Department of Computer Science

MSc Data Science and Analytics

Academic Year 2021-2022

EMOTION ANALYSIS OF NEWS ARTICLES AND SUBSEQUENT CORRELATION WITH COVID CASES IN 2021

Shivangi Dubey
2039934

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397
Fax: +44 (0) 1895 251686

ABSTRACT

In light of the recent pandemic, the sentiments and emotions of the public have been quite influential and impacted the norm at every level, global, community and individual. This has caused us to focus on maintaining calm and led us to focus on everything that is a factor in balancing this entropy throughout society. The entropy is on a significant scale shown by the emotions or sentiments prevailing at any given time, and one of the factors that control it is the medium of information dissemination, i.e. news. In this research account, we have attempted to quantify the significance of emotions by a correlational study between the emotions of the news texts published and the covid cases on the same and the following days. It employs the algorithm of emotion dynamics which is officially suggested on the website of NRC lexicon (EmoLex) for emotion analysis.

ACKNOWLEDGEMENT

I want to express my most profound appreciation to my advisor, Dr David Bell, for his intangible support, knowledge and expertise shared with me. His weekly calls and meetups, paired with timely support and feedback, made this thesis a possibility.
I would like to extend my sincere thanks to my cohort members and library assistants for their advice and suggestions.
Lastly, I'd like to mention my parents and friends for their emotional support and their belief in me. Their utmost belief in me has kept my spirits high throughout my dissertation
process.

I certify that the work presented in the dissertation is my own unless referenced

Signature.....Shivangi Dubey.....

Date.....13-09-2022.....

TOTAL NUMBER OF WORDS: 9571

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 *Introduction*

The recent pandemic has caused us to re-evaluate significant aspects of our lives on an individual and community level. The importance of public discourse has caused us to re-evaluate and strategise how content is delivered to the public. It is because the pandemic has already stressed our health system, and any more carelessness that might impact the mental health of the masses can wreak havoc in sensitive times where all kinds of information, regardless of whether it is reliable or not, can destroy the delicate balance of entropy within the community.
Emotions play a big role in sustaining the body's immune system and can affect the spread rate of infections. Thus a major focus on information dissemination must be dealt with.
The acknowledgement of emotions in the social strata has led the research community to investigate the causes and impacts of sentiments and emotions at an unbelievable pace. The keyword ‘covid 19’ yields research papers in the categories analysed below:

These categories appear to constitute the total publications under the ‘computer science’ umbrella when they were refined for results in the last three years. The Artificial Intelligence publications are 13.75% of the total.
To study the pandemic’s impact on community dynamics, data representing every form of public expression or the cause of any influence on this expression is bound to be analysed. The sentiment of this expression is a significant window into the kind of mass reaction it is capable of triggering. Hence, sentiment analysis is a convenient way of studying the most apparent impact and effectively reading between the lines of this data. At least the recent advances in this field have made it possible to do both, the latter to a limited extent. Some explore the sentiment (positive, negative and neutral) analysis of the text (Yu *et al.*, 2020), while others showcase the methods of identifying trends of the not-so-obvious emotions like anger, fear, sadness and joy. (Lwin *et al.*, 2020)

Recent research in sentiment analysis has been focused on social media data. Twitter and Facebook data have been exploited for various purposes with the help of sentiment analysis in natural language processing. From an attempt to gain insight into the public sentiments (Samuel *et al.*, 2020) to map the determinants of vaccine uptake (Baj-Rogowska, 2021), there have been umpteen efforts to study the available data from these social media websites to understand the situation better and encourage better public management via solutions devised after every study.

Another source of data publicly available is that of the News. The discourse of news has changed media from on-paper text to an email in our inboxes and has also multiplied in respect of the amount of information it entails and delivers
(Taj, Shaikh and Fatemah Meghji, 2019).
Research concerning news data includes the detection of fake news using sentiment analysis, for the most part. Fake news detection fundamentally improves sentiment analysis and polarity determination (Samonte, 2018). It is a crucial area of research since unverified media updates may cause the spread of misinformation and wreak havoc in sensitive times like these. Another area where news data has been studied using sentiment analysis is the determination of stock prices. (Agarwal, 2020)

However, limited research has been done on utilising the potential of news and the sentiment it delivers to study the reaction it may induce. A few attempts have been made to identify public sentiments concerning the pandemic (Liu *et al.*, 2020). Still, any evident correlation between the resulting emotions and an important metric that may confirm its impact has been lacking.

1.2 Research aim and objectives

The research aims to exploit a source of publicly available data, i.e. News articles,

- to extract any dominant emotions that might be evident from the text.
- After preprocessing, a textual analysis of the text is done to identify any trends in the same.
- An attempt to correlate the presence of specific emotions to the number of new covid patients in the subsequent days is made using Pearson R correlation. Since the impact may not be immediately visible, a slight delay is considered to allow time for the absorption of the news material and a consequent reaction.

CHAPTER 2: LITERATURE REVIEW

The chapter discusses the method employed to study the citations (citespace), the importance of public discourse, how it can trigger nervousness and anxiety amongst the public and states studies exploring the role of language in strategising information dissemination.
It later discusses the state-of-the-art techniques related to the concepts being used in the study. An introduction to natural language processing, proceeding with sentiment analysis, compares different approaches for the same and then dives into emotion analysis. The NRC lexicon and emotion dynamics have been explained to lay out a background for analysing news articles.

2.1 Citation Analysis

Citespace is a software platform used to detect and visualise emerging trends and transient patterns in scientific literature. (Chen, 2006)

The papers in the review were collected using the Web of Science platform (only for the last three years) using the keywords:
Covid
Sentiment
Lexicon
Emotion
Health
The following are the details of the total papers collected and the time taken for the software to run the cluster analysis.
Distinct Valid References: 50561 (97.9124%)
Distinct invalid references: 1078 (2.0876%)
Parsing time: 1 min 14 sec
Total Run time: 21 sec

The details of the cluster network are:
Nodes: 132
Links: 692
The visualisation has been included.
The clusters generated have specific keywords associated with them:

- The most prominent group has the keyword ‘natural language processing
- The second largest cluster has the keyword ‘media’ assigned.
- The third largest cluster’s keyword is called ‘real-time monitoring.
- The fourth largest cluster’s keyword is called ‘corona.

The keyword in the above clusters was chosen based on the most frequently used word in papers included for clustering.
The most cited papers generated were:

- The most cited paper was by (Samuel *et al.*, 2020). It had a total of 16 citations.
- The second most cited was a book by Devlin J, which could not be accessed.
- The third most cited paper was (Lwin *et al.*, 2020), with nine citations.
- The fifth most cited paper was (Abd-Alrazaq *et al.*, 2020), with six citations.

The bursts are papers that are pretty influential during a specific period. The report of these papers and the periods in which these papers were popular are:

Top 5 References with the Strongest Citation Bursts				
References	**Year**	**Strength**	**Begin**	**End**
Wang Y, 2016, P 2016 C EMPIRICAL M, V0, P606, DOI 10.18653/v1/D16-1058, DOI	**2016**	**2.32**	**2019**	**2020**
Poria S, 2016, KNOWL-BASED SYST, V108, P42, DOI 10.1016/j.knosys.2016.06.009, DOI	**2016**	**2.1**	**2019**	**2022**
Devlin J, 2018, P 2019 C N AM CHAPT, V0, P0	**2018**	**2.19**	**2020**	**2022**
Pontiki M, 2016, INT WORKSH SEM EV, V0, P19, DOI 10.18653/v1/S16-1002, DOI	**2016**	**2.19**	**2020**	**2022**
Cambria E, 2016, IEEE INTELL SYST, V31, P102, DOI 10.1109/MIS.2016.31, DOI	**2016**	**2.11**	**2020**	**2022**

The rest of the bursts are included in the appendix of the report.
The papers were studied and then reviewed in the literature to study state-of-the-art practices in the research field.

2.2 Background

Public Discourse

The pandemic of 2019 has prompted the government to take some inevitable measures to prioritise the safety of the people. These interventions, which are necessary for the absence of widespread vaccinations, can cut off social support networks, restrict access to services, and make people feel anxious and unsafe. (Finlay *et al.*, 2021)

Strategising crisis communications has attracted attention in recent years for a good reason. Fighting misinformation and eradicating uncertainty remain the main objectives of these studies, focusings on the critical role of language in keeping nerves calm during these dire and unprecedentedtimess. (Chepurnaya, 2021) (Xue *et al.*, 2020)

News remains the primary source of information for most even today. Newspapers were prevalent not long ago; however, media like websites, blogs and social media have taken their place. But written language has held its importance over the years and has maintained its role as the primary medium of information dissemination.
Thus, news articles on media websites or official government announcements percolate through to the deep ends of society and potentially directly affect or spread the word about the common public perception on a large scale.

Emotions and Physiological aspects of the body.

The link between an emotion felt and physiological changes that occur in the body is corroborated by several theories that have been proposed since the earliest of times.

Emotion is part and parcel of an individual's personality. Psychology deals with the concept of emotions and their inter-relatability with stimuli (events) via different theories of emotion. It is a composite of feeling, cognition, action, expression and physiology. There have been many proposed over time. Some deal with only specific fields like stress or coping; others try to encapsulate explanations for every emotional response. A summary of emotion theories is as follows:

- Charles Robert Darwin studied emotions and gave the idea that emotions are not just related to facial expressions, but there are a number of discrete emotions that don't affect the facial muscles. He studied Duchenne's work on manipulating facial expressions by applying electrical pulses to the facial muscles.
- Duchenne had proposed a theory that there are individual facial muscles that are responsible for each emotion.
- Darwin opposed Duchenne's view as he thought that not all emotions are linked with the facial muscles. He also has an idea that there might be a small number of emotions that are uniform across all cultures.
- He did a test in his home where he invited 20 guests and asked them to recognise the emotions of Duchenne's photographic images, each of which depicted a distinct emotion. Darwin selected some of the expressions of Duchenne to include in his Expressions. Darwin then went on and included some of these images in his book, Expressions. (Snyder *et al.*, 2010)
- Paul Ekman reviewed Darwin's book, Expressions, in which he discussed Darwin's five major contributions, reasons as to why he was able to make these deductions and some major questions that Darwin did not address.
- Darwin proposed emotions as discrete sets that were distinct from one another. Recent studies and research suggest that Darwin's deductions regarding emotions as separate distinctive sets were correct.
- Paul further described Darwin's contributions, with the first being that Darwin described variations in emotions rather than including them in a family of related emotions. Paul Ekman opposes the idea that hatred is an emotion and suggests that it is a psychological state rather than an emotional state and that hatred leads to anger.
- The second contribution of Darwin was that he regarded the face as the primary recogniser for emotions. Recent studies also suggest that the face is one of the most resourceful factors in recognising emotions.
- Paul Ekman also states that Darwin did not take the duration of the expression into consideration. Paul Ekman's research suggests that the contraction in facial muscles reaches a plateau within a few seconds. These expressions are called snapshot expressions. What follows after that are aggregate signals which comprise a series of expressions. These signals depict the intensity of the emotions rather than the emotion itself.
- Darwin's third deduction was that facial expressions are uniform throughout the planet. This deduction was made from his short experience of travelling around the world. Paul Ekman also noticed that this deduction was not in favour of his evolution theory because the uniformity of the emotions would support the contrary theory of Adam and Eve.
- Darwin also suggested that while facial expressions are uniform throughout the globe, gestures were, on the contrary, culture specific traditions.
- Darwin's fifth deduction in his book was the explanation of movements of facial muscles that depict an emotion. He gave an example that in the anger emotion, the upper lip is raised to expose the canine teeth that would imply a threat. It also depicted a preparation to strike on the prey.

