Media Content Analysis of Covid-19 Virus Using Natural Language Processing Techniques

MSc Research Project
Data Analytics

Anaelle Rouxel
Student ID: X15022421

School of Computing
National College of Ireland

Supervisor: Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Anaelle Rouxel

Student ID: X15022421

Programme: Master of Science in Data Analytics **Year:** 2020

Module: Research Project

Supervisor: Catherine Mulwa

Submission Due Date: 17th August 2020

Project Title: Media Content Analysis of Covid-19 Virus Using Natural Language Processing Techniques

Word Count: 9,196 **Page Count**…25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Anaelle Rouxel

Date: …17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Media Content Analysis of Covid-19 Virus Using Natural Language Processing Techniques

Anaelle Rouxel
X15022421

Abstract

Covid-19 outbreak in December 2019 spread worldwide during the first half of 2020, affecting populations and economies. The pandemic created an emergency context and emphasized on the lack of knowledge in the domain of crisis informatics. The impact of coronavirus on media usage and population emotional reactions will contribute to the current state of art. This research project objectives are to assess the public's interests and responses to Covid-19 and assess the use of social media and news media in communicating on the emerging virus, using Natural Language Processing (NLP) techniques. Extracted topics relate to the pandemic development with infection cases updates and protection measures. Themes are more varied in news media (environment, entertainment, economy, vaccine) whereas Twitter data evoke behaviours instructions and more negative latent topics (search for the virus origin, testing). LDA models achieved a coherence score of 0.381 on tweets and 0.475 on news corpus. The sentiment analysis showed the importance of the neutral class, 100% of news articles and 90.2% of tweets fall into this category. The distribution showed 7.2% of tweets are positive and 2.6% are negative. Statistical paired t-test comparing tweets scores means before and after text pre-processing confirmed the operation impacts polarity results. Tweets were more distributed to the neutral class after pre-processing. Lexicon-based emotion detection showed a dominance of fear in tweets against trust evoked in news, whereas sadness and anticipation emotions are similarly present in both corpuses analysed. The project also featured a literature review and concluded on the research gaps concerning media content analysis using NLP techniques.

1 Introduction

From December 2019, the outbreak of coronavirus affecting populations worldwide gave rise to greater amount of information being shared online with velocity. The first two patients diagnosed with Covid-19 were reported on 8th December 2019 in the region of Wuhan in China. As of 1st August 2020, the virus spread across 213 countries, contaminating 17.8 million people and taking more than 683,000 lives[1]. The amount of media attention dedicated to Covid and the vocabulary used to describe the virus reinforced the fear in many countries with confirmed cases.

[1] Figures dated from 1st August 2020, https://www.worldometers.info/coronavirus/

The project is motivated with personal interests in semantic and accurate language use, combined with the domain of mental well-being. This research is an opportunity to learn about the field of Natural Language Processing, regarding how human language is computed for text mining and machine learning applications. Moreover, the Covid-19 crisis is a novel topic of general interest with a rapidly evolving situation since its outbreak. With the reduction of people's movements and lockdowns, online media became the main source of information for instant updates on the situation and communication between stakeholders at multiple levels (local, national, international).

1.1 Research Question

Covid-19 virus outbreak led to high volume of text content being generated online, it is unprecedented given the suddenness of events and the emergency context it created. This is a great opportunity to gain insights on a current phenomenon by applying unsupervised machine learning and performing a content-based analysis with NLP techniques. Based on findings from the literature review, the domain of crisis informatics and veracity assessment of online content data are recent and relatively immature to date. As a result of these gaps in the literature and the uniqueness of the topic of research, the project focuses on answering the following research question (RQ).

RQ: *"How can media (social and news) content analyzed using natural language processing techniques (LDA topic modelling, emotions detection, sentiment analysis) provide insights on information discussed and emotional reactions in regards to Covid-19 pandemic?"*

And the two sub-research questions of the research which were answered are as follows:

SUBRQ1: *"To what extent natural language text pre-processing can improve the computation accuracy of polarity scores on Twitter data ?"* Difference in polarity scores will be evaluated using statistical paired t-test.

SUBRQ2: "*To what extent sentiments and emotions evoked in the information on Covid-19 spreading through both Twitter and The Guardian news media differ?"* Frequency scores of emotions detected using a lexicon-based approach will be compared. Difference in sentiment scores will be evaluated using statistical Welch's t-test on unpaired samples with unequal variances.

1.2 Objectives and Contribution

The research project aims at discovering hidden patterns in text data from Twitter social media and The Guardian news media to give insights on areas of interests and sentiments towards Covid-19 pandemic. A critical literature review on NLP techniques for media content analysis is carried out as **Objective1**. Performing data collection is **Objective2** and pre-processing datasets is **Objective3**. **Objective4** involved the exploratory analysis as preliminary work to modelling. **Objective5** involved the extraction of news and social media topics with LDA topic modelling algorithm. **Objective6** consisted in a sentiment analysis with polarity computation

from text feature, both media are analysed and compared to give insights on areas of interests and sentiments towards the Covid-19 pandemic. **Objective7** was to measure the impact of text data pre-processing task on computing polarity to answer ***SUBRQ1***. Sentiment scores were computed before and after pre-processing Twitter data, results were compared using paired t-test statistical test. **Objective8** was to measure the difference of sentiments between Twitter and The Guardian using Welch's non-parametric t-test to answer ***SUBRQ2***.

The novelty of the research project is to use NLP for the comparative content-based analysis of two media data sources tackling a current worldwide concern that is Covid-19 pandemic. The main contribution of the research is to create new knowledge in the domain of crisis informatics. This refers to studying the forms of interaction and usage patterns of media during emergency and crisis events (Reuter, et al., 2018). This domain of research focuses on the role of technology in supporting the collaboration of people and government agencies during emergencies situations (Zhang, et al., 2020). Minor contribution is the recognition of the importance of neutral class in sentiment analysis.

First, the existing literature will be reviewed, second the methodology will be explained. Then the design specification will be presented. Exploratory analysis will be presented in Chapter 4. Implementation, evaluation and results obtained from machine learning and NLP techniques will be detailed in Chapter 5. To finish, findings and outcomes will be discussed in Chapter 6 and the conclusion will end with future work recommendations.

2 Literature Review

Covid-19 pandemic context significantly impacted populations and economies where lockdowns were imposed. In this unprecedented situation, online media became a crucial source of information and outlet for people deprived from social relationships. The World Health Organisation (WHO) coordinated a response to the phenomenon of "infodemic" caused by this pandemic. This term describes an overabundance of information, either true or false, occurring during an epidemic (WHO, 2020). The topic of coronavirus dominated media and conversations since its outbreak. Media in this research refer to both news media and social media platforms, unless specified to differentiate the two.

2.1 Media Communication During Crisis

With social networks and highly connected population, information spread online rapidly. The use of social media increases greatly in context of emergency and crisis events. According to Traylor et al.(2019), information (especially false) is initially distributed over social media, such as Twitter and Facebook, and later passed on to mainstream media platforms such as traditional radio, television on online news websites (Traylor, et al., 2019). This shows traditional and official sources of information react later than the public. This finding is aligned

with the results of the research from Liao et al. (2020) on Covid-19 showing the public response was seen earlier on Weibo[2] than the government agency accounts (Liao, et al., 2020).

2.1.1 Crisis Informatics

Crisis informatics is the field of research of personal communication and information technology that explores the forms of interaction and usage patterns during emergency and crisis events. Two relevant types of use for the research project scope are presented in sections 2.1.2 and 2.1.3.

2.1.2 Communication from and to the Public

Content generated by citizens for self-support to communities refers to virtual communication of citizens with each other via social media. Such platform enables people to coordinate among each other, share information, help each other. This provides a network of social relationships and a supporting climate face to a perceived threat. People tend to react rationally to a crisis, rarely panic or loot, and are not helpless (Helsloot & Ruitenberg, 2004). Another function of social media is for users to express solidarity as in the 2011 Egyptian uprising, offer support and give emotional encouragement as in the earthquake that affected Japan in 2010. In their research on Ebola messaging via Twitter, Wong et al. (2017) grouped posts into four categories: "information giving, news update, event promotion, and preparedness (Wong, et al., 2017). Moreover, when uncertainty is caused by extra information and misinformation because of disorganised and chaotic online behaviours, the trend noted in reaction is a larger amount of collaboration on the social platform (Valecha, et al., 2013). A quantitative analysis of tweets with frequency plot of specific keywords in Ebola-related posts during the crisis revealed that rumours spread like true news on the social platform (Jin, et al., 2014).

