patterns of bias and fake accounts in order to create a scoring or rating depending on the posts, follows, followers, likes and retweets of the account.

The Target Audience for the Deliverable(s):

The target audience would be exclusively twitter users and potentially twitter as a company due to the nature of the application. It would cater to any users who have an interest in making sure that the information they are receiving is truthful and trustworthy on a platform that lends itself to misinformation and opinion.

The Work to be Undertaken:

The work that needs to be undertaken can be broken down into 4 sections:

1. Investigation – Looking into previous academic work that has been done on the platform that relates to my own project. This also will help to build a greater understanding of how fake news is being spread and the effective ways in which it is being combatted.

2. Creating a Twitter Crawler – This will allow me to gather the necessary data set from twitter in order to both find the trends amongst the accounts which I am observing and to train my application on.

3. Designing and building an AI – Creating an AI which uses what it has learned from the data set which can observe an account or trend and analyse it in order to determine a rating of how trustworthy it is as a source.

4. Testing, Analysis and Evaluation – With a functioning application it will be important to test in order to find out the shortcomings and where it fails in creating an accurate rating, should this be an issue. After testing I will analyse what has been discovered and what has been ascertained from the project and evaluate how this was carried out and what could have been improved or changed in hindsight.

Additional Information / Knowledge Required:

Twitter Crawler: Creating a crawler to collect data from accounts that are relevant to my project

Natural language processing: Formatting the collected data in such a way that it can be read and interpreted by the AI in order to determine information from what is written (e.g. is a comment negative or positive?)

AI and determining weightings: While I've worked with AI in the past, this seems to be a very different use of it compared to what I have done before. Using the AI to take the information it is given and use it to give a reasonable decision on an account is something that I will need to do more research into.

Information Sources that Provide a Context for the Project:

Organisations:

- Knight Foundation (https://www.knightfoundation.org/features/misinfo)
- Botsentinel – Trollbot (https://botsentinel.com/trollbot-network)

References:

- Atodiresei, C., Tănăselea, A. and Iftene, A. (2018). Identifying Fake News and Fake Users on Twitter.
- Agrawal. R, de Alfaro. L, Ballarin. G, Moret. S, Massimo Di. P, Tacchini. E, Della. V, Marco. L. (2019). Identifying Fake News from Twitter Sharing Data: A Large-Scale Study
- Montangero. M, Furini. M. (2015) TRank: Ranking Twitter Users According to Specific Topics