- Darwin also claimed that these emotional expressions were not limited to humans and that there were also applied to animals. He rejected his professor, Sir Charles Bell's, theory that emotional expressions were only bestowed to men by God.
- Darwin's method to display the images of the facial expressions to people and then ask what emotion they depicted from each expression is still implemented to study the facial expressions. This method is known as a judgement study.
- Paul Ekman stated that Darwin did not invent a method to measure facial movements and that he did not specify the limits of each family of emotions. Paul Ekman noticed that we, till now, do know the number of variations of emotions, nor do we know the quantity of variations that are related to the difference in social experience. Paul also noticed that Darwin did not provide an explanation for falsified facial expressions that hid their genuine emotions. According to Paul, Darwin did provide an explanation of the role of facial expressions and emotions when a subject is being deceptive.(Ekman, 2009)

Darwin's Theory of Emotion

The theory proposes that every emotional expression is a part of survival. Emotion is influenced by feelings, thoughts, behaviour and psychological and physiological response to stimuli. Darwin said the feelings were temporary and involuntary.

Based on the priority of reactions, some early theories are:

James Lange Theory

Emotions are a result of physiological reactions to events. They are interpretations of experiences an individual has had. This theory has some evidence to its credit where a brain scan revealed some basic emotions elicit distinct patterns of activity in the brain's neural network. Another experiment proved that perception of the internal physical state affects how people experience emotions. Yet another study demonstrated that different facial expressions about other emotions caused physiological changes like the individual's heart rate and skin temperature.

But there are flaws to this theory and other theories of emotion, like the Cannon-Bard *Theory* of emotion, which proposes precisely the opposite of James Lange's theory but does relate physiological reactions to emotions.

Muscle paralysis or loss of sensation does not imply an inability to experience emotions. Also, electrical stimulation to the same part of the brain or similar physiological reactions causes different feelings. An Individual's mental state, cues in the environment and reactions to other people influence a person's emotional response.

The above was proven via experiments to disprove James-Lange and Cannon-Bard's theory.

(Cherry, 2020)

Lazarus Appraisal Theory

Lazarus's work focused on stress and emotions induced by coping with it. He proposed that stress consisted of harm, threat and challenges. It caused coping mechanisms that, in turn, affected appraisal, which could either be emotion-focused or problem-focused. Several notes from this theory include

- There is no distinction between male and female stressors, contrary to what is believed.
- Coping strategies can be stable and different depending on the stage of the

emotional encounters.
- Emotion reaction also depends on the outcome modality,i.e. whether the encounter affects the individual or the society.

(Strongman, 2003)

The relationship between the emotional and physiological body has quite a dearth of research behind it. It has been recently corroborated by a large number of papers that stress can induce a coping mechanism where the immune system of the body can get suppressed. One way can be due to the variability of perceptions of stress among individuals and other can be a chronic response of the body to exhibit immunosuppressive characteristics. In both cases it is pretty evident that it leads to mounting physiological stress, leaving it vulnerable to infections.(Amal I. Khalil, Rawan E.Nasr, and Rahaf E.Nahr, 2020)

Many theories explain stress in different contexts. Causes of stress are defined to be allostatic or pathogenic. In the allostatic case, the load accumulated due to stress and life events (Guidi *et al.*, 2021), while the pathogenic load is when a body is fighting an infection and expending its resources to meet the demands of an illness. In both these cases the problem only arises when the resources of the body fall short of the demand of the crisis.(Amirkhan, 2021)

Both these studies were done recently and have concluded the negative consequences of any kind of stress on the physiological aspects (weakening the immune system). Hence, stress regulation is an of utmost importance on an individual level. On a national level, the correct form of information dissemination to reduce anxiety seems to be one of the things that can be and should be dealt with sensitivity.

Natural Language Processing

The increase in online activities of people due to limited outdoor activities owing to the spread of covid has caused the availability of massive amounts of structured and unstructured data: blogs, Social Networks, e-commerce websites, news reports and platforms for opinion expressions. The variety of text data publicly available has catalysed the onset of research around the use of language and improved the cognisance of machines when it comes to interpreting contemporary language. (Park *et al.*, 2021) (Gupta and Joshi, 2018)
Some subtasks of Natural Language Processing are sentiment Analysis, Opinion Mining, Polarity detection, and Emotion Recognition. These are some significant areas of publication in the textual data processing.

Sentiment Analysis

Sentiment analysis is the first step of emotion detection and is very popular when analysing texts for applications like customer relationship management and recommendation systems. It involves categorising the sentiment of the text at hand based on the connotation of the words that comprise it.

Some papers use the terms like Opinion Mining and Sentiment Analysis interchangeably (Medhat, Hassan and Korashy, 2014). Others refer to Emotion detection as an Affective computing (Nandwani and Verma, 2021), which ideally

encapsulates the entire multidisciplinary field, focused on exploring how technology can inform an understanding of human affect. (Daily *et al.*, 2017)

Sentiment Analysis is done using knowledge-based and rule-based approaches. Knowledge-based approaches are lexicon-based, while rule-based are machine and deep learning methods.

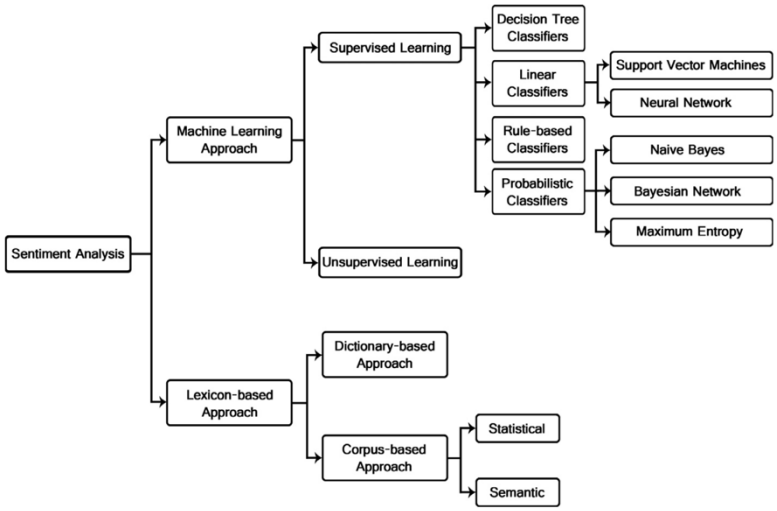


Figure 2 Sentiment classification techniques.

Sentiment Analysis Approach	Lexicon Based Approach	Machine-Learning Approach
1	Known as knowledge-based methods.(Cambria, 2016)	Known as Rule-based methods.
2	Economic and accessible.	Complexity limits the usability of these approaches.
3	No pre-requisite of annotated texts.	Involves novelty of labelling text.(Al-Moslmi *et al.*, 2018)
4	Has low recall as recall is dependent on coverage of database used. (Al-Moslmi *et al.*, 2018)	Better Accuracy
5	Poor Recognition of affect when linguistic rules are involved.(Cambria, 2016)	Feature extraction methods work great to extract linguistic rules for training of sentiment classification models.(Nandwani and Verma, 2021)
6	Cannot handle negation or long term dependencies on their own.	Deep Learning, dependency parsing and RNN (Recurrent Neural Network) with attention models handle long term dependencies well.(Nandwani and

		Verma, 2021)

Lexicon Based Approaches	Dictionary Based Approach	Corpus Based Approach
1	Adaptable and straightforward.(Nandwani and Verma, 2021)	Complex and domain specific.
2	Demonstrates less accuracy.(Nandwani and Verma, 2021)	Exhibits better accuracy.
3	Generalisation is easier broad variety of applications.(Nandwani and Verma, 2021)	Domain specificity limits generalisation.

Lexicon-based approaches have been around for a long time, and current research focuses on several tasks related to lexicon-based sentiment analysis. The baseline approach of sentiment analysis using lexicon involves building a database of words with their sentiment polarity, which can then be used to calculate the sentiment of a text with those words. Now the sentiment can be calculated at various levels of the text that involves (Ravi and Ravi, 2015):

- Sense level
- Clause level
- Link level
- Phrase level
- Concept level
- Document-level
- Sentence level
- Aspect level
- Word level

There are several tasks associated with lexicons, some papers attempt to introduce new lexicons (Al-Moslmi *et al.*, 2018), and others design algorithms to update existing algorithms to improve the accuracy of sentiment analysis. Like (Viegas *et al.*, 2020), they propose a new hypothesis, thus questioning a common assumption of representation of a word in vector space, that words with a similar representation in vector space have identical meanings. They corroborated the hypothesis using different open-source datasets and hence devised a new effective algorithm to update an existing lexicon.
Some popular, easily accessible lexicons are VADER, General Inquirer, TextBlob, NRC,

and Affect lexicon. (Cambria, 2016)

A study analysed a number of tweets by calculating the word frequencies of single and double words. Also, Latent Dirichlet Allocation was used for topic modelling of 167,073 tweets. Sentiment analysis was done using a text-blob dictionary-based approach in python. (Abd-Alrazaq *et al.*, 2020)
Another paper focussing on sentiment analysis and news topic classification used the WordNet lexical approach via the Rapid Miner tool. (Taj, Shaikh and Fatemah Meghji, 2019)

The Machine Learning approach (including the deep-learning method) has also been researched extensively.
(Aygun, Kaya and Kaya, 2021) discussed the sentiment around the vaccines developed in response to the covid 19 outbreak. They used BERT and BERTurk (deep-learning) models to cover four aspects of the six most frequently used vaccines and perform sentiment analysis.
Another research performing sentiment analysis on covid tweets using BERT and CT-BERT was done. They combined the BERT approach with an auxiliary-sentence approach to improving the classification. (Lin and Moh, 2021)

Some recent work like (Hussain *et al.*, 2021) includes feature extraction using a weighted ensemble of lexicons and then combining it using a rule-based ensemble method with a pre-trained BERT (deep learning) model to give a better score for neutral and negative texts (than individual lexicons). This kind of sentiment analysis is categorised under hybrid approaches and utilises the best of both worlds (lexicon-based and Machine Learning-based approaches for sentiment analysis).

Typical areas of application of sentiment analysis include affective tutoring, affective entertainment or troll filtering and spam detection. Some government intelligence applications of systems that help real-time sentiment mining of texts available over the Web include monitoring hostile communications or modelling cyber-issue diffusion. (Cambria, 2016)

Emotion Recognition

Emotion Recognition is a task that requires sentiment analysis as a preliminary step, followed by polarity determination and, finally, emotion detection. The emotion is labelled based on the polarity of the sentiment (sentence, aspect or document level) and the usage of an appropriate model that can assign an emotion label.

There is minimal research available regarding the detection and study of emotions. Sentiments have been analysed repeatedly, with a rare focus on any emotions that might emanate from them.
One is a paper by (Lwin *et al.*, 2020). They have deployed the CrystalFeel algorithm, which uses features extracted from several lexicons like Opinion Lexicon, AFINN, NRC-Emolex, NRC Hash-Emo and E2I lexicon. It then uses a composite of these features to predict the intensity of emotions. Further, it can generate a Pearson R correlation with the dates of the text (primarily Tweets), which can then help identify any trends that the emotions may be exhibiting over time.

NRC Lexicon, popularly known as EmoLex, has been around since 2010 for sentiment

analysis and emotion detection. (Mohammad and Turney, 2010)
It consists of 14,182 unigrams with emotion associations and 25000 senses. It is the most extensive lexicon with such information. It has been used in healthcare for understanding pandemic response toward influenza vaccinations (Roe *et al.*, 2021), depression detection (Tshimula, Chikhaoui and Wang*, 2022), hate-speech detection (Chiril *et al.*, 2022), cyberbullying (Talpur and O'Sullivan, 2020) etc.