2.1.3 Communication from Authorities to Citizens

Crisis communication from authorities to citizens increasingly include social media into their official communication to reach out the largest audience in a rapid and efficient manner. Messages disseminated in this fashion relate mainly to instructions on how to behave during emergencies (Reuter, et al., 2018) and to correct misinformation caused by the chaotic use of social media" (Kaewkitipong, et al., 2012). Researchers recommend the automation of cross-media checks to verify the relevance of posts pursuant to crises (Kaufhold & Reuter, 2016). The challenge of the information assessment is the velocity for checking, the absence of truthful sources to check against, and the timeframe: what is deemed true at an instant may be considered as false at a later instant. Public often turn to social media for information. Tang et al. (2018) reviewed the communication approaches on social media regarding outbreaks of emerging infectious diseases. They identified three objectives for data mining experiments: (1) assess the public's interest in and responses to the disease, (2) examine the use of social media in communicating and (3) evaluate the accuracy of disease-related medical information on social media (Tang, et al., 2018).

[2] Sina Weibo is a major social media platform in China.

2.1.4 Content-based Features

As defined by Shu et al.(2017), content features refer to information that can be directly extracted from text, there are linguistic features such as vocabulary, syntax, semantic (Shu, et al., 2017). This is the meaning in language, and there are three linguistic features to use for content-based approaches.

- Syntactic features relate to the number of content words and the frequency of specific Part-of-Speech patterns. Complexity of sentences indicate the reliability of the information (Vosoughi, et al., 2017).
- Lexical features refer to the actual word usage. Expressions can be analysed by combining *n* adjacent words (bigrams, trigrams or *n*-grams length). Researchers Traylor et al. (2019) used machine learning methods that take into account text mining to assess the likelihood an article with quotes is fake, their model was 63.33% precise (Traylor, et al., 2019).
- Semantic features concern sentiment analysis or opinion mining, they are often extracted by NLP techniques. The aim is to extract features based on opinions and emotions expressed in the text. It is also possible to extract topics with Latent Dirichlet Allocation (LDA) algorithm (Blei, 2003). Machine learning (ML) and deep learning approaches for fake news detection were successful with neural networks and "word embeddings" which is a language modelling and features learning technique in NLP.

2.2 Text Mining with Natural Language Processing Techniques

News shared online must be understood from a linguistic perspective, therefore an analysis of natural language is necessary. This motivated to review literature for text mining techniques considering NLP and unsupervised text data analysis.

2.2.1 Stance Analysis Approaches

Stance detection consists in evaluating the position (stance) of the text towards a target or a set of targets (Kucuk & Can, 2020). Problems related to the automatic analysis of all human affects includes sentiments and emotions stance-based classification.

Emotion recognition uses stance detection to categorize text according to common emotion classes (joy, sadness, anger, disgust, anxiety, surprise, fear, and love) (Kucuk & Can, 2020). Such analysis can be carried out using the NRC Emotion Lexicon. It is a crowdsourced list of English words and their associations with eight basic emotions and two sentiments (negative and positive) (Mohammad & Turney, 2010). Sentiment analysis refers to computing a polarity score to classify a piece of text or document into a positive, negative or neutral class (Bold, 2019). The technique consists in finding key words in the text and mapping them to a dictionary that will assign scores or weights to the words (Lane, et al., 2019). Sentiment dictionaries contains set of rules (called lexicons), text is classified by analyzing words, grammar construct, rules of language and semantics (Beigi, et al., 2016). Dattu & Gore (2015) classify sentiment analysis on Twitter data in three techniques: lexical analysis, ML based and hybrid analysis. To perform a sentiment analysis on Twitter data, they labeled it to apply supervised algorithms SVM and Naïve Bayes classifiers, both ML models achieved 89% accuracy. Research shows

the introduction of the neutral category can even improve the overall accuracy by learning better the distinction between positive and negative stances (Koppel & Schler, 2006) (Taboada, et al., 2011). In the case of unlabeled data, lexical analysis is suitable given it uses a dictionary of pre-tagged lexicons (Dattu & Gore, 2015). Unsupervised ML consists in grouping unsorted data by learning from its hidden structure, patterns, similarities and differences, without having previous information on the data (Mittal & Patidar, 2019). To date, limited research has been done on emotions and sentiments analysis related to medicinal matters and using context-based approach to assess vocabulary (Zeng-Treitler, et al., 2008).

A study of Weibo posts related to Covid indicated that personal posts were more likely to show empathy to affected people and blame other individuals or government, and express worry about Covid epidemic (Liao, et al., 2020). As stated by Aslam et al. (2020) in their analysis of 141,208 news headlines on coronavirus outbreak, the media attention dedicated to Covid-19 between January and June 2020 does not reflect the death rate from this virus remains proportionately low compared with other viral infections, such as influenza (flu) or HIV. The results obtained from the sentiment analysis show that 51.66% of total news headlines are associated with negative sentiments. A smaller portion generate positive sentiments (30.46%) and the remaining 17.87% are categorized as neutral news. Their findings from emotions analysis show the language describing the virus is rather negative with words such as “deadly virus”, “public health emergency”, and “outbreak”. This lexicon increases the sentiment of fear and negative emotions in countries with confirmed cases. The researchers also highlighted the fact that sentiment trajectory score over time started from the negative region and declined in a first period. Emotional valence never turned positive or neutral, a sharp increase towards more negativity was noticed around mid-April 2020 (Aslam, et al., 2020).

2.2.2 Topic Modelling Techniques

A semantic analysis of datasets can be done with topics extractions. This unsupervised NLP method consists in analysing the link between terms present in a document to extract themes. Several techniques are available such as Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA).

LSA was designed in 1990 (Deerwester, 1990), it is an effective technique to use in terms of rapidity and simplicity to obtain results. However, drawbacks are critical: the components generated as output lack of interpretability and hinder the identification of topics. The representation of results is not efficient and LSA works best to give accurate results when applied to very large set of documents and vocabulary (George & Birla, 2018). LDA is a more recent method, effective and popular. It uses the dataset as training data to calculate the Dirichlet distribution of document-topic distribution (Blei, 2003). It computes a term-topic matrix from a collection of documents. The algorithm output can be easily interpreted by humans, topics extracted from a corpus are composed of a list of words that are most strongly associated with a given topic identified (George & Birla, 2018).

In their analysis of tweets during the Ebola outbreak, Odlum & Yoon (2015) identified topics discussed on social media Twitter included risk factors, prevention education, disease trends

and compassion (Odlum & Yoon, 2015). Also in Liao et al. (2020) analysis of Weibo posts, common themes identified in personal and government content shared included updates on "epidemic situations, general knowledge of the new disease, and policies, guidelines, and official actions" (Liao, et al., 2020). However, government posts were morel likely to express instrumental support and praise people or organisation.

Topic modelling results can be evaluated by coherence and perplexity metrics (Vayanskya & Kumar, 2020). The *k* number of latent topics to extract is adjusted to increase the performance of the model. It is recommended to initially set a high *k* and then optimise the model following a first set of results obtained (Lane, et al., 2019). Perplexity measures "how surprised a model is" when new unseen data is tested on the model (Kapadia, 2019). A low perplexity score indicates the probability model is good at predicting the sample (Fola, 2019) however this metric presents limitations according to literature and measuring a change in perplexity scores is suggested to compare models (Zhao, et al., 2015). Therefore the coherence measure is seen as a better metric for the quality of topics, the higher score the better the model (Röder, et al., 2015).

2.2.3 Deep Learning with Neural Networks

Traditional ML approaches are based on manually designed and time-consuming features extraction task that may result in biased features (Ma, et al., 2016). Therefore, deep learning (DL) methods have an advantage over traditional ML methods where the identification of relevant features for the analysis pose a challenge. DL approaches can learn hidden patterns from simpler inputs both in text and content variations. The method does not model relevant input features as for ML but model the network itself in a way that enables the tasks to be solved efficiently (Bondielli & Marcelloni, 2019). Most widely implemented artificial neural networks are Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). CNN is now gaining popularity in NLP and have been employed in 2017 by Chen et al. with single and multi-word embeddings for solving stance and veracity classification of tweets (Chen, et al., 2017). Volkova et al. (2017) used both RNN and CNN approaches to identify suspicious and trusted news posts. They employed both word sequences and linguistic and network cues deception. The evaluation of both methods shows average precision close to 1.00 and outperforms baselines (Volkova, et al., 2017). It has been demonstrated that hybrid models, that consists in a mixture of RNN and CNN, outperform all baselines too for fake news detection (Wang, 2017).