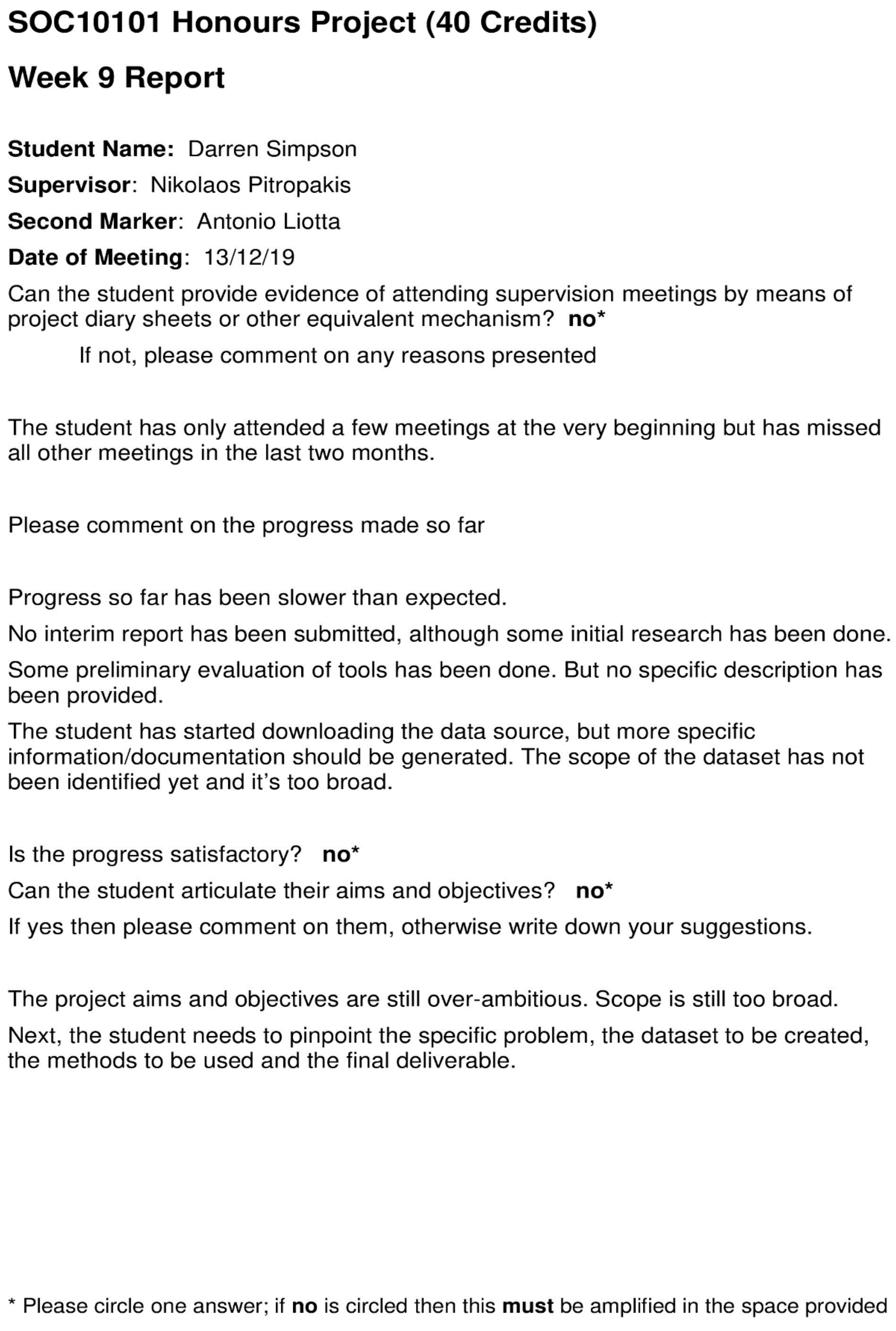
The Importance of the Project:

I feel that in this current political age we are seeing social media and the internet being manipulated in ways that traditional media could not be in the past. With fake news being published and pushed not only by those who stand to benefit from it, but also by those who are unaware of it being falsified, it is more important now than ever to check your sources and think critically. However, due to clickbait articles and fast paced nature of social media, users have begun to take news at face value from headlines and minimal research. I hope that this project can be used in order to take some of the strain away from users in order dismiss obvious malicious sources or to give them pause for thought when viewing posts from an account.

The Key Challenge(s) to be Overcome:

In this project, I believe that my main challenge will be the use of natural language processing as it is a concept I am fairly unfamiliar with and different from anything I have done in the past. Along with this, I feel that creating an accurate and relevant weighting for the AI to judge accounts on may be one of the bigger challenges.

8.2.4. Interim Meeting Report



SOC10101 Honours Project (40 Credits)

Week 9 Report

Student Name: Darren Simpson

Supervisor: Nikolaos Pitropakis

Second Marker: Antonio Liotta

Date of Meeting: 13/12/19

Can the student provide evidence of attending supervision meetings by means of project diary sheets or other equivalent mechanism? **no***

If not, please comment on any reasons presented

The student has only attended a few meetings at the very beginning but has missed all other meetings in the last two months.

Please comment on the progress made so far

Progress so far has been slower than expected.

No interim report has been submitted, although some initial research has been done.

Some preliminary evaluation of tools has been done. But no specific description has been provided.

The student has started downloading the data source, but more specific information/documentation should be generated. The scope of the dataset has not been identified yet and it’s too broad.