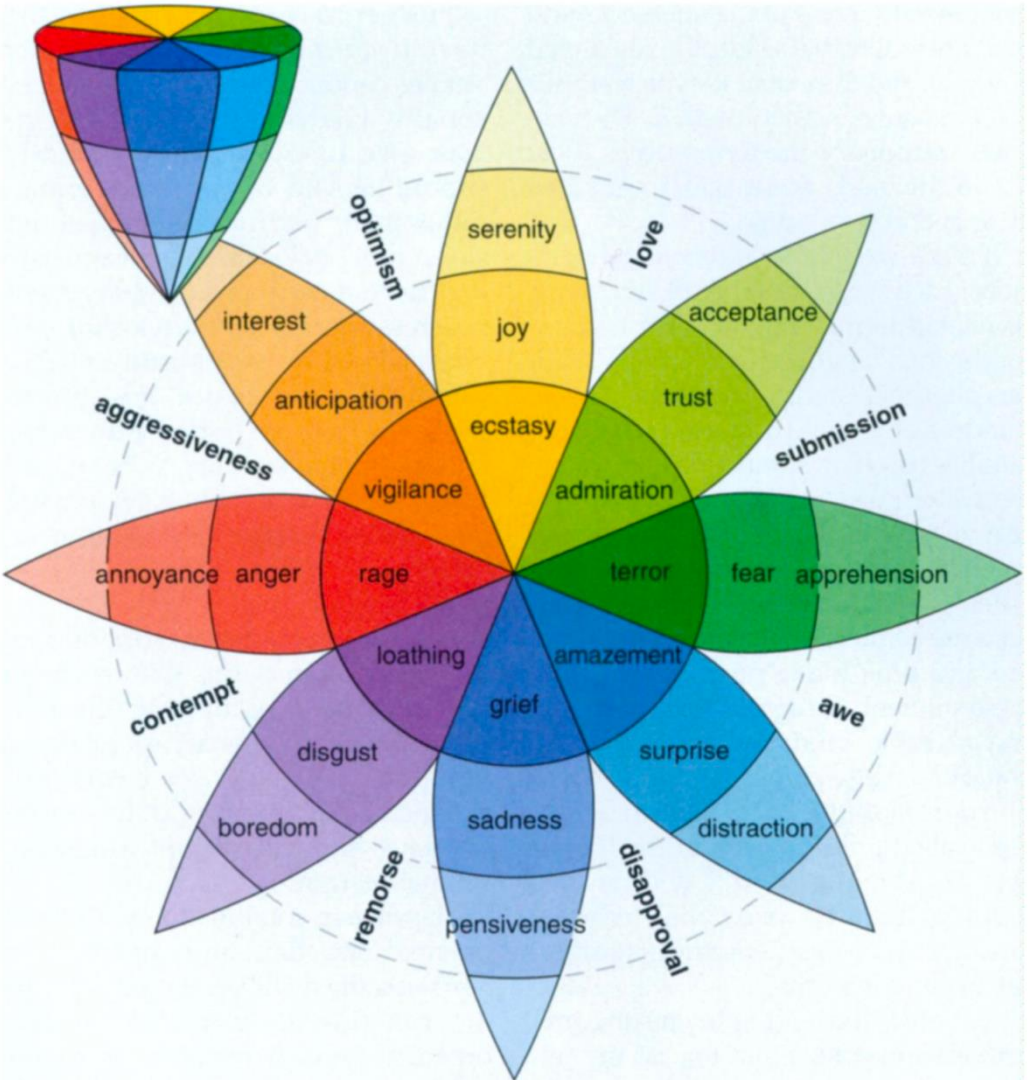
In a couple of these applications, the lexicon has been used to enrich text embeddings via lexicon-based features to be used in conjunction with machine learning approaches for different applications rather than sentiment detection. Hence, it is one of the most widely used lexicons for sentiment and emotion recognition.

Plutchik's Wheel of Emotions

(Plutchik, 2001)

The emotion-word associations in the NRC lexicon include eight primary emotions recognised by the psychoevolutionary theory:
Anger, disgust, sadness, surprise, fear, trust, joy, anticipation

Emotions, according to Plutchik, are feedback processes rather than a series of processes.
The emotions model proposed by him is analogous to the colour model. It is a circumplex model with shades of emotions (and colours) merging into ' imperceptible gradients'. The mixture of emotions at varying intensities gives several other emotions, called 'primary dyads'. A representation is a figure below:



Plutchik's Circumplex Model of Emotions

Emotion Dynamics

(Hipson and Mohammad, 2021)

NRC lexicon publication specifies an official way to analyse word-emotions in a text using 'Emotion dynamics'. It has been recently used to examine the use of words and emotions associated with dialogues of characters from a movie to chart a trajectory of variation of emotions exhibited throughout the narration. It also visualises the movement of the character's emotions from a baseline model representing the entity most often in the story narrated by the movie. This is done using the valence, arousal, and dominance lexicon (NRC-VAD), essentially a variation of the basic NRC lexicon. Along with the emotion and sentiment scores, it also contains the scores that define the degree of valence (pleasure or displeasure), arousal (excitation or calmness) and dominance (power or weakness) inflicted by the word used.

In this paper, the code from this 'Emotion Dynamics' has been exploited to plot the

datewise trends of the eight primary emotions of Plutchik's model in the news articles collected for the year 2021, using the keyword 'covid'.

Research Gap

A very limited account of publications on sentiment and emotional analysis can be seen in the era before the pandemic. The major focus during this post-pandemic period is on the analysis of sentiment trends using social media data, i.e. Facebook and Twitter. It can be due to the fact that their data is very easily accessible via APIs. (Lwin *et al.*, 2020)(Samuel *et al.*, 2020) (Hussain *et al.*, 2021)

Reliable news is, however, mainly disseminated via official texts from either the government or the media websites. These have been exploited previously for their utility in predicting stock prices (Rai, Kasturi and Huang, 2018) but very rarely in-depth for emotion recognition. Attempts at sentiment analysis for stock price predictions, however, have been made before. (Agarwal, 2020)

2.3 Related Work

Recent works that deploy lexicon based approaches are (Bhat *et al.*, 2020), where a simple sentiment analysis of tweets is done using a lexicon based approach to assess the attitudes of people during the time of lockdown. They conclude that most tweets indicate a positive attitude towards the whole situation and look forward to adjusting to this new norm in the future.

Another work is by (Raamkumar, Tan and Wee, 2020), where strategies for pandemic measures dissemination by the Public Health Authorities via social media are studied, owing to their increased use of this medium. Responses to the posts were studied using polarity, frequency, likes and dislikes of the users. The purpose of analysis was to highlight the impact the posts may have caused and it focused on suggesting improvements on these practices.
A paper (Pastor, 2020), discussed the effect of quarantine due to Covid in Phiippines via analysing tweets. He found that most tweets were negatively polarized which is an important revelation as far as educating public policy is concerned. Critical concerns like food shortage came to the surface during the qualitative analysis.

Another recent paper (Qanita Bani Baker *et al.*, 2019) analysed opinions in Arabic tweets which deployed machine learning algorithms to test the performance of the approach used to collect, label, filter and analyse the influenza related texts.

Not many works related to news text analysis for the purpose of sentiment and emotion analysis were found due to lack of academic research in the field. This makes this work quite unique in its contribution.

2.4 Summary

The chapter starts with citation analysis details of the papers published in the field of interest via the platform citespace. It then presents the review of literature related to strategizing public discourse, natural language processing, sentiment analysis and its two approaches, emotion analysis and theories of emotion. It also reviews literature from emotion dynamics algorithm of the NRC lexicon and similar works in the field.
It lays a background to discuss basic concepts building up to mention recent research in each of the fields. It then mentions research gap and any related works for the same.

Chapter 3 Methodology

The chapter consists of the following parts. Research Question mentions the research question answered using the study design. The methodology concluded the process of designing and the design of the method used to answer the question. It has sub-parts like research philosophy, approach to theory development, research approach, correlational study design and summary of certain statistics concepts being used in the study like p-value, alpha and ANOVA test.

3.1 Research Question

This study attempts to test the hypothesis that emotions arising from the spread of information via news articles (public discourse) have a physiological impact on people leaving them more susceptible to the risk of infection.

This is done via testing a correlation between the emotion that is likely to be conveyed by the language of the articles and the number of covid cases arising in the subsequent days after publishing the article.

3.2 Research approach

Research Onion (Saunders *et al.*, 2019) is one of the popular methods used to finalise a methodology to address a research question. The diagram below shows the layers of the process, beginning from the outermost.

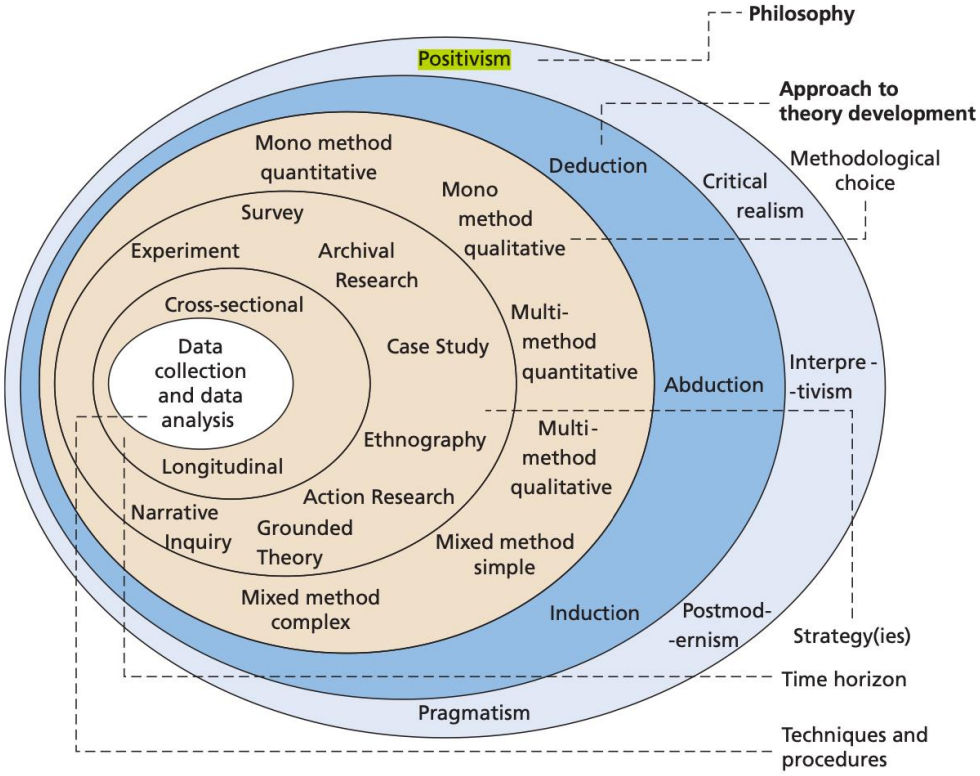


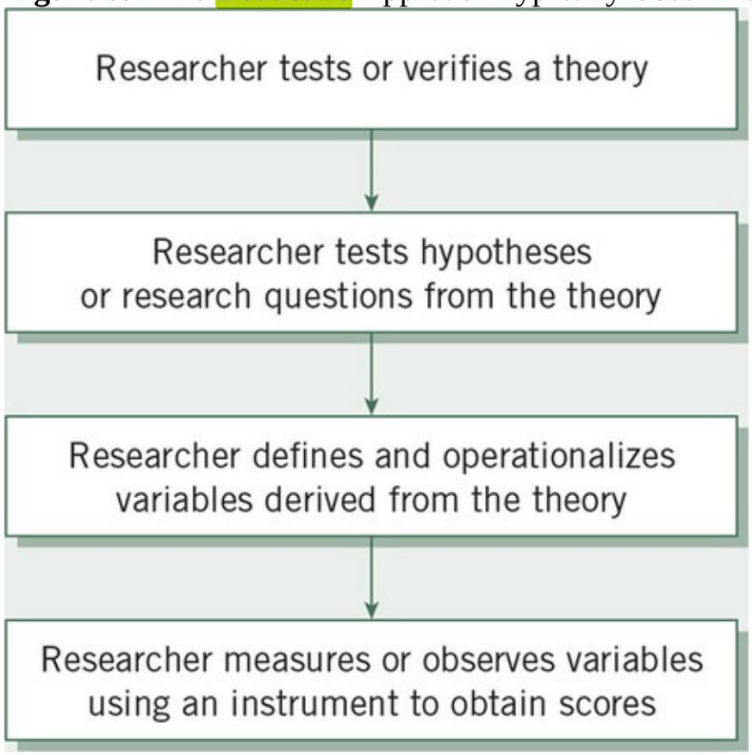
Figure 4.1 The 'research onion'
Source: ©2018 Mark Saunders, Philip Lewis and Adrian Thornhill

The outermost layer is Research Philosophy:
This is mainly related to the research scientist's beliefs that are reflected through the work. The ontologies and epistemologies are finite, and the philosophies arising from these may be:

1. Empiricism
2. Positivism
3. Pragmatism
4. Objectivity
5. Induction
6. Refutation
7. Occam's Blade
8. Constructivism

The nature of this study (quantitative) undertaken conforms to the post-positivist philosophies of scientific computing. This philosophy adheres to positivist beliefs and argues the traditional idea of being strictly optimistic about our approach toward knowledge.(Creswell, 2018)
The next layer is that of the approach to theory development. Since the hypothesis of this study emanates from the theory that the use of web and social media for discussions on popular topics may lead to an elongated existence of particular views, thus carrying the potential to twist public opinions on things and hence induce polarisation field(Bisiada, 2022), it is pretty clear the approach here is deductive, rather than inductive which follows an opposite pattern.