2.3 Identified Gaps

To date, social media data and news media data have not been analysed in conjunction regarding epidemics. At the early stage of Ebola outbreak in 2014, Odlum & Yoon analysed tweets to give insight into social media and public health outbreak monitoring and information (Odlum & Yoon, 2015). Analysis of Ebola information circulating on Twitter and Sina Weibo in 2014-2015 assisted public health agencies to develop their social media communication strategies (Fung, et al., 2016). More recently with Covid pandemic, content analysis of social media Weibo was done to investigate Chinese public engagement and government

responsiveness in the communications during the early epidemic stage (Liao, et al., 2020). Aslam et al. (2020) carried out sentiments and emotions analyses on more than 140,000 headlines from the 25 top English news sources to measure the impact of Covid-19 outbreak on mental wellbeing (Aslam, et al., 2020). There was no emphasis on the neutral sentiment class in their experiment. Therefore, to the best of the candidate's knowledge, there is no comparative NLP analysis of two sources of information, being a social media and news media, on Covid-19 pandemic. Focusing on semantic analysis to extract themes evoked in the two media as well as comparing the polarity of text between them (social media vs. newspaper) and within the same source (before and after pre-processing Twitter social media data) will contribute to the existing knowledge and fill gaps in literature. There is no past research with the objectives of assessing the public's interest in and responses to emerging virus and assessing the use of social media and news media in communicating on Covid-19. Evaluating the accuracy of Covid-related medical information on media is out of scope of the project.

Based on findings from the Literature Review and completing **Objective1**, the sentiment analysis on unlabelled data will be performed using with a lexicon-based approach to extract polarity with unsupervised technique. To discover hidden semantic structures in text body and allow for comparison between Twitter and The Guardian, LDA topic modelling is retained as being the most suitable technique to implement for the research project. To define the context, the pandemic situation developed while this research project was carried out.

3 Research Methodology

This chapter explains the scientific approach used to carry out the project from its start to end, the architectural technical design and the process to collect data in the data persistence layer.

3.1 Covid-19 Methodology Approach

The data mining approach selected to tackle the problem is called Cross Industry Standard Process for Data Mining, and commonly known as CRISP-DM. This methodology serves as a guideline to lead the research project. It is deemed to be the appropriate approach as it starts with a business understanding of the problem. Then data must be gathered, prepared and explored. Therefore CRISP-DM methodology provides a structured approach to planning data mining project by performing tasks in a certain sequence. The process is made of six stages in the following order: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.

The method is a modified CRISP-DM methodology designed for a Media Content Analysis of Covid-19 Virus Using Natural Language Processing Techniques (Fig.1). The methodology has been adapted to this specific data analytics project, the research and identification of hidden patterns with machine learning. There will be no deployment of the model in an operating system per se.

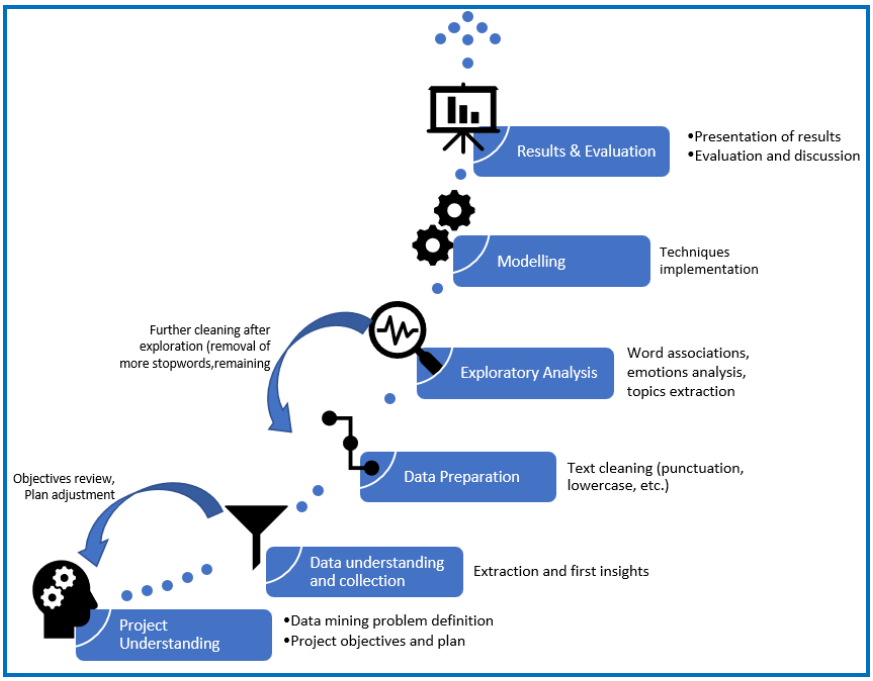

Figure 1: Covid-19 Methodology

Project understanding: First, the project requires a business understanding through setting objectives and defining clearly the data mining problem with the suitable plan to resolve it. The project objectives are related to analyse unexplored and very recent text data from media using NLP techniques, data mining for modelling topics and sentiment analysis from information spread on Covid-19 through the media platforms selected.

Data understanding and collection: The next step is to collect data and define the scope. In this project two different data sources are used: Twitter and The Guardian[3] (more precisely the web section dedicated to Covid-19 news). Data was extracted using Python programming language, in the period ranging from 24th March 2020 to 9th July 2020.

Data preparation: Data is cleaned and briefly explored. The outcome of this step is the construction of the required data. In this phase of research data cleaning was done using Python. The aim was to transform raw data (text content feature) into a format that was exploitable for analysis and the extraction of valuable insights. Cleaning consists in removing stop words, punctuation, special characters, hashtags, URL, convert to lowercase, etc.

Exploratory Analysis: Using RStudio, Python and Tableau to complement with visuals, this phase is closely related to Data Preparation. Exploration leads to going back and forth to the previous stage of data preparation until the data input to modelling is in the suitable form (Vorhies, 2016). Preliminary findings will be presented at this stage of exploratory analysis.

[3] https://www.theguardian.com/world/coronavirus-outbreak

Modelling: The approaches implemented are selected based on findings from the literature review and additional readings from reliable machine learning and technical sources[4]. Natural Language Processing techniques such as Emotion recognition, Sentiment Analysis and LDA topic modelling are implemented.

Results and Evaluation: Results and insights gained from the analysis will be presented, evaluated and discussed. As a last stage, the project objectives will also be evaluated against the outcomes.

3.2 Design Specification

The project is executed using a three-tier architecture: first with the visualisation tier, then the computational tier and third the data persistent tier, as depicted in Figure 2. The first component of visualization corresponds to the client layer. This is the user interface and the presentation of results to stakeholders. The second component of computational tier is the business logic layer. It captured NLP techniques implemented for exploratory analysis and modelling. The third component is the data persistence in this research project. It is a complex and work-intensive layer represented by the back-end data. Data was scraped from online sources using Python programming language, then formatted and cleaned in order to constitute the foundation of the text content-based analysis. This phase will be detailed in the next section 3.3 Data Persistence.

Tools used for the project are Python for scraping, processing and cleaning data as well as modelling. RStudio is used for Exploratory Analysis. Tableau is used for data visualisation as part of the first brief exploratory analysis and presentation of findings.

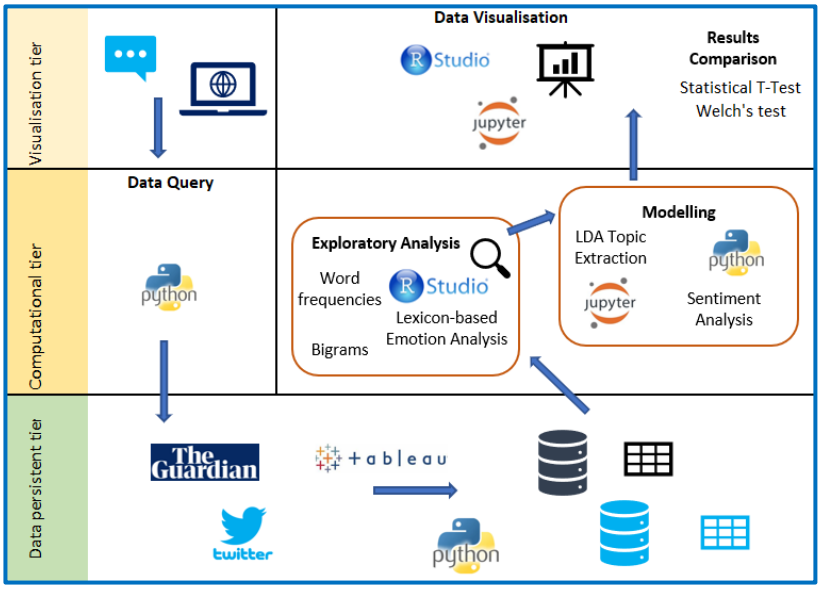

Figure 2: Design Specification

[4] https://medium.com/ and https://machinelearningmastery.com/

3.3 Data Persistence

This section presents the back-end data used for the project: a collection of news articles and tweets covering the topic of Covid-19 virus and completing **Objective2**.