Is the progress satisfactory? **no***

Can the student articulate their aims and objectives? **no***

If yes then please comment on them, otherwise write down your suggestions.

The project aims and objectives are still over-ambitious. Scope is still too broad.

Next, the student needs to pinpoint the specific problem, the dataset to be created, the methods to be used and the final deliverable.

* Please circle one answer; if **no** is circled then this **must** be amplified in the space provided

Does the student have a plan of work? **no***

If yes then please comment on that plan otherwise write down your suggestions.

The plan has not been produced yet. The student has some general ideas of the work he wishes to perform. However, objective, plans, tasks, milestones, and deliverables should be specified.

Does the student know how they are going to evaluate their work? **no***

If yes then please comment otherwise write down your suggestions.

The student has started looking into the project scope and possible methodology. However, since the deliverable is not clear, and the methods is undefined, the evaluation method has not been considered yet. The above should be the priority now.

Any other recommendations as to the future direction of the project

The project is now behind schedule. It is essential to complete an interim report, complete the literature review, define objectives, deliverables, tasks and methods.

It is strongly recommended to attend periodic supervisory meetings.

EDITED BY Prof. Antonio Liotta and sent over email.

Signatures: Supervisor Second Marker

Student

The student should submit a copy of this form to Moodle immediately after the review meeting; A copy should also appear as an appendix in the final dissertation.

* Please circle one answer; if **no** is circled then this **must** be amplified in the space provided

9. References

- Omnicoreagency.com. (2020). • *Twitter by the Numbers (2020): Stats, Demographics & Fun Facts*. [online] Available at: https://www.omnicoreagency.com/twitter-statistics/

- Atodiresei, C., Tănăselea, A. and Iftene, A. (2018). Identifying Fake News and Fake Users on Twitter. *Procedia Computer Science*, 126, pp.451-461.

- Castillo, C., Mendoza, M. and Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5), pp.560-588.

- AGRAWAL, R., DE ALFARO, L., BALLARIN, G., MORET, S., PIERRO, M.D., TACCHINI, E. and DELLA VEDOVA, M.,L., 2019. *Identifying Fake News from Twitter Sharing Data: A Large-Scale Study.* Ithaca: Cornell University Library, arXiv.org.

- *European Commission, Brussels (2018): Flash Eurobarometer 464 (Fake News and Disinformation Online). TNS opinion, Brussels [producer]. GESIS Data Archive, Cologne. ZA6934*

- de Alfaro, L., Polychronopoulos, V., & Shavlovsky, M. (2015). *Some Like it Hoax: Automated Fake News Detection in Social Networks*. Santa Cruz: School of Engineering, University of California.

- About verified accounts. (2020). Retrieved from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

- Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M., & Alamri, A. (2016). Reputation-based credibility analysis of Twitter social network users. *Concurrency And Computation: Practice And Experience*, *29*(7), e3873. doi: 10.1002/cpe.3873

10. Bibliography

- de Alfaro, L., Polychronopoulos, V., & Shavlovsky, M. (2015). *Some Like it Hoax: Automated Fake News Detection in Social Networks*. Santa Cruz: School of Engineering, University of California.

- Bovet, A., & Makse, H. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1). doi: 10.1038/s41467-018-07761-2

- Naive Bayesian. (2020). Retrieved from https://www.saedsayad.com/naive_bayesian.htm

- Tweepy. (2020). Retrieved from http://www.tweepy.org/
- Shu, K., Mahudeswaran, D., & Liu, H. (2018). FakeNewsTracker: a tool for fake news collection, detection, and visualization. *Computational And Mathematical Organization Theory*, *25*(1), 60-71. doi: 10.1007/s10588-018-09280-3
- Yada, K. *Data Mining for Service*.
- Natural Language Processing - Python - Tutorialspoint. (2020). Retrieved from https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_python.htm
- Twitter Scraping for Sentiment Analysis - PromptCloud. (2020). Retrieved from https://www.promptcloud.com/twitter-scraping-sentiment-analysis/
-