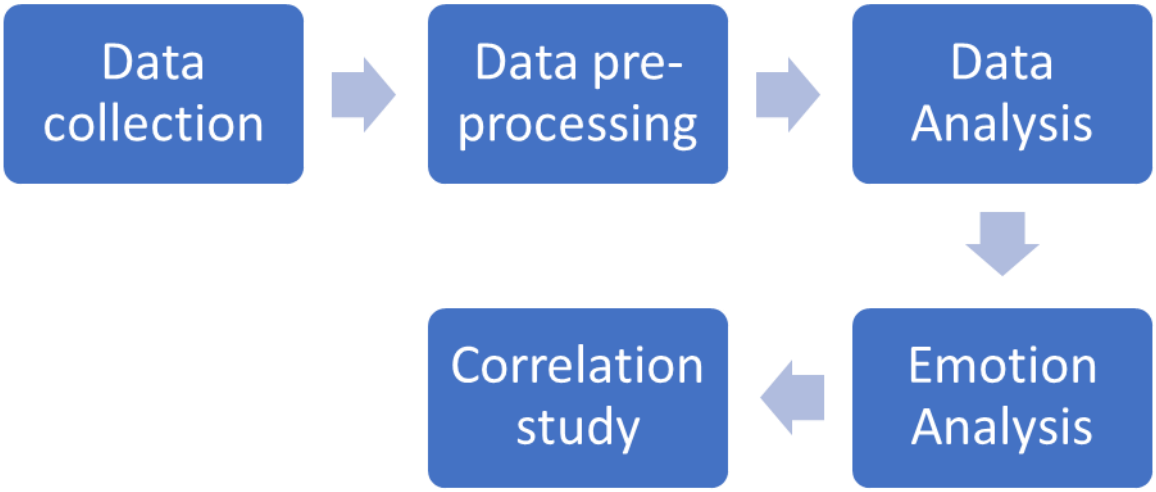
Figure 3.4 The Deductive Approach Typically Used in Quantitative Research



Research methodology There are three kinds of research methodologies, two of which are on extreme ends of the continuum, i.e. Qualitative and Quantitative. The mixed method approach sits in between the two. Qualitative research poses open-ended questions, while quantitative addresses closed-ended to queries. The mixed methods approach is a mix of these approaches to answer more complex research questions.

This report's study is quantitative, which tests the theory that public discourse has a significant emotional impact by quantifying its effect with a correlational experiment design. It explores the emotions arising from news articles published and tries to quantify the reaction via the number of covid 19 cases arising in the subsequent days. It proposes a correlation between the two. However, the causality is indeterminable due to the lack of inclusivity of published texts in the previous year owing to legal restrictions against web data scraping and other influencers like social media. The method is the ‘mono’ method, i.e. quantitative.

Correlational study design



The study explores any correlational existence between the publishing of news articles and fluctuations in the number of cases reported during the pandemic of covid 19 to highlight any impact public discourse might have via the affective study of news articles.
The experiment assumes that the effect of social media or other influencers on the public's emotions is negligible.

The steps involved in the study are

- Data collection
- Data pre-processing
- Data Analysis
- Emotion Analysis
- Correlational study

Data Collection

The cross-sectional study starts with collecting news headlines using the GoogleNews API key, and the text was scraped from the web using the news article 'URL' provided by API. The keywords 'covid' and the location 'UK' were used.
A total of 8071 articles were collected between 31-12-2020 and 31-12-2021. A few dates were exempted as scraping headline text was limited by the news website's restricted access. 15-12-2021 and 16-12-2021 are some of these dates.

Data Preprocessing

The text from the articles was preprocessed using the nltk library of python. The process was adapted utilising an amalgamation of techniques keeping in mind the specificities of the raw data collected via scraping. Also accounted for are the prerequisites of the data fed into the 'Emotion Dynamics' pipeline to get the emotion valences for the texts of the news headlines.

Data Preparation

The first step is replacing special characters or punctuation that do not affect sentiment analysis but only help in text normalisation and noise reduction. (Steven Bird, Evan Klein, and Edward Loper, 2009) (Fehle, Schmidt and Wolff, 2021)
The second step of changing the text all to lowercase does not affect the sentiment

of the text. Still, it helps with the inconsistencies found in the web data, i.e. spelling errors or incorrect capitalisation that may hinder accurate sentiment analysis. (Fehle, Schmidt and Wolff, 2021)
The third step is tokenisation, an application domain-oriented task (Steven Bird, Evan Klein, and Edward Loper, 2009). Here, it has been done to perform word-level emotion analysis using the NRC lexicon via the “Emotion dynamics” algorithm. Even though this is not a pre-requisite of the data being fed into it, tokenisation is done here to ensure the removal of any special characters that might not have been removed by previous code and conversion of text into a refined corpus.
The fourth step is removing stop words, which are crucial for dimensionality reduction, optimising computational costs and improving the algorithm's performance. Stop-words also might not carry a lot of sentiment/emotion information.
(Fehle, Schmidt and Wolff, 2021)

Lemmatization has been done to reduce words to their base form. It is because lexicons often contain sentiment or emotion valences of the base form of the word and not inflexions. It is preferred over stemming as truncation of suffixes may not always give the base form of the word, hence impeding the matching of words between text and lexicon.
(Fehle, Schmidt and Wolff, 2021)

A common concern with lemmatisation is, if we pass just the token to the lemmatiser in the NLTK package, it defaults the word to noun and hence errors may occur which may not give the result as desired. To counter that problem, the pos tagging is done and the passed to the lemmatiser along with the token for more accurate results.
It does take longer to process than without the pos tagger.

Other preprocessing steps are recommended and are famous, like replacing contractions numbers with text conversions recommended for better text analysis later in various sentiment analysis works. As for this text, the pre-processing pipeline was coded keeping in mind the specificities of its domain. (Steven Bird, Evan Klein, and Edward Loper, 2009)

Some additional steps are taken in the preprocessing file to ensure the data complies with the requirements of the Emotion Dynamics algorithm. (Hipson and Mohammad, 2021)

Data & Emotion Analysis

The text of the news articles was analysed using parameters like word count, article length and word clouds.
The polarity was also calculated using a text blob to determine a preliminary sense of polarity distribution amongst texts.
Analysis was also carried out after the emotion valences of texts were determined using word clouds and frequency charts.
Firstly the overall emotion of the text was aggregated by finding the highest valence and the emotion corresponding to it.
The word clouds from texts corresponding to the highest overall emotion count were plotted, which indicated the inclusion of many stopwords. This was handled by the ‘CountVectoriser’ of python to filter the main words and display the unigrams and

bigrams.
The bar plot was plotted for the top frequency unigrams and bigrams to display the most common words in the texts corresponding to the high frequency of overall emotion.

Correlational Study

A correlational design is different from an experiment design since there are no manipulated measured variables. Here the variables being measured are:

- Emotion valences of the news texts
- New cases are recorded on the government website.

Amongst the various data collection methods in a correlational study, the nature is "archival".

Correlation statistical test is done to measure the relationship and its significance between two numerical continuous variables. The correlation is measured using a coefficient whose value ranges between [-1,1].

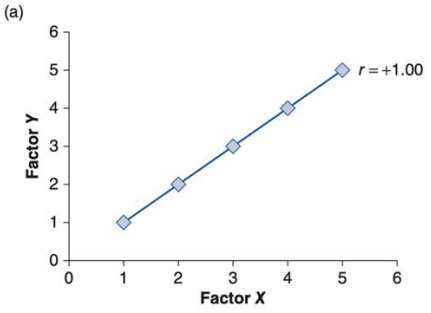
Direction and strength of the correlation

A positive sign signifies a positive correlation. Examples of a correlation of 1 and close to one are represented below in the figure.
The value when close to 1, indicates that increase in one variable implies an increase in the other variable. A value of the correlation coefficient less than zero and close to -1 indicates a decrease in one variable's value when the value of other variable is increased.

Figure 8.3 A Perfect Positive (a) and a Positive (b) Linear Correlation

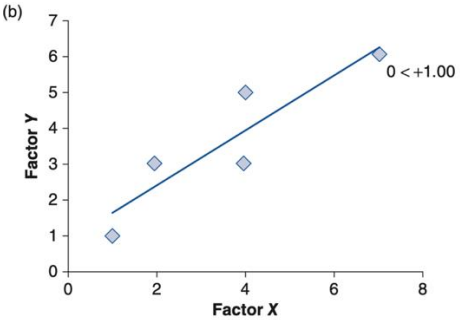
(a)

X	Y
1	1
2	2
3	3
4	4
5	5



(b)

X	Y
1	1
2	3
4	3
4	5
7	6



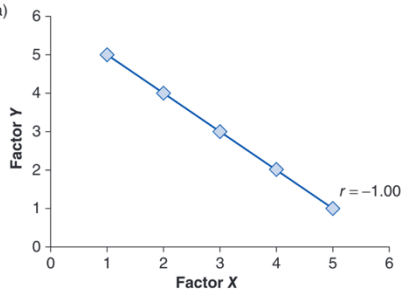
Both the table and the scattergram show the same data for (a) and (b).

A negative value signifies a negative correlation. The value of -1, or a value close to -1, is represented below.

Figure 8.4 A Perfect Negative (a) and a Negative (b) Linear Correlation

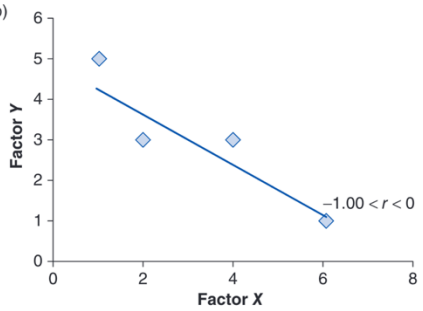
(a)

X	Y
1	5
2	4
3	3
4	2
5	1



(b)

X	Y
1	5
2	3
2	3
4	3
6	1



Both the table and the scattergram show the same data for (a) and (b).

The expression gives the formula for Pearson correlation.

$$r = \frac{\text{variance shared by X and Y}}{\text{total variance measured}}$$

The *Pearson correlation coefficient* measures the direction and strength of the linear relationship between two factors. The data for both elements are on an interval or a ratio scale of measurement.
The variance in the numerator, called covariance, is the amount or proportion of the total variance shared by X and Y. The larger the covariance, the closer data points will fall to the regression line.(Price, Chiang and Jhangiani, 2014)

P-value
It is the probability of obtaining results as unusual or extreme as observed results,given a chance model that embodies a null hypothesis. (Bruce, Bruce and Gedeck, 2020)

Alpha
The probability threshold of unusualness that chance results must surpass for actual outcomes to be deemed statistically significant.(Bruce, Bruce and Gedeck, 2020)

The alternate hypothesis in this study is that there is a significant correlation between the values of emotional valence and new covid cases reported.
A lag of 0 to 14 days is explored to find significant correlation values based on the p-value of the correlation. The alpha value is kept at 5 per cent, i.e. any correlation for which the p-value is 0.05 or less is considered significant.