3.3.1 Twitter Data

3.3.1.1 Data Collection

Social media data was obtained through Twitter API. The hashtags "covid" and "coronavirus" were selected to scrap tweets during a defined period. The choice for these specific keywords was supported by the verification on 20th March 2020 of popular hashtags associated with the virus5.

Instructions were run with Anaconda Prompt to scrap up to 10,000 tweets for each keyword and export results in unstructured json files. The method is depicted in Fig.3. The period was bounded from 1st December 2019 to 23rd March 2020 for the first wave, and up to 23rd June for the second batch[6]. The virus was first detected in China on 17th November 2020 but cases were officially reported to the WHO for the first time on 8th December as confirmed by The Guardian newspaper (Davidson, 2020). This information supported the start date selection for capturing relevant conversations on Twitter from the beginning of the pandemic period.

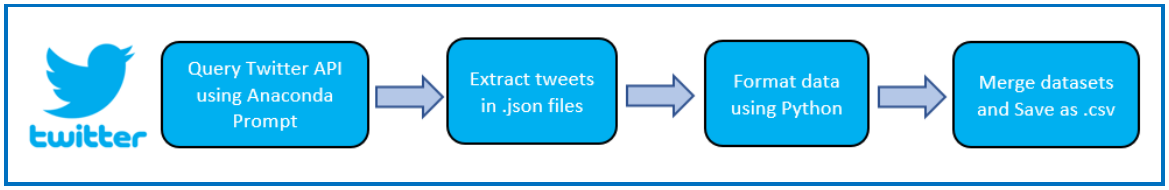

Figure 3: Collect tweets using Anaconda & Python

3.3.1.2 Formatting

Tweets were extracted in separate json files and formatted into a data frame using Python to keep the fields of interests. Features selection was based on findings from the literature review and previous research (Castillo, et al., 2011) (Della Vedova & al., 2018). These features were: count, tweet_id, tw_timestamp, parent_tweet_id, user_id, lang, likes, retweets, sentiment1. The feature "sentiment1" corresponds to sentiment score computed on raw tweet message[7], and the feature "lang" corresponds to the language of the tweet. This characteristic was extracted from multiple tweet information contained in the feature "text_html" using regular expression operations in Python (also called "Regex"). Each tweet was labelled with a two-letters code in a new field of the data frame (i.e. English tweets are labelled "en", Spanish tweets are "es", etc). This simplified the analysis of tweets per language and allows to filter on English claims for the project.

[5] Further details in Configuration Manual, section 4.1.1 Scrapped Twitter Data
[6] A response was returned for #coronavirus only and nothing for #covid. Therefore, three json files resulted from two scraping waves.
[7] Further details in section 5.3.1 Data Pre-processing Effect on Tweets Polarity

Three data frames were merged into one for a brief exploration of the scrapped tweets with Tableau to check for languages and duplicates. The quantity of English tweets was deemed sufficient to proceed with the dataset. From the 11,319 tweets scraped, 10,799 tweets remained after the removal of duplicated (based on tweet_id) and then 6,660 English tweets were retained to make the dataset.

3.3.1.3 Pre-processing

The cleaning process executed takes in account the English language and semantic specificities to remove stop words and lemmatize tokens (words)[8]. It also considers characteristics of language written on social media (informal, slang, emoticons, etc). The quality of text pre-processing is key to obtain clean data and extract meaningful information. A Python library dedicated to pre-process tweets was used in the first place to remove URL, hashtags, reserved words (RT for retweets), emojis and smileys. However, data obtained was not deemed cleaned enough. Therefore, the following key steps were carried out in addition:

1. Remove URL and punctuations: characters such as ? ! “ ” - _ etc.
2. Convert text to lowercase
3. Remove digits and words containing digits: this is ideal to remove parts of URL combining text characters and digits
4. Remove extra spaces and words of 1-character length
5. Remove stop words of the English language. Frequent words not adding values to the meaning of documents were discarded (i.e. “a”, “and”, “but”, “for”, “the”, etc.)
6. Tokenize words: the aim is to split text into individual words or expressions (sentences), which are called “tokens”. In this report, a token refers to a word. Strings of text were transformed to list of tokens.
7. Lemmatize words: the aim is to convert words into their base form, considering the context and position tag. Lemmatization is the preferred method to stemming, the latter is indeed simpler but with less performing results as the words suffix are truncated. A simple example illustrates the two techniques and their differences in Table 1, it supports the decision to lemmatize.

Table 1: Examples of stemming and lemmatization

Token	Suffix	Token Description	Stem	Lemma
Studies	-es	Third person, singular, present tense of verb study	Studi	Study
Studying	-ing	Gerund of the verb	Study	Study

3.3.2 News Articles Data

3.3.2.1 Data Collection

The second dataset consists of official news information scrapped from The Guardian website. This newspaper was chosen as being a reliable source of information in English language.

[8] Detailed step by step in the Configuration Manual, section 4.3 Pre-processing. Process is suitable for English tweets only.

Articles were scraped at several points in time between 21st June and 9th July 2020, using Python script. News headlines listed on the dedicated webpage to coronavirus were targeted in order to obtain the entire content of articles, which relate a large quantity of information regarding Covid-19. Scraped articles were saved into a csv file combining all articles of a particular extraction date. Files were appended to form a single data frame from collected data (Fig.4).

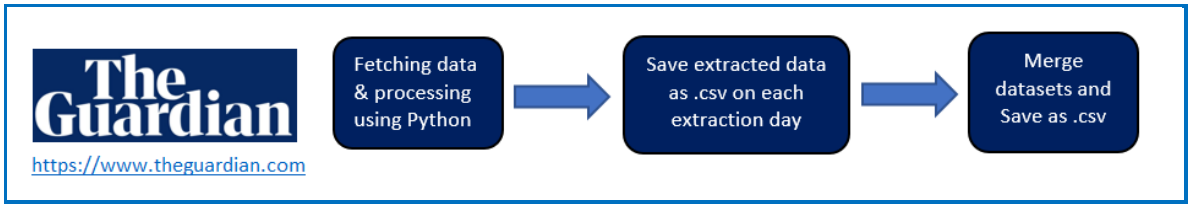

Figure 4: Collect news articles from TheGuardian.com using Python

3.3.2.2 Pre-processing

News articles may remain listed on the webpage for several consecutive days; therefore they can be duplicated in scraped data. Text content extracted in its raw form with special and unwanted characters (spaces, etc). constituted a proxy to a unique identifier for each article. Duplicates were identified among 227 articles in their non-processed form and dropped from the data frame before starting the cleaning process on the remaining 158 unique articles.

Data was cleaned using a similar methodology as Twitter data in order to implement a consistent approach to allow for stance comparison. Steps adapted to the text were carried out successively using manually build instructions to ensure the process followed was under control: (1) Remove strange characters, (2) Expand contractions (for instance, "wouldn't" was transformed to "would not"), (3) Lowercase the text, (4) Remove digits and words containing digits, (5) Remove punctuation, (6) Remove extra spaces, (7) Remove stop words, (8) Tokenize and (9) Lemmatize words.

Required output from pre-processing tweets and articles is lemmatized text (**Objective3** completed), which will be the basis for exploratory analysis and modelling data. Sample checks were done to assess pre-processed data and validate the quality. After a thorough and continuously reviewed cleaning method, tweets and articles still showed very few inconsistent words in Exploratory Analysis. To the best of the candidate's efforts and knowledge, data quality standards were considered sufficient to proceed.

4 Exploratory Analysis

Using RStudio and Python, an exploratory analysis was performed to gain insights on text features. This is preliminary work to modelling techniques for topic extraction and sentiment analysis.

4.1 Word Frequency and Association

4.1.1 Word Frequencies

Word clouds Fig.5 and Fig.6 depict a maximum of 50 words with frequency above 100. Frequencies indicate that news media used the term “covid” (302 times) to describe the virus as opposed to “coronavirus” that is essentially used on social media (1,408 times, excluding hashtags removed at pre-processing). This usage unbalance between the two terms was already indicated by the number of tweets scraped with hashtags “covid” and “coronavirus” (section 3.3 Data Persistence). The 10 most frequent words in tweets evoke negative content (“virus”, “outbreak”, “spread”) whereas news words convey a more neutral content as can be expected from general information source (“people”, “work”, “government”, “time”, “theatre”, “health”)[9].