A spearman correlation test is done to see if the correlation and lag values exhibit any significant relationship.
A spearman correlation coefficient is used when both variables' scale is either ranked or ordinal. Here the lag (in days) is an ordinal variable. And the correlation coefficient value between the emotional valence and the covid cases after a specific lag is a continuous variable. Since it is still ordinal, a spearman test deems appropriate. Also, there is no need for us to check the distribution for the spearman test as it fits for non-parametric variables.(MacFarland and Yates, 2016)

Anova Test

Anova test is an analysis tool used in statistics that splits an observed aggregate variability found inside a dataset into two parts : systematic factors and random factors. The systematic factors have a sif=gnificant influence on the target variable while the random factors don’t.
One way Anova is done between an independent continuous variable and a dependent nominal variable. If the p-value is less than alpha, the effect of the systematic variables is concluded significant.

Anova test here is done between the emotion categories, with whose valence the correlation was found significant when tested against covid cases in the subsequent days.

3.3 Summary

This chapter successfully delineates the research question that is answered in this thesis and the approach adopted to answer the question. The entire design of the correlational study is described. Introduction to statistical concepts used is also done.

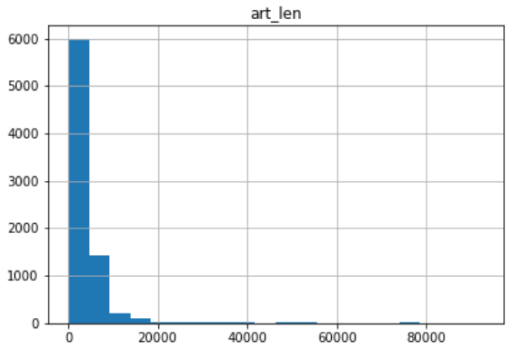
Chapter 4 Discussion & Findings

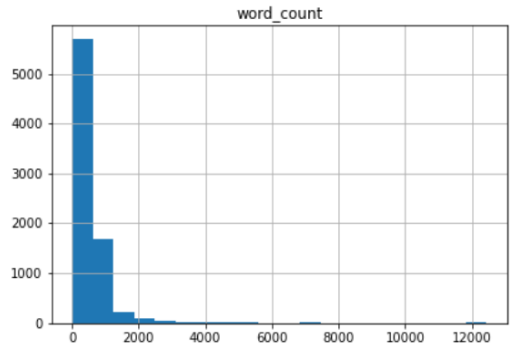
4.1 Visualisations

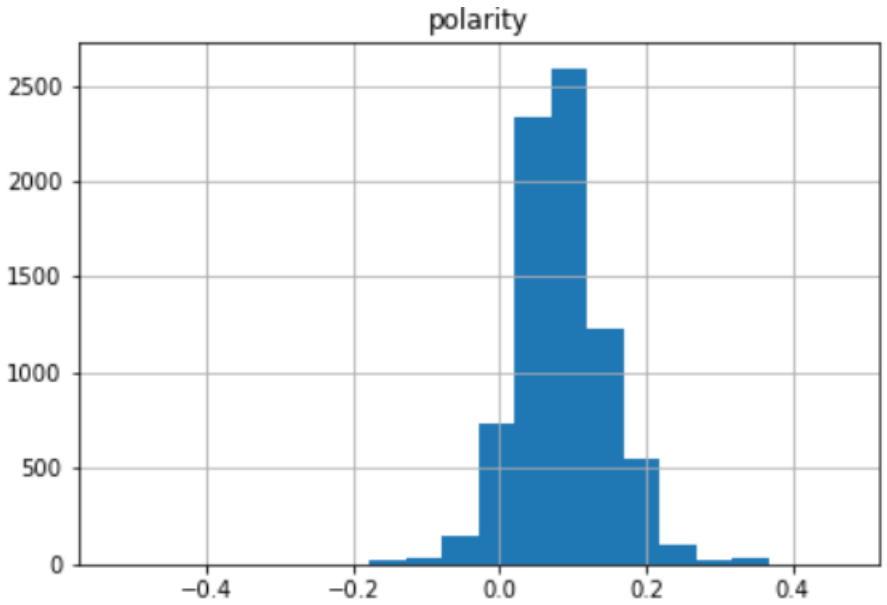
This chapter discusses the visualisations used for the analysis of texts using the python code. It consist of preliminary text analysis on the basis of polarity and emotion. It then discusses the comparisons between the most frequent unigrams and bigrams in the same. The findings of the correlational study are then presented, which is followed by further analysis based on the results of the correlational study.

Text Analysis and Results

The preliminary text analysis of the news articles includes a basic assessment of each piece that involves counting words, calculating character lengths of the paper and polarity for an introductory study of things.
Following were the distributions of the word count, character length and polarity amongst the entire dataset of 7791 articles.







The distribution of polarity is normal and centred at slightly more significant than zero, meaning the texts generally are of the polarity positive. This is a positive

indication of discourse management on the part of the media.
The word count distribution helps us conclude that most articles have words less than 500, while the article length distribution exhibits that the texts have less than 10000 words on average.

The total number of positive, negative and neutral texts as per the nltk package in python is:
Positive 7315
Negative 466
Neutral 10

4.2 Discussion (Analysis)

Polarity classification

The polarity classifications of the texts as per text blob and NRC are comparable. Even though the numbers of positive and negative readers differ slightly in the two categories. The top 20 unigrams in the positive and negative texts remain the same. However, the percentage of words in both texts differs.

Positive corpus from text_blob has the following composition (top 20 words only):

```
covid 0.19583642061127915%
say 0.14344024879406486%
people 0.1173592777495778%
vaccine 0.10246023025325428%
case 0.07555436320931254%
health 0.06859196940321557%
test 0.06612407067233675%
uk 0.0563067898887032%
new 0.05441597889406838 6%
government 0.05039673254281592%
use 0.04808159233035462%
make 0.04522331071907542%
pandemic 0.0450739468344005%
week 0.04369572553490007%
country 0.04265017834217562%
work 0.042297136432943985%
virus 0.041533343840856315%
year 0.0408442331911061%
day 0.04082726002239304%
number 0.04003631036036447%
```

Positive corpus from NRC sentiment polarity has the following percentages of top 20 words:

```
covid 0.22838280901749866%
say 0.16727882813372172%
people 0.13686341607476102%
vaccine 0.11948827049012242%
case 0.08811087156010518%
health 0.07999138566489128%
test 0.07711334263913774%
uk 0.06566451123545242%
new 0.06345946314006767%
government 0.05877225138973453%
use 0.05607235408084331 6%
make 0.05273904977866385%
pandemic 0.05256486285550958%
week 0.05095759261004060 4%
country 0.0497382841479607%
work 0.04932656960232334%
virus 0.04843584101801172%
year 0.04763220589527724%
day 0.04761241926736 98%
number 0.04669001299276094 6%
```

Negative texts from text_blob and NRC classification consist of the following top 20 unigrams.

```
covid 0.28586176935482316%
say 0.13959647193482505%
people 0.12228317500216415%
case 0.11779454246406688%
test 0.0874000878489511%
health 0.07925642595840321%
vaccine 0.06835546122302412%
day 0.06713711810554057%
uk 0.06534166509030166%
week 0.06380270536295404%
data 0.06219962231363358%
new 0.05886520957104704 5%
death 0.0583522229952645%
pandemic 0.05751861980961786%
infection 0.05296586394954777 6%
rate 0.05033680774866223%
number 0.04982382117287969%
report 0.04764362822580387%
government 0.04738713493791259%
virus 0.04681002504015723%
```

```
covid 0.07730084455247528%
say 0.03774875248782832%
people 0.033067005509549205%
case 0.031853219255921286%
test 0.023634152338497937%
health 0.021431997278344425%
vaccine 0.018484230662390903%
day 0.01815477439354904%
uk 0.01766925989209787%
week 0.017253104605139725%
data 0.016819609514558326%
new 0.01591793972614901%
death 0.015779221297162967%
pandemic 0.015553803850060636%
infection 0.014322677792809 46%
rate 0.013611745844255965%
number 0.013473027415269916%
report 0.012883474092079211%
government 0.012814114877586187%
virus 0.012658056644976884%
```

Emotional Dynamics classification

[NOTE: 'trust' and 'fear' texts refer to the corpus comprising combined texts where trust and fear are dominant emotions.]

The texts, when classified according to the highest valence of the emotions, give the following frequencies of emotions in texts:

```
avgLexVal_anger 15
avgLexVal_anti 803
avgLexVal_disgust 4
avgLexVal_fear 1920
avgLexVal_joy 7
avgLexVal_sad 205
avgLexVal_trust 4837
```

Fear and trust are the emotions with the highest valence in most texts. To further analyse which words are common in each of these texts, the following percentage composition of texts with several top frequency words were found:

Trust corpus

```
covid 0.1793124319231402%
say 0.13843259731138743%
vaccine 0.10911129163430729%
people 0.10441357159612902%
health 0.07031311194738106%
test 0.06980414986307137%
use 0.05706482889280032%
case 0.05352754240684811%
government 0.05259641792561404%
uk 0.05159857610731456%
new 0.04822924710918443%
make 0.04747598322440612%
work 0.04576078100028252 4%
pandemic 0.04223367375601650 6%
data 0.04197919271386167%
country 0.04012148110613137%
vaccination 0.03917990125015848%
week 0.03903739186655177%
need 0.03837574115694919 5%
year 0.03784133096842404%
```

Fear Corpus

```
covid 0.22363721403755218%
say 0.15671025526530039%
people 0.14432863546869265%
case 0.13961435028410188%
vaccine 0.08685673181385035%
death 0.07415729503850618%
uk 0.07170745582741267%
new 0.07096588287702761%
infection 0.06818498431308362%
health 0.06679453503111163%
test 0.05751163077718434 6%
number 0.05643899847394881%
week 0.05612118149521235%
pandemic 0.05598875775407216 6%
virus 0.05549878991185345 4%
rate 0.05539285091894130 5%
government 0.05516773055900298%
variant 0.05514124581077494 6%
day 0.05469100509089830 4%
country 0.05108907933188514%
```

- ‘covid’ appears in both, but in ‘trust’ corpus has a slightly higher percentage out of total words.
- ‘say’ in the ‘trust’ corpus constitutes 0.1384%, while in ‘fear’ is 0.1567%.
- ‘Vaccine’ has a significantly higher percentage in the ‘trust’ corpus(0.109%) than in fear texts (0.08%).
- ‘people’ appears more often in the ‘fear’ corpus(0.144%) than in ‘trust’(0.1044%).
- ‘health’ constitutes a higher percentage in the ‘trust’ corpus(0.07%) than in the ‘fear’ (0.067%) corpus.
- ‘Government’ has a slightly less percentage in the ‘trust’ texts(0.052% > 0.055%).
- The ‘test’ word forms a higher per cent in the ‘trust’ texts(0.069%) than in the ‘fear’ texts(0.057%).
- The word ‘new’ has an interestingly higher percentage in the fear texts(0.048% <0.0709%).
- ‘Week’ forms 0.039% of the trust corpus and 0.056% of the fear corpus.
- Unique top frequency words in the ‘trust’ texts are ‘make’,’ work’,’ data’,’ vaccination’,’ need’, and’ year’.
- Unique top frequency words in the ‘fear’ texts are ‘virus’,’ rate’,’ variant’,’ day’,’ death’,’ infection’, and’ number’.

The differences between the words used for articles based on the highest emotion valences can be a real insight into what triggers a positive reaction in public and which words are like to wreak havoc when times are this sensitive.

Another comparison here can be made between the bigrams of the texts.

For their trust texts, the percentage of top frequency words are:

```
covid vaccine 0.021111747257165154%
public health 0.01539101342952445 1%
email address 0.009817878606333555%
test positive 0.009151138275887885%
covid case 0.007588624677057 1954%
prime minister 0.006764106100475528%
film tonight 0.006662313683613594%
covid pandemic 0.006077007286657472%
covid test 0.005827615865345734%
external site 0.005751271552699283%
social care 0.005674927240052832%
bbc news 0.005629120652464962%
covid vaccination 0.005349191506094643%
new variant 0.005176144397429355%
care home 0.005079441601410517%
wed like 0.005033635013822646%
fully vaccinate 0.004982738805391679%
vaccination programme 0.004906394492745229%
million people 0.004824960559255682%
delta variant 0.004814781317569489%
```

And for the fear texts the percentage of top words in the entire corpus are:

```
covid case 0.019492774695835907%
covid vaccine 0.015996787929734902%
infection rate 0.015387638720490003%
public health 0.014884428504157311%
new case 0.014050158934974117%
test positive 0.012792133394142315%
covid death 0.008845905908164666%
delta variant 0.008369180440059982%
positive covid 0.008236756698919793%
new variant 0.008183787202463717%
people die 0.007733546482587073%
bbc news 0.007667334612016977 5%
care home 0.007614365115560902 5%
prime minister 0.00737600238150856%
seven day 0.006925761661631917%
covid infection 0.006780095546377708%
fully vaccinate 0.006594702308781442%
external site 0.006568217560553404%
covid patient 0.006554975186439386%
film tonight 0.006515248064097328 5%
```

Interesting revelations from the above two outputs are:

- ‘Covid vaccine’ (0.0211%) has a higher percentage in the trust corpus, while ‘covid case’ occupies a more significant percentage of the fear corpus (0.0194 %).
- ‘public health’ is higher in percentage in the ‘trust’ corpus (0.01539%) than in the ‘fear’ corpus(0.01488%).
- ‘prime minister’ is another one with a significant percentage difference in percentage and is higher in fear ‘texts’ (0.0067%) compared to ‘trust’ texts (0.0073%).
- ‘fully vaccinate’ bigram has a slightly higher percentage in ‘fear’ texts (0.0065%) than in ‘trust’ texts (0.0049%).
- ‘test pos’ appears higher in ‘fear’ texts (0.01279%) than in ‘trust’ texts. (0.00915%)
- The phrase ‘covid death’ appears only in the fear texts, while the ‘covid test’ appears only in the ‘trust’ corpus.
- ‘external site’ constitutes 0.00575% in the trust corpus while 0.00656% in the fear corpus.
- ‘new variant’ occupies 0.008183% of the fear corpus while only 0.005176% of the ‘trust’ corpus.
- ‘care home’ bigrams are found in 0.00507% of the trust corpus, while fear corpus has a slightly higher percentage of 0.0076.
- ‘delta variant’ appears more often in the fear corpus(0.00836%) than in the ‘trust’ corpus(0.004814%)
- Unique bigrams in the trust corpus include ‘covid pandemic’, ‘covid test’, ‘covid vaccination’ and ‘vaccination program’.
- Unique bigrams in the fear corpus include ‘infection rate’, ‘new case’, ‘positive covid’,’ people die’,’ covid infection’ and ‘covid patient’.

Moreover, suppose we were to discuss the nature of emotions that seem most dominant in texts from Plutchnik's wheel of emotions addressed previously. In that case, it is noted that trust and fear are not very distantly positioned on them. Hence, it is essential how a choice of a few words can shift the whole notion of discourse. The anxiety or faith emanating from such texts can further lead to distressing physiological impacts on the health of the masses, worsening the situation as health remains the main focus of news dissemination.

Findings and Discussion(Correlation Test)

This test is done by combining the emotional valence dataset and the covid cases dataset repeatedly for a lag period of (0-14) days.
The Pearson correlation coefficient was calculated along with the p-value. Only the results with a p-value below alpha=0.05 were considered significant and stored in a different dataset.
In a new data frame, the following values of significant correlation results were saved:
Emotion value, correlation value, p-value and lag (in days) when the correlation was significant.

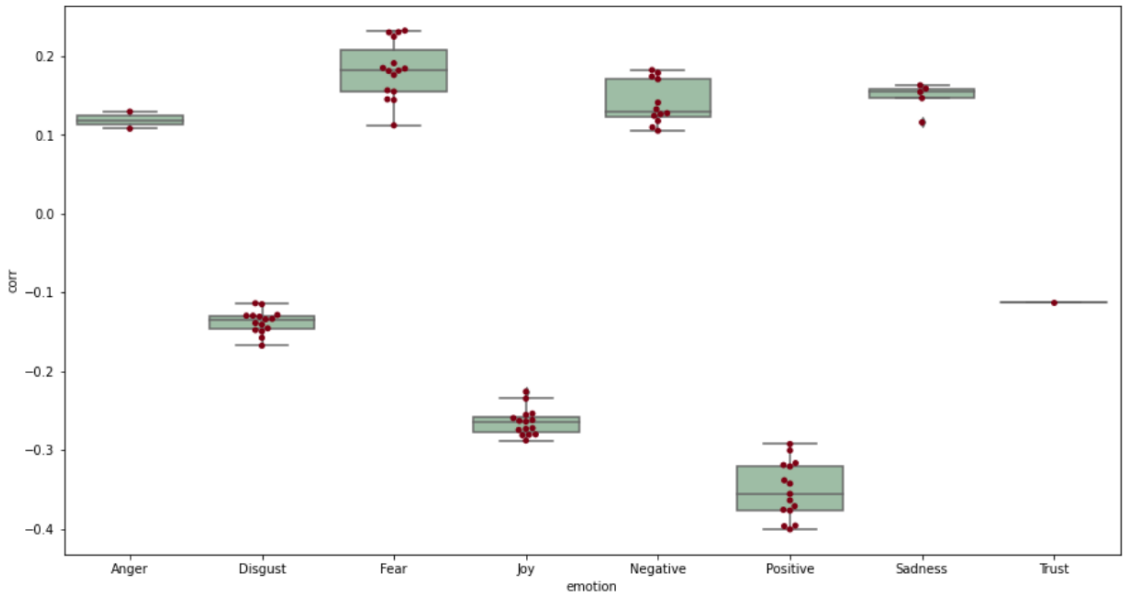
The frequency of the lags(in days) when the correlation between the emotional valence and covid cases was found significant are:

```
0 days 5 freq
1 days 7 freq
2 days 7 freq
3 days 6 freq
4 days 6 freq
5 days 6 freq
6 days 5 freq
7 days 4 freq
8 days 4 freq
9 days 5 freq
10 days 5 freq
11 days 5 freq
12 days 5 freq
13 days 6 freq
14 days 4 freq
```

The frequency of emotions where the correlation is significant is:

```
Anger days 2 freq
Disgust days 15 freq
Fear days 15 freq
Joy days 15 freq
Negative days 12 freq
Positive days 15 freq
Sadness days 5 freq
Trust days 1 freq
```

The most frequent lags (in days) where the significant correlation with specific emotion valences are 1,2,3,4,5,13.
The most frequent emotions where the correlation is significant are:
Disgust, Fear, Joy emotions and Positive sentiment.
An ANOVA test was done to hypothesise that the mean correlation values for different emotions differ. Following were the results.



	sum_sq	df	F	PR(>F)
C(emotion)	0.125560	7.0	97.409865	3.944385e-34
Residual	0.013258	72.0	NaN	NaN

The p-value was relatively low. Hence the differences between the means of correlation values(squares, to be exact, for a normal distribution) based on emotion were found to be significant.

Another spearman test was done to examine the hypothesis that there is a significant correlation between lag values (in days) and correlation values (squares). Following were the results:

```
corr,p_val=scipy.stats.spearmanr(sig_cor_data['corr'],sig_cor_data['lag'])
print(corr,p_val)
```

```
-0.17065567617311833 0.1301542791432275
```

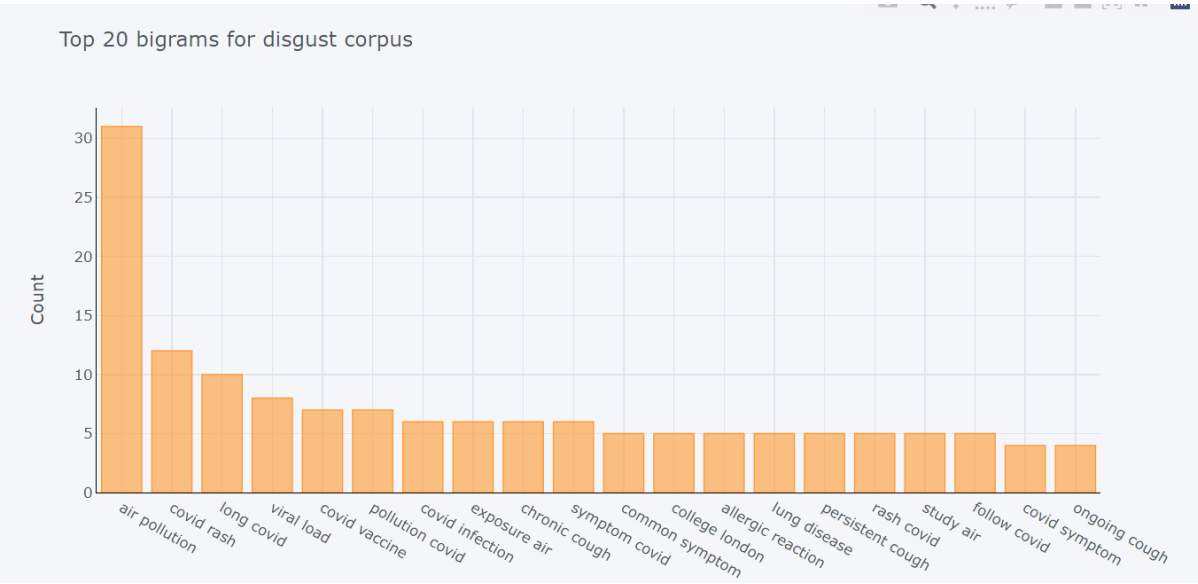
The p-value is high (not less than the value of alpha=0.05), so the alternate hypothesis mentioned above is rejected. The null hypothesis that there is no significant correlation between the lag values and the correlation values calculated previously(here, squares are not taken as the spearman test is a non-parametric test) cannot be rejected.