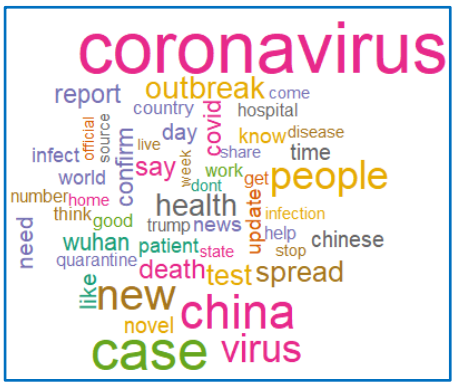

Figure 5: Frequent words in tweets

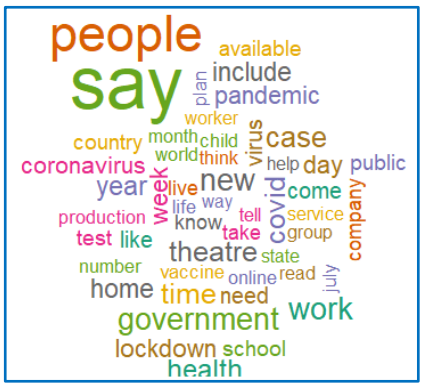

Figure 6: Frequent words in news

4.1.2 Bigrams Analysis

In NLP, a word is a gram and a pair of adjacent words is called bigrams. These pairs provide insights on text content by retrieving words frequently occurring together. For both tweets and news articles, the 20 bigrams with the highest occurrence frequency were queried using Python[10]. Pairs were interpreted, compared, and potential themes of discussions in the two corpuses identified are presented in Table 2.

4.1.2.1 Interpretation for tweets

The words "case" and "coronavirus" occur often and form several combinations with "new", "confirmed" or "death". The lexical field is rather pessimistic, it conveys negative meanings and a sentiment of fear from the population who generated these tweets data. The written content seems more personal, the main worry is related to the number of new cases and the outbreak. It must be noted that tweets were scraped between December 2019 and early April 2020, when the pandemic was spreading worldwide, and little was known about the virus.

[9] Refer to Configuration Manual section 5.2.1 Word clouds and frequencies
[10] Refer to Configuration Manual section 5.3. Bigrams analysis on tweets and news articles

4.1.2.2 Interpretation from news articles

The word associations uncover themes surrounding the official communication from authorities ("public health", "prime minister", "local authority", "Boris Johnson", "chief executive", "world health") regarding precaution measures to apply ("social distance", "face mask", "wear mask"). The UK Prime Minister Boris Johnson stands out from news articles as an individual given that the data source is a British newspaper. It must be noted that news articles were scraped from the end of June until the start of July 2020, a different timeframe from tweets data and lockdowns had ended in most of the countries following the first pandemic wave.

Table 2: Themes inferred from bigrams analysis

Twitter	The Guardian
Development of the pandemic crisis with comments on the outbreak, positive testing, the number of new cases, death toll.	- Promoting precaution measures, - Economy with the lockdown and closure of businesses such as pubs and restaurants.
Citation of official sources of information such as public health, health official, world health organisation.	Development of the pandemic crisis with updates on the number of new cases, there is no mention of the death toll here (vs. tweets).
The outbreak on the cruise ship is mentioned. It refers to the contamination of a very high portion of passengers and their isolation on the Diamond Princess boat until it docked on the Japanese coast.	The segments of population particularly affected: elder people in care home who are more vulnerable to developing critical symptoms and die; and young people affected by home schooling, the absence of socialization or the lack of opportunity to join the job market.
Hong Kong is cited; it could be for both the way the government handled the crisis and the impact it had on the demonstrations and rising of the population against the authorities.	Information on international stage and the outbreak of cases in New York.

4.2 Emotions Analysis

The National Research Council Canada (NRC) Word-Emotion Lexicon is used to identify emotional tone towards Covid-19 in the two datasets. It calculates the presence of eight emotions and two sentiments (positive, negative) and their corresponding valence in large units of texts (Mohammad & Turney, 2010). The focus is on exploring emotions before modelling the sentiment polarity as presented in section 5.2 of this report.

The distribution of emotions from 6,660 tweets vs. 158 news articles is represented with frequency scores as percentage of the corpus to allow for a comparison between tweets in Fig. 7 for tweets and news Fig. 8. The top 4 emotions in tweets are, by order of importance: fear (19.51%), trust (17.82%), anticipation (17.07%) and sadness (13.77%). They are the same in news corpus but in a different order with trust first (22.52%) and fear (15.91%) in third position. These findings are in line with the research from Aslam et al. where fear, trust, anticipation, sadness, and anger were the main emotions evoked by the news headlines (Aslam, et al., 2020).

Negative emotions (anger, disgust, fear, sadness) are more prevalent in the tweets (49.72% vs. 42.28% in news). News evoke more positive emotions (joy and trust) than tweets (33.28% against 25.38% in tweets). The two neutral emotions of anticipation and surprise represent the same share in both datasets (respectively 24.44% for tweets and 24.91% for news)[11].

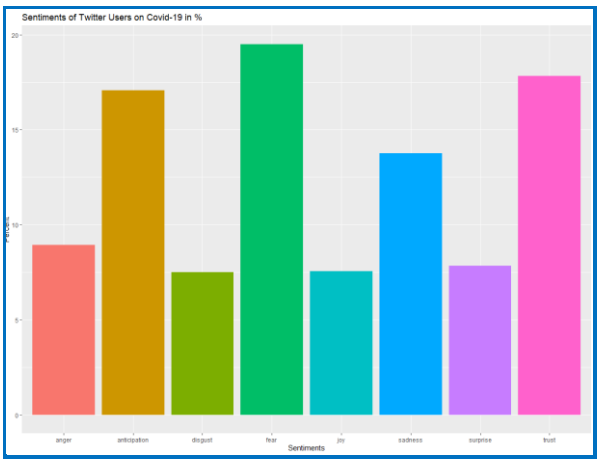

Figure 7: Emotions in tweets

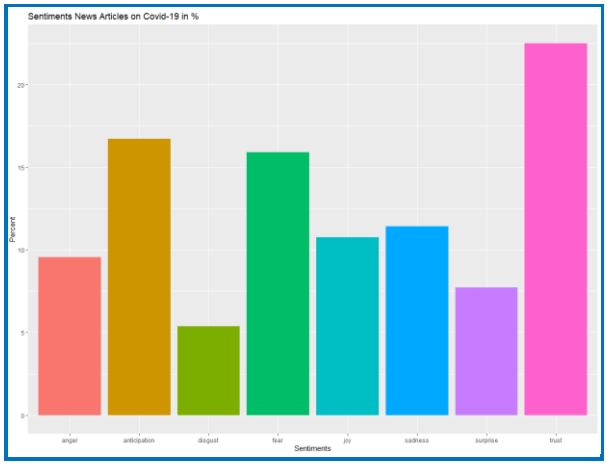

Figure 8: Emotions in newspaper

In overall, based on the eight emotions scores computed, we can conclude that news articles written by journalists and published by news media tend to convey more positive text content than short posts written on Twitter platform by the public. From this emotion detection analysis, we can see the population is more negative about Covid-19 when sharing posts and information informally on the platform. This exploratory analysis completes **Objective4**.

5 Implementation, Evaluation and Results of Topic Modelling and Sentiment Analysis

5.1 Models and Evaluation metrics

Implementation consists in two unsupervised modelling techniques to analyse the text-content feature and involves: (1) topic modelling with Latent Dirichlet Allocation model to gain insights on sub-topics related to coronavirus being discussed in media; and (2) a lexicon-based sentiment analysis for investigating text pre-processing impact on scores computations, and comparing sentiments evoked in news and tweets. LDA model will be assessed using coherence score. Sentiment analysis results will be evaluated with confusion matrix, accuracy (a), precision (b), recall (c), as well as Student's t-test on dependent and independent samples.

(a) Accuracy = (TP + TN) / (TP + TN + FP + FN)

(b) Precision = TP / (TP + FP)

(c) Recall = TP / (TP + FN)

[11] Refer to Configuration Manual section 5.4. Emotions Detection

5.2 Latent Dirichlet Allocation Topic Modelling

Latent Dirichlet Allocation algorithm was implemented to investigate and compare themes discussed in tweets and news. The optimum *k* number of topics was selected based on coherence scores of models tested, the aim is to select the one with the highest score and for the most number of topics (Fola, 2019). For news articles, the best coherence metrics of 0.475 was obtained with *k=5*. For tweets, 10 topics were extracted with the model that generated a coherence of 0.381. Using Python language and JupyterLab, interactive visualization graphs were built to present topic modeling results to stakeholders[12].