Post Analysis and Findings

The corpus of disgust and joy was also analysed after the correlation test. Following were the results:

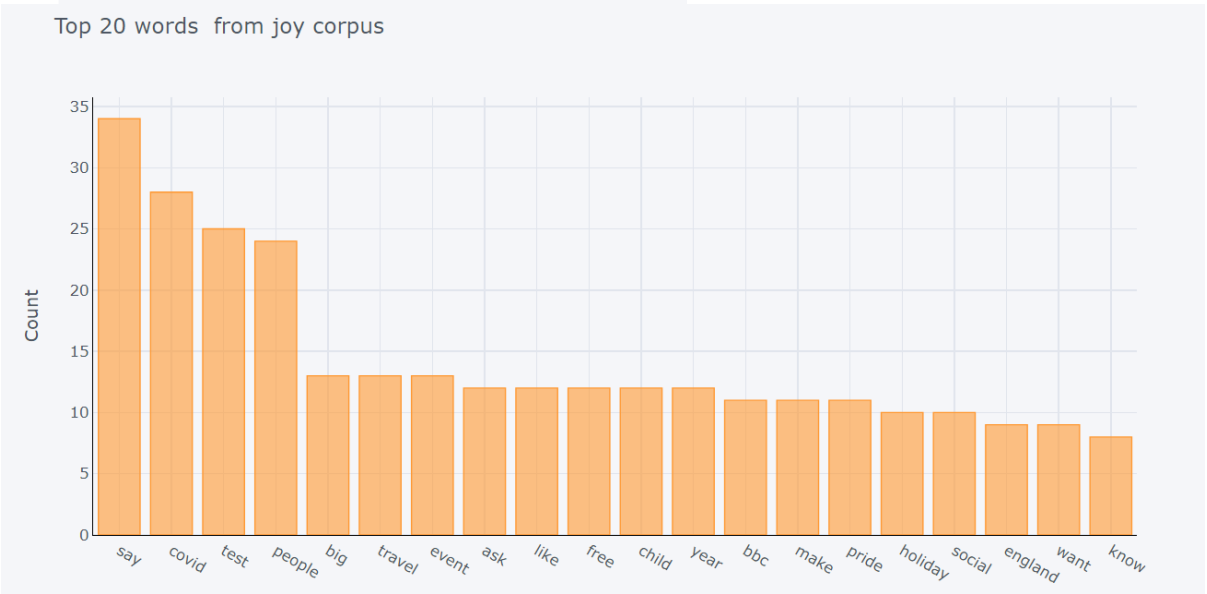
For Disgust, the top 20 unigrams and their percentages are:

```
covid 0.6252150969370196%
cough 0.24664448778249398%
rash 0.23517265114144772%
air 0.1835493862567397%
symptom 0.17781346793621658%
pollution 0.17781346793621658%
study 0.12045428473098543%
people 0.10324652976941608%
infection 0.09751061144889298%
disease 0.09751061144889298%
virus 0.09177469312836985%
viral 0.06883101984627739%
like 0.06883101984627739%
link 0.06883101984627739%
vaccine 0.06883101984627739%
cause 0.06883101984627739%
report 0.06883101984627739%
lung 0.06883101984627739%
patient 0.06309510152575427%
london 0.06309510152575427%
```



For Joy, the percentages and the graph are

```
say 0.22957461174881838%
covid 0.18906144496961513%
test 0.16880486615800135%
people 0.16205266711681296%
big 0.08777852802160703%
travel 0.08777852802160703%
event 0.08777852802160703%
like 0.08102633355840648%
year 0.08102633355840648%
ask 0.08102633355840648%
child 0.08102633355840648%
free 0.08102633355840648%
make 0.07427413909520594%
bbc 0.07427413909520594%
pride 0.07427413909520594%
social 0.0675219446320054%
holiday 0.0675219446320054%
england 0.06076975016880487%
want 0.06076975016880487%
know 0.05401755570560432%
```

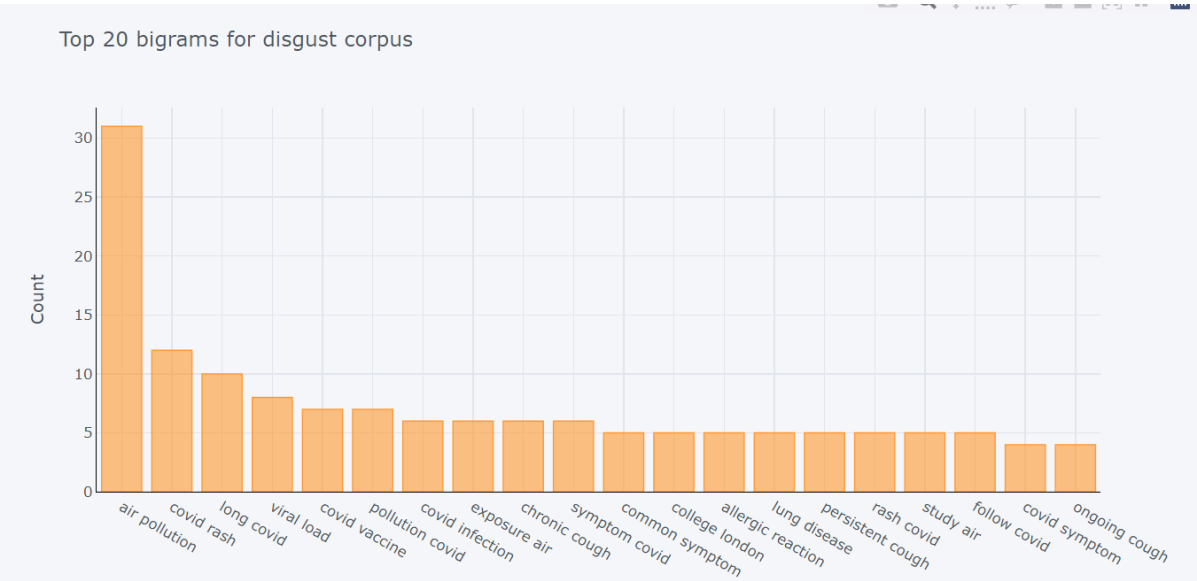


Findings (Unigrams)

- The word ‘covid’ appears more often in the disgust texts (0.625%) than in the ‘joy’ texts (0.189%).
- Words like ‘cough’, ‘air’, ‘symptom’,’ pollution’,’ study’,’ people’, ’infection’,’ virus’,’ viral’,
  ’vaccine’, ’cause’ and ‘report’ are unique to the corpus of disgust.
- The ‘people’ appears much more often in the ‘joy’ corpus (0.1620%) than in the ‘disgust’ corpus (0.103%).
- Other words unique to the joy corpus are: ‘say’, ‘test’, ‘big’, ‘travel’, ‘event’,’ year’,
  ‘ask’,’ child’,’ free’,’ make’,’bbc’,’ pride’,’ social’,’ holiday’,’ England,’ want’,’ know’.

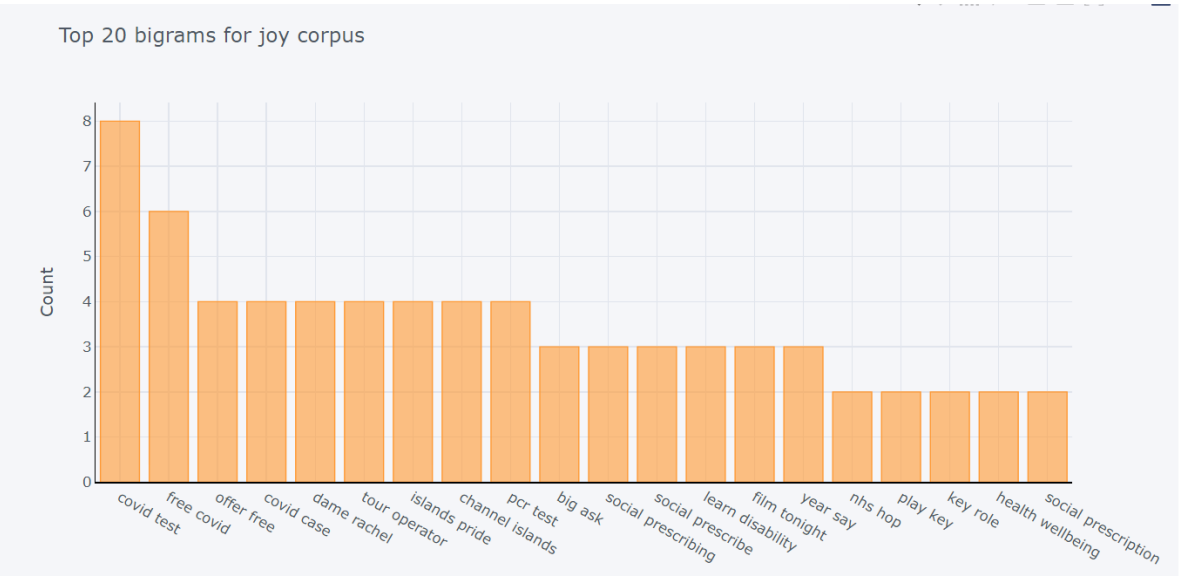
Bigrams from the corpora of disgust and joy are:

```
air pollution 0.17781346793621658%
covid rash 0.06883101984627739%
long covid 0.05735918320523115%
viral load 0.045887346564184923%
covid vaccine 0.04015142824366181%
pollution covid 0.04015142824366181%
covid infection 0.034415509923138694%
chronic cough 0.034415509923138694%
symptom covid 0.034415509923138694%
exposure air 0.034415509923138694%
college london 0.028679591602615576%
common symptom 0.028679591602615576%
persistent cough 0.028679591602615576%
lung disease 0.028679591602615576%
follow covid 0.028679591602615576%
rash covid 0.028679591602615576%
allergic reaction 0.028679591602615576%
study air 0.028679591602615576%
ongoing cough 0.022943673282092462%
covid symptom 0.022943673282092462%
```



Bigrams from the joy corpus are:

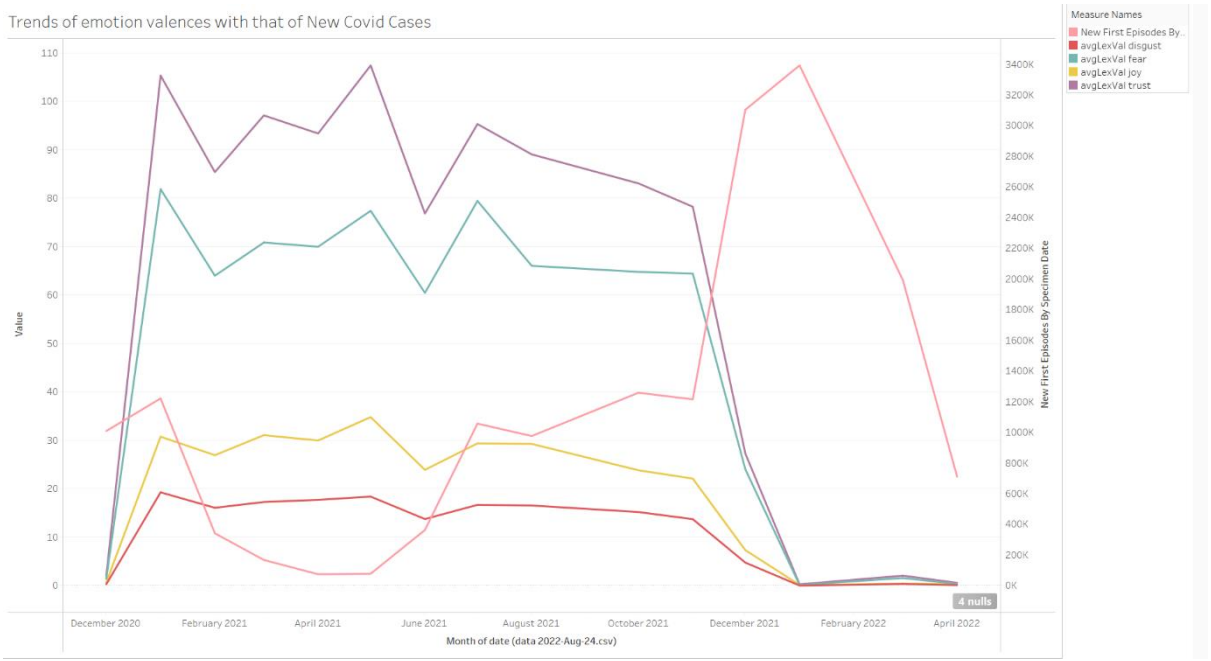
```
covid test 0.0540175557056043 2%
free covid 0.0405131667792032 4%
dame rachel 0.0270087778528021 6%
pcr test 0.0270087778528021 6%
channel islands 0.0270087778528021 6%
islands pride 0.0270087778528021 6%
covid case 0.0270087778528021 6%
tour operator 0.0270087778528021 6%
offer free 0.0270087778528021 6%
social prescribe 0.0202565833896016 2%
social prescribing 0.0202565833896016 2%
learn disability 0.0202565833896016 2%
big ask 0.0202565833896016 2%
year say 0.0202565833896016 2%
film tonight 0.0202565833896016 2%
social prescription 0.0135043889264010 8%
health wellbeing 0.013504388926401 08%
play key 0.013504388926401 08%
key role 0.013504388926401 08%
nhs hop 0.013504388926401 08%
```



Findings(Bigrams)

- The phrases for both these corpora are pretty unique.
- For disgust, the phrases like ‘covid-rash’,’ covid infection’, and chronic cough’ seem common.
- For Joy, the phrases are related to ‘social contexts’ like ‘social prescribe’,’health-wellbeing’,’covid-test’ and ‘PCR-test’ seem common.
- The counting scale on the disgust corpus graph hints toward a higher number than that for joy, indicating more texts and volume of the corpus.