A list of 10 most relevant keywords was extracted for each principal component *k* in order to infer themes of discussion on media. To limit bias for this interpretation task, topic labels were assigned after the contribution of a panel of 5 people for assessment. In both corpuses, the recurrent topic relates to infection cases, updates and development of the pandemic ("new", "death", "case", "week", "people", "test", "confirm", "infection", "outbreak", "spread", "positive"). Based on the panel's feedback, news topics appeared as easier to infer with list of words from consistent vocabulary fields. The other news topics are varied (environmental impact of Covid with increased plastic use, cultural entertainment, vaccine and lockdown impact on the economy) among the only 5 components extracted. These findings are different from what was seen in crisis informatics research where topics covered by official sources relate mainly to instructions on how to behave during emergencies (Reuter, et al., 2018) that was seen in the current project as a recurrent topic in tweets (appearing twice). On the opposite, the exploratory analysis with bigrams (section 4.1.2) showed that word pairs from news suggested prevention measures such as "wear mask" and "social distance".

On the other hand, topics evoked in tweets were more difficult to infer. The model returned several overlapping components depicted on the interactive graph[13]; this is a sign of similarity between topics extracted from tweets. Keywords were sometimes difficult to relate to each other without the influence of specific words (for instance, the topic inferred from tweets_topic_1 is the outbreak on Diamond Princess cruise ship: i.e. "follow", "vaccine", "cancel", "*ship*", "supply", "crisis", "*cruise*"). Moreover, other keywords showed outliers from the Covid-19 overall topic (i.e. from tweets_topic_9 "*spy*", "today", "ask", "question", "hope", "*video*", "possible")[14]. Inferred topics relate mostly to government responses to Covid-19 with prevention measures, updates on the pandemic and search for a vaccine (news_topic indexes 0 and 3, tweets_topic indexes 2, 5, 7). These themes are evoked by authorities in the news or on social media according to literature (Liao, et al., 2020) whereas findings of the current analysis show they are discussed by the public on social media. Topics from tweets evoke the idea of mutual assistance and solidarity during the pandemic outbreak ("information", "protect", "prevention", "stay", "home", "wait", "help"). This insight confirms findings from the literature review on crisis informatics, where a reaction on social media when facing

[12] Detailed results in Configuration Manual section 6.1. LDA Topic Modelling
[13] Refer to Fig.135 in Configuration Manual, section 6.1.1 Topics Extraction from Tweets
[14] Refer to Table 5, in Configuration Manual, section 6.1.3. Comparison

uncertainty is a larger amount of collaboration (Valecha, et al., 2013). In overall, topics on Covid-19 expressed in tweets convey more negative semantic and reactions ("death", "symptom", "risk", “fear”) than in news data, even if there is a trust in the authorities responses (tweets_topic_8: “pandemic”, “good”, “hand”, “trump”). LDA topic modelling completes **Objective5**.

5.3 Sentiment Analysis

Polarity scores for each tweets and news articles were computed using Python TextBlob lexicon to complete **Objective6** with sentiment classification. It is an unsupervised technique to predict the sentiment of text documents by analysing words position, associations (bigrams, trigrams etc.), context, and part-of-speech elements. Sentiment is computed by associating tokens from documents with positive and negative polar words from the lexicon dictionary.

5.3.1 Data Pre-processing Effect on Tweets Polarity

At the stage of tweets data collection, the sentiment score was computed with TextBlob on the raw ‘text’ feature scraped from Twitter and labelled ‘sentiment1’ in the dataset. After pre-processing tweets, TextBlob was used again to compute a new polarity score (feature named ‘sentiment2’ in the data). Polarity scores obtained were categorised into three classes of positive (score > 0.3), negative (< -0.3) and neutral sentiments (score > -0.3 and <0.3). The same bounds were applied to categorise ‘sentiment1’ feature. This feature served as actual input to the confusion matrix assessing the prediction model accuracy. The multi-class confusion matrix presented evaluates the performance of sentiment analysis computed after an extensive data pre-processing task to clean tweet posts.

5.3.2 Prediction Results

A prediction accuracy score of 86.68% was obtained, with 5,773 true positives presented in Fig.9 (sum of cells A + E + I highlighted in blue). The accuracy metric shows how good a classification model is at predicting the correct category (Fig.10). The high percentage achieved confirms tweets sentiment scores computed before and after text pre-processing fall into the same class in 86.68% of 6,660 instances. Precision measures the proportion of instances correctly predicted out of all instances predicted by the model. Neutral tweets have a high precision of 0.96, this means if a result is predicted as belonging to the neutral class, it is 96% sure this class prediction is correct.

		Actual Sentiment Class		
		Negative (-1)	Neutral (0)	Positive (1)
Predicted	Negative (-1)	133 A	38 B	3 C
	Neutral (0)	237 D	5354 E	415 F
	Positive (1)	1 G	193 H	286 I

Figure 9: Sentiment Analysis Confusion Matrix

Accuracy Score : 0.8668168168168168
Report :

	precision	recall	f1-score	support
-1	0.36	0.76	0.49	174
0	0.96	0.89	0.92	6006
1	0.41	0.60	0.48	480
accuracy			0.87	6660
macro avg	0.57	0.75	0.63	6660
weighted avg	0.90	0.87	0.88	6660

Figure 10: Evaluation of Sentiment Model

Recall, also called sensitivity or true positive rate, measures what the model predicted correctly to what actual labels are. High recalls are obtained for the neutral class (0.89) and negative class (0.76), it is lower for the positive class (0.60). High recalls tend to lead to a higher number of false positive measurements and a lower accuracy (NillsF, 2020). The model selected shows high recalls together with a high accuracy.

In overall, sentiments towards Covid-19 in tweets tend to be neutral. Sample checks on data were performed to highlight the reasons of such results. Sentiment analysis present several limitations: (1) text can contain multiple sentiment by combining both positive and negative polarity in a same sentence or document (the longer the document, the more sentiments may be evoked); (2) a piece of text may not convey sentiment at all and be neutral (i.e. "The virus continues spreading" returns a polarity of 0.00; and "The virus spreads quickly worldwide" returns a score of 0.33). These sample sentences in the context of Covid-19 express negative information (when interpreted by humans) with the virus propagation context that is not captured by the sentiment-based lexicon; (3) the removal of stop words impacts sentiments expressed and therefore polarity scores. For instance, the removal of modifiers such as "very" reduces the negative polarity in the example "The situation is very serious" that scores -0.43 whereas "The situation is serious" returns a weaker negative polarity of -0.33. Pre-processed tweets show 90.18% of them are neutral whereas, whereas before text pre-processing, 83.86% were categorised as neutral.

To compare the same population of tweets before and after the pre-processing, and statistically measure the impact on sentiment scores computation, a Student t-test on dependent samples is calculated using Python (paired t-test). It is assumed that both samples come from normally distributed populations with equal variances. The hypotheses tested are as follows:

H_0: $\mu_d = 0$ (the true means difference μ_d is equal to zero)
H_1: $\mu_d \neq 0$ (μ_d is not equal to zero)

With a level of significance $\alpha = 0.05$, the t-statistic obtained is 4.431 for this 2-tail test. It is greater than the critical value of 1.96 from t-distribution table[15]. The null hypothesis is rejected; therefore, text pre-processing impacts the computation of polarity scores. The hypotheses testing is not about sentiment scores but about the processes producing the data and determine which one is more consistent (Solutions, 2020).

We can conclude the tedious and complex data pre-processing tasks modified the text feature and polarity scores assigned to tweets (**Objective7** completed). This is also translated in the change of distribution of tweets across the positive, negative and neutral categories as depicted in Fig. 11 (before) and Fig. 12 (after pre-processing). Tweets are plotted by sentiment category on the x-axis, and with their corresponding score (float numbers from -1 to 1) on the y-axis.

[15]https://www.gradecalculator.tech/wp-content/uploads/2019/12/T-Table-T-Distribution-Critical-Values-Table-Large.jpg

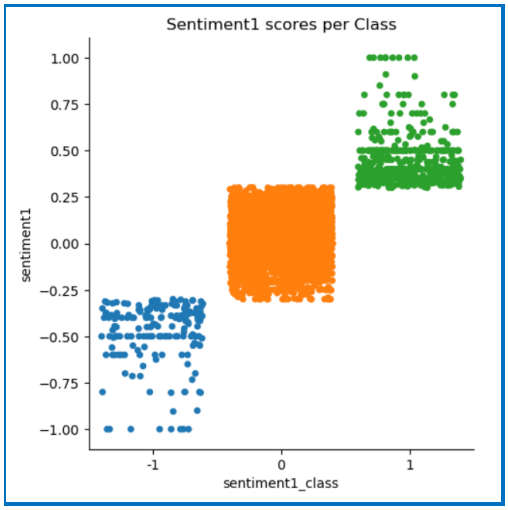

Figure 11: Sentiment1 distribution

Figure 12: Sentiment2 distribution

5.3.3 Tweets and News Comparison

Polarity scores from news articles range from -0.122 to 0.219, they were all classified into the neutral category (bound from -0.3 to 0.3 scores). A t-test on two independent samples (unpaired test) is done to assess the difference of sentiments evoked in tweets and news by testing the average polarity from tweets (μ1) to average sentiment from news (μ2). Populations have unequal variances; samples are of different sizes and the news articles distribution of scores does not follow a normal distribution. Therefore, the non-parametric Welch's t-test is used. Hypotheses stated are as follows:

H_0: μ1 = μ2 (the two populations means are equal)
H_1: μ1 ≠ μ2 (the two populations means are different)

With a significance level α = 0.05, the t-statistics computed using Python is 3.113. For this 2-tail test with a critical value of 1.96 from t-distribution table, we reject the null hypothesis. Therefore, the average polarity from tweets (μ1) is not equal to the average sentiment from news (μ2) [16]. **Objective8** is completed.

6 Discussion

During the Covid-19 crisis, we have seen opposite reactions from what Helsloot & Ruitenberg (2004) described in the context of crisis informatics. Spread of rumours and fake news online essentially (or a new phenomenon via messaging apps like WhatsApp), irrational reactions to the pandemic crisis, where people panicked and stocked up food and other items creating shortage of supplies. We can assume the lockdowns implemented in several countries created an over-use and reliance on social media, to relay and amplify news. Moreover, they reinforced worries regarding an unknown virus and uncertainties towards the short-term future.

[16] Sentiment scores mean for news articles (0.060) is slightly higher than the mean calculated for tweets sentiment scores (0.041).

In the context of Covid-19, it is assumed from findings that themes discussed, in both social media and news media, are informative on the virus, provide regular updates on the propagation and infection cases, prevention measures and advice to adapt lifestyle during lockdown. This is aligned with previous research on Ebola outbreak where conversation categories were for: giving information, news update and preparedness (Wong, et al., 2017). News topics modelled evoke more varied themes than tweets by covering entertainment, environment and economic plans. On the opposite, tweets topics are informative and negative in overall, they convey fear where affected people blame other individuals or government, and express worry about Covid-19 pandemic (Liao, et al., 2020). But tweets express also ideas of solidarity among the population during the crisis. These positive empathic reactions were seen in previous researches in the field of crisis informatics (Wong, et al., 2017).

Results observed from the sentiment analysis are not aligned with findings from reviewed literature where sentiment scores from news headlines were "severely weighted towards the negative side" (Aslam, et al., 2020). Research project shows that 100% news articles studied evoke neutral sentiments, as it can be expected from information sources. However, 90.18% of pre-processed tweets fall into the neutral category and reveal little polar sentiments (7.21% of positive tweets and 2.61% are negative). The distribution across three sentiment classes is unbalanced with a clear prevalence of neutral stances in tweets. This reflects a significant limitation in extracting valuable insights on people's mental health and feelings towards Covid-19 given that the majority of tweets expresses a neutral stance. However, it demonstrates the importance of the neutral class that should not be ignored. It can even improve the overall accuracy, as reviewed in the literature where polarity problems are best handled with three classes (Koppel & Schler, 2006) (Taboada, et al., 2011) and demonstrated in an additional approach of categorizing sentiments into classes with alternative cutoff points[17].When detecting emotions, it was noted that two neutral feelings of anticipation and surprise represented the same share in both datasets. Negative emotions (anger, disgust, fear, sadness) were more predominant in both media, as revealed by previous research on emotions evoked by headlines covering Covid-19 news (Aslam, et al., 2020). It must be noted that two sentiment lexicons were used for the emotion detection (NRC) and sentiment analysis (TextBlob lexicon), we can assume words are mapped to sentiments in different ways in each lexicon and generate different results.

Sentiment analysis on Twitter data presents limitations for analysing human and emotional patterns as presented in section 5.3.2 of this report. Moreover, the quality and reliability issues associated with this type of data have been discussed by scholars. A major contribution is the work from Danah Boyd and Kate Crawford in 2012 where they highlight the fact that not all the population use Twitter social media. In addition, not all Twitter materials are made public. The company provides a sample selection of tweets via the API, therefore scraped output is biased (Boyd & Crawford, 2012). Results of tweets sentiment analysis should be taken as an indication for interpretation of behaviours, sentiments and topics discussed on the platform.

[17] This second approach was not retained for the technical report. The accuracy score obtained was 80.12%. It is presented in the Configuration Manual, section 7.2. Extra Sentiment Analysis on Tweets.

Findings should not be considered as representative of the overall population as only a sample – not representative – of the population use Twitter and is vocal on the network (Boyd & Crawford, 2012).

7 Conclusion and Future Work

The research project focused on analysing social media and news media content using NLP techniques to provide insights on information discussed and emotional reactions regarding Covid-19 pandemic. Objectives 1 to 4 were completed with the literature review, data collection, data pre-processing and exploratory analysis. To solve the RQ, the implementation of LDA topic modelling met with Objective 5 in section 5.2 Emotions detection (section 4.2) and sentiment analysis implemented in section 5.3 met Objective 6. Objectives 7 and 8 were completed with the statistical analysis of sentiment scores before and after prep-processing tweets to answer SUBRQ1 (section 5.3.2), and between news and tweets scores to answer SUBRQ2 (section 5.3.3). Key findings confirm topics discussed in media during Covid-19 relate to information giving, update on the pandemic and prevention measures the population should adopt during such pandemic. Media content analysed with NLP techniques indicate Twitter convey more negative themes and sentiments than expressed in The Guardian news articles. Hidden patterns show the prevalence of the neutral class in sentiment analysis while Covid-19 crisis context and its recurrent topic dominance in the media during months seemed to have created a negative and fearful environment, based on findings from emotion recognition.

The research significantly contributes to enhance knowledge in the domain of crisis informatics by analysing the use of media and emotional reactions regarding Covid-19 pandemic. This crisis brought a new context for experiments and observations. A minor contribution of the research is highlighting the importance of the neutral class in sentiment analysis as it was done in scarce literature reviewed in the past. Leaning outcomes for the candidate are a deeper knowledge of coding with Python, the domain of crisis informatics, semantic analysis and in particular the natural language processing techniques. Further personal learning directions include artificial intelligence, neural networks and the recent field of Natural Language Generation with GPT3.

Challenges in terms of time and coding technical constraints brought some limitations to carry out the project without implementation flaws. Text data cleaning is a complex task that is acute when language used on social media is informal, contains abbreviations, slang and new words as in the context of an emerging virus. The project research could be complemented by sentiment analysis using VADER lexicon (Valence Aware Dictionary and Sentiment Reasoner), better suited for social media content but not attuned for newspaper text corpus[18]. Sentiments and emotions reactions to Covid-19 evoked in tweets and news could be tested by measuring cosine similarity to assess how similar documents from datasets are irrespective of

[18] Additional implementation with examples of sentiment scores computation with VADER available in Configuration Manual, section 7.2 Extra Sentiment Analysis on Tweets

their size similarities. This approach of stance-based similarity detection could be extended to assess veracity of tweets against news sources. Proposing an unsupervised method for veracity assessment of information, with in mind the willingness to bring innovation to the field of automatic fake news detections. The identification of false information and influencing content is critical nowadays due to the consequences on target audience, for making inaccurate or altered decisions, or manipulating their opinion.

Analysing the trustworthiness of text content is recommendation for future work. This idea was inspired from previous research from (Castillo, et al., 2011) on Twitter data and gap in the existing knowledge and business applications. More recently, on May 27th, 2020 Twitter implemented for the first time a fact-check label to a Tweet from the U.S. President Donald Trump. As reported in the news, the content of his post was described as "unsubstantiated" and this label acted as a warning for readers. The platform is moving towards assessing the reliability of information shared, in May 2020 it introduced a new policy on misleading amid the coronavirus pandemic (BBC News, 2020). From March 2020, Twitter announced to broaden their policy guidance to address content shared on their platform that went directly against the information from official or public health authorities, both local and global. This would take the form of potential labels and warning messages to advise readers to look for additional information or clarifications of the claims made (Roth & Pickles, 2020).

Acknowledgments

I would like to thank work colleagues for their support during the research project and part-time Masters course. Especially the four members who agreed to contribute to this project as the panel of assessors to infer topics modelled.

References

Aslam, F. et al., 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. *Humanities and Social Sciences Communications.*

BBC News, 2020. *Twitter tags Trump tweet with fact-checking warning.* [Online] Available at: https://www.bbc.com/news/technology-52815552

Beigi, G., Maciejewski, R., Hu, X. & Liu, H., 2016. An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment Analysis and Ontology Engineering,* pp. 313-340.

Blei, D. N. A. J. M., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research,* Volume 3(Issue 4-5), pp. 993-1022.

Bold, A., 2019. *Sentiment Analysis - The Lexicon Based Approach.* [Online] Available at: https://alphabold.com/sentiment-analysis-the-lexicon-based-approach/

Bondielli, A. & Marcelloni, F., 2019. A survey on fake news and rumour detection techniques. *Information Sciences,* Volume 497, p. 38–55. doi: 10.1016/j.ins.2019.05.035..

Boyd, D. & Crawford, K., 2012. Critical Questions For Big Data. *Information, Communication & Society,* 15(5), pp. 662-679.

Castillo, C., Mendoza, M. & Poblete, P., 2011. *Information credibility on Twitter.* pp. 675-684, s.n.

Chen, Y.-C., Liu, Z.-Y. & Kao, H.-Y., 2017. *Ikm at semeval-2017 task 8: convolutional neural networks for stance detection and rumor verification.* s.l., SemEval-2017, pp. 465-469.

Dattu, B. S. & Gore, D. V., 2015. A Survey on Sentiment Analysis on Twitter Data. *International Journal of Computer Science and Information Technologies,* pp. 5358-5362.
Davidson, H., 2020. *First Covid-19 case happened in November, China government records show - report.* [Online] Available at: https://www.theguardian.com/world/2020/mar/13/first-covid-19-case-happened-in-november-china-government-records-show-report
Deerwester, S. D. S. T. F. G. W. L. T. K. H. R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science,* pp. vol. 41(issue. 6), pp. 391.
Della Vedova, M. & al., e., 2018. *Automatic Online Fake News Detection Combining Content and Social Signals.* s.l., s.n., p. p. 272. doi: 10.23919/FRUCT.2018.8468301.
Fola, S., 2019. *NLP with LDA: Analyzing Topics in the Enron Email dataset.* [Online] Available at: https://medium.com/datadriveninvestor/nlp-with-lda-analyzing-topics-in-the-enron-email-dataset-20326b7ae36f
Fung, I. C.-H. et al., 2016. Social Media's Initial Reaction to Information and Misinformation on Ebola, August 2014: Facts and Rumors. *Public Health Reports,* pp. 461-73.
George, L. E. & Birla, L., 2018. *A Study of Topic Modeling Methods.* Madurai, India, India, IEEE, pp. 09-113.
Helsloot, I. & Ruitenberg, A., 2004. Citizen response to disasters: A survey of literature and some practical implications. *Journal of Contingencies and Crisis Management,* p. 98–111.
Jin, F., Wang, W., Zhao, L. & Dougherty, E., 2014. *Misinformation Propagation in the Age of Twitter.* 47(12), pp. 90–94. doi: 10.1109/MC.2014.361., IEEE Computer Society.
Kaewkitipong, L., Chen, C. & Ractham, P., 2012. *Lessons learned from the use of social media in combating a crisis: A case study of 2011 Thailand flooding disaster. In.* Orlando, USA, s.n., pp. 1-17.
Kapadia, S., 2019. *Evaluate Topic Models: Latent Dirichlet Allocation (LDA).* [Online] Available at: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
Kaufhold, M.-A. & Reuter, C., 2016. The self-organization of digital volunteers across social media: The case of the 2013 European Floods in Germany. *Journal of Homeland Security and Emergency Management,* p. 137–166.
Koppel, M. & Schler, J., 2006. The Importance of Neutral Examples for Learning Sentiment. *Computational Intelligence,* pp. 100-116.
Kucuk, D. & Can, F., 2020. Stance Detection: A Survey. *ACM Computing Surveys,* pp. Vol. 53 Issue 1, p1-37.
Lane, H., Howard, C. & Hapke, H., 2019. *Natural Language Processing in Action.* New York, USA: Manning.
Liao, Q. et al., 2020. Public Engagement and Government Responsiveness in the Communications About COVID-19 During the Early Epidemic Stage in China: Infodemiology Study on Social Media Data. *Journal of Medical Internet Research,* p. 22.
Ma, J. et al., 2016. *Detecting rumors from microblogs with recurrent neural networks.* s.l., s.n.
Mittal, A. & Patidar, S., 2019. Sentiment Analysis on Twitter Data: A Survey. *Computer and Communications Management.,* pp. 91-95.
Mohammad, S. & Turney, P. D., 2010. *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,* Ottawa, Ontario, Canada: National Research Council Canada.
NillsF, 2020. *Confusion matrix, accuracy, recall, precision, false positive rate and F-scores explained.* [Online] Available at: https://nillsf.com/index.php/2020/05/23/confusion-matrix-accuracy-recall-precision-false-positive-rate-and-f-scores-explained/

Odlum, M. & Yoon, S., 2015. What can we learn about the Ebola outbreak from tweets?. *American Journal of Infection Control,* pp. 563-71.
Reuter, C., Hughes, A. L. & Kaufhold, M.-A., 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human-Computer Interaction,* pp. 280-294.
Röder, M., Both, A. & Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures. In: *Web Search and Data Mining.* New York, NY, USA: ACM, pp. 399-408.
Roth, Y. & Pickles, N., 2020. *Updating our Approach to Misleading Information.* [Online] Available at: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
Shu, K. et al., 2017. *Fake news detection on social media: a data mining perspective.* s.l., s.n., p. 22–36.
Solutions, S., 2020. *Paired Sample T-Test.* [Online] Available at: https://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/
Taboada, M., Brooke, J., Tofiloski, M. & Voll, K. D., 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics,* pp. 267-307.
Tang, L., Bie, B., Park, S.-E. & Zhi, D., 2018. Social media and outbreaks of emerging infectious diseases: A systematic review of literature. *American Journal of Infection Control,* pp. 962-972.
Traylor, T., Straub, J., Gurmeet & Snell, N., 2019. *Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator.* p. 445. doi: 10.1109/ICOSC.2019.8665593, Semantic Computing (ICSC), 2019 IEEE 13th International Conference on, p. 445. doi:.
Valecha, R., Oh, O. & Rao, R., 2013. *An exploration of collaboration over time in collective crisis response during the Haiti 2010 Earthquake.* Milan, Italy, Association for Information Systems, pp. 1-10.
Vayanskya, I. & Kumar, S. A., 2020. A review of topic modeling methods. *Information Systems,* p. Volume 94.
Volkova, S., Shaffer, K., Jang, J. & Hodas, N., 2017. *Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter.* s.l., s.n., pp. 647-653.
Vorhies, W., 2016. *CRISP-DM – a Standard Methodology to Ensure a Good Outcome.* [Online]
Available at: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome
Vosoughi, S., Mohsenvand, M. & Roy, D., 2017. *Rumor gauge: predicting the veracity of rumors on twitter.* s.l.:ACM Trans. Knowl. Discov. Data (TKDD) 11 (4).
Wang, W., 2017. *"Liar, liar pants on fire": a new benchmark dataset for fake news detection.* s.l., s.n., pp. 422-426.
WHO, 2020. *1st WHO Infodemiology Conference.* [Online] Available at: https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference
Wong, R., Harris, J. K., Staub, M. & Bernhardt, J. M., 2017. Local Health Departments Tweeting About Ebola: Characteristics and Messaging. *Journal of Public Health Management and Practice,* pp. 16-24.
Zeng-Treitler, Q. et al., 2008. Estimating consumer familiarity with health terminology: a context-based approach. *Journal of the American Medical Informatics Association,* pp. 349-56.
Zhang, Y., Suhaimi, N., Azghandi, R. & Joseph, M., 2020. *Understanding the Use of Crisis Informatics Technlogy among Older Adults.* Honolulu, HI, USA, s.n.

Zhao, W. et al., 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, p. 1–10.