The following graph was plotted between trends of emotions and covid news cases in 2021 using Tableau desktop.

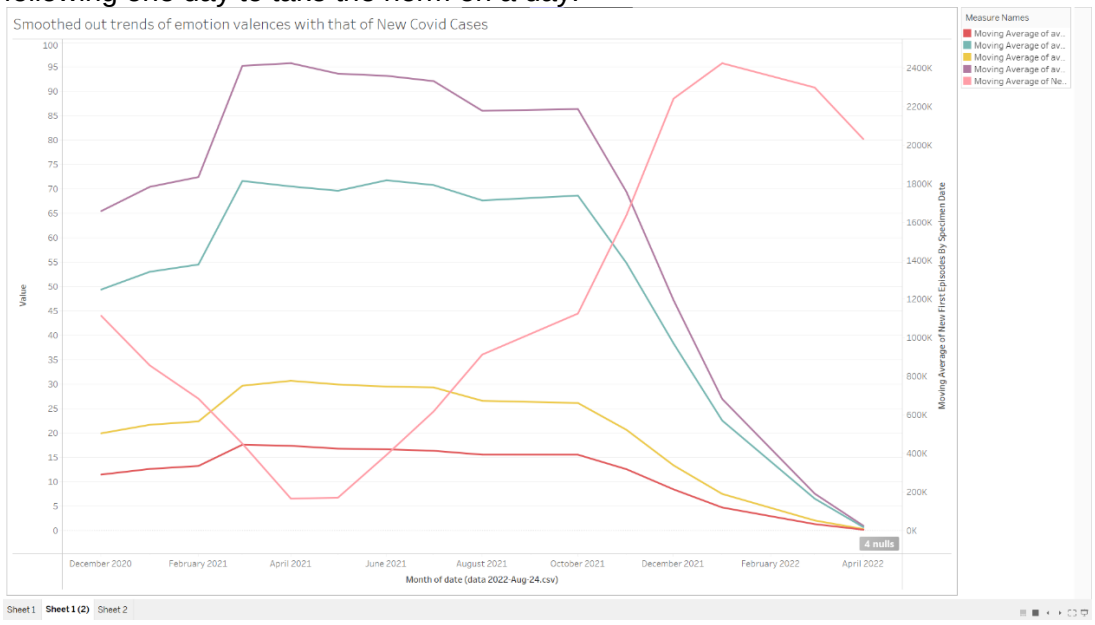


The trends of only the following emotions are considered due to their dominance in the texts scraped online (as given by 'Emotion Dynamics').

- From visual inspection, the dip in cases from January 2021 to April 2021 is accompanied by a drop in the trends of disgust and fear. However, the movements of joy and trust do rise during this time.
- The drastic increase of cases from November 2021 to December 2021 is accompanied by a rise in disgust and a slight rise in joy valences too. Fear valence rises drastically, and trust shows an increase but does achieve its highest overall.

For better trend visualisation, a graph of moving averages of the valences was plotted.

Moving averages takes into account the values of the previous two days and following one day to take the norm on a day.



According to this plot, the following was observed:

- Maintained trends in disgust and joy accompany the fall in cases in April 2021, but a rise in the valence trends of joy and trust.
- The rise of cases from June 2021 to December 2021 is accompanied by some upward trend in trust and fear but maintenance and downward trends in disgust and joy.

The file with the visualisation from tableau has been included in the appendix of the report.

The answers to the aims of the research are:

- The emotions for the news articles were successfully extracted using the emotion dynamics algorithm of the NRC lexicon.
- Post preprocessing of texts, a textual analysis of content of the articles was done using unigrams, bigrams and word clouds.
- A correlation between the emotional valences and covid cases was tested ehich resulted in 80 significant correlation values, on which the emotion category had a significant impact. This was demonstrated using ANOVA test.

4.3 Summary

In conclusion, this chapter presents the findings of the study using visualisations and discussions in the order they were conducted. It describes the correlational study results and the observations made during the analysis of the texts. It concludes by the research aims of the study being answered in pointwise fashion.

Chapter 5 Conclusion

This chapter consists of 3 parts, mainly a conclusion that wraps up the entire study conducted and its findings, the limitations encountered and the future work that can be done, keeping in mind the possibilities with the present algorithms.

5.1 Summary of the dissertation

In the first chapter, an introduction is done to the study to lay a background for the study included in this research. The second chapter, 'Literature Review,' was done using the Citespace software, which is free for citation analysis. The most cited papers were then used for assessing the latest research in sentiment and emotion analysis. The papers were clustered and hence analysed based on the keywords suggested by the software for each cluster.
The methodology comprises the third chapter and explains how the course of the study was decided on by taking the 'Research Onion' approach. It also explains the design of the correlational study conducted and finally progresses into the analysis, findings and discussion part of the study.
The study itself includes accumulating a new dataset of 7791 news headlines and their text. After pre-processing the texts via the coded pipeline in python, the dataset was assessed using the 'Emotion Dynamics' algorithm to analyse them and calculate their emotional valences.
Along with the emotional valence, the polarity values were also saved using the same algorithm. The categorised texts were combined to form a corpor based on the dominant emotion in the texts. The unigrams and bigrams in the corpora were compared using percentages.
Post emotion dynamics algorithm execution, the text was analysed using unigrams and bigrams based on the dominant emotion of the texts.
The correlation study included testing the correlation between the emotion valences and the covid cases, it concluded with 80 significant correlation values with different lag times.
The correlation study also included tests to see if the correlation mean values varied significantly with lag(in days) or emotion. The test was conclusive, and the results were then discussed in the report.
Post the correlation study, the analysis was done with texts of emotions with significant correlation coefficient value with covid cases. The analysis was again done using unigrams and bigrams.

5.2 Research Contributions & Limitations

In conclusion, the dataset presented can be used for several research purposes, and the study concludes how the texts can convey different emotions based on a varying choice of words used in the articles published for news dissemination.
It can be used for government policy drafting keeping in mind the public sentiments and emotions. It treads towards the solving the crisis spread in certain parts of the world due to negligence of the impact language can have on the entropy in the society. The algorithm can make help us manage mental stress and other issues amongst masses in real time and can help deal with their trauma by completely eradicating it in the first place.

Limitations

The study was conducted by scraping websites for news headlines and texts. However, not all texts for the headlines found by the Google API were scraped due to the scraping restrictions on the websites or the unavailability of outdated articles. These restrictions are studied well in the social sciences and have some state-of-the-art solutions associated with them. (Luscombe, Dick and Walby, 2022) These have not been included while coding for scraping the texts of headlines.
The emotion valences have been calculated using the NRC lexicon, which has cons. According to this recent paper by (Zad, Jimenez and Finlayson, 2021), the NRC lexicon has some pejorative, inaccurate and non-sensical labels for words which in turn leads to a questionable emotional analysis of texts. It also mentions and corrects POS tags which are ambiguous in the lexicon and hence lead to wrong results since the words are context-dependent most of the time.
The emotion dynamics algorithm uses the NRC lexicon, which does not calculate the neutral valence of the article. This also limits our analysis by only including positive and negative sentiment valences of the texts.
Comparisons between the polarity classification of texts by text blob and NRC lexicon have been made, which do not yield any significant results, as far as the top most frequent words in the classified texts are considered. There can be some differences if further the frequency of all the words in the corpora is considered.
The trends here have been analysed only, including the news articles, and the inclusion of any other factors is not considered. Some major crises or announcements that may trigger a certain rise in covid cases have not been taken into account.
Social media is undoubtedly the most used platform today, and many people rely on it for news also. Hence the trends in covid cases could be primarily affected by public opinions gaining momentum on these platforms. It is why many have based their research on sentimentally analysing texts from social media. Many have analysed the texts of discussion forums found on news media sites. These texts are collected to give a more profound understanding of what prevails in the public space on the web and hence work better towards the emotional analysis of texts. (Hussain *et al.*, 2021)(Abd-Alrazaq *et al.*, 2020)
The time series element of the data is an essential factor to consider while modelling the trends in covid cases. There have been multiple attempts at time-aware modelling for stock price predictions, but here the time-series trends are not studied and do not form the focus of this study. (Feng *et al.*, 2021)
The keyword ‘covid’ has been used to extract headlines and thereafter the texts, which may limit the search for articles as other headlines (especially that of the falling economy) of emotional significance can affect the mental state of the public too. More keywords might broaden the domain of texts being analysed.
The headlines have only been taken and analysed that were published in the United Kingdom. Any generalisations about these trends can only be made if a similar study is conducted as a part of further research for other countries and compared to improve the quality of public discourse texts.
The authenticity of texts scraped using the headlines has not been put to test. There have been many works to weed out fake news using machine and deep learning approaches for polarity (sentiment) analysis. (Samonte, 2018)

5.3 Future research and development

Further research in this field could include analysing the emotional valences throughout the article length to analyse the abrupt rise and falls in the polarity closely. This can be done using the emotion dynamics code mentioned on the GitHub website.
Time series analysis for both covid cases and emotional valences can be very interesting to explore the predictions for future cases. This can aid us in improving readiness for future pandemics or peaks and ebbs during these unfortunate times.
Trend analysis can be incorporated to note essential periods in the explored timeline accurately. This is usually done for stock prices but has been explored for covid cases due to the extremity of the pandemic.
Topic modelling is another task of natural language processing that can be attempted on this new dataset of 7791 articles, which is capable of giving us sizeable results.
Comparison of performance with other lexicons like EmoSentic Net, DepecheMood, and Topic Based DepecheMood will be part of the research with this dataset.
Currently, I'm exploring the CrystalFeel algorithm for sentiment analysis of texts. The algorithm is made for tweets, but a comparative study on news articles with other algorithms like the one used in the study seems quite intriguing.

References

Abd-Alrazaq, A. *et al.* (2020) 'Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study', *Journal of Medical Internet Research*, 22(4), p. e19016. Available at: https://doi.org/10.2196/19016.

Agarwal, A. (2020) 'Sentiment Analysis of Financial News', in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 312–315. Available at: https://doi.org/10.1109/CICN49253.2020.9242579.

Al-Moslmi, T. *et al.* (2018) 'Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis', *Journal of Information Science*, 44(3), pp. 345–362. Available at: https://doi.org/10.1177/0165551516683908.

Amal I. Khalil, Rawan E.Nasr, and Rahaf E.Nahr (2020) 'Relationship between Stress, Immune System, and Pandemics of Coronaviruses' COVID19: Updates Narrative Review'.

Amirkhan, J.H. (2021) 'Stress overload in the spread of coronavirus', *Anxiety, Stress, & Coping*, 34(2), pp. 121–129. Available at: https://doi.org/10.1080/10615806.2020.1824271.

Aygun, I., Kaya, B. and Kaya, M. (2021) 'Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic with Deep Learning', *IEEE Journal of Biomedical and Health Informatics* [Preprint]. Available at: https://doi.org/10.1109/JBHI.2021.3133103.

Baj-Rogowska, A. (2021) 'Mapping of the covid-19 vaccine uptake determinants from

mining twitter data', *IEEE Access*, 9, pp. 134929–134944. Available at: https://doi.org/10.1109/ACCESS.2021.3115554.

Bhat, M. *et al.* (2020) 'Sentiment analysis of social media response on the Covid19 outbreak', *Brain, Behavior, and Immunity*, 87, pp. 136–137. Available at: https://doi.org/10.1016/j.bbi.2020.05.006.

Bisiada, M. (2022) 'Discourse and Social Cohesion in and After the Covid-19 Pandemic', *Media and Communication*, 10(2), pp. 204–213. Available at: https://doi.org/10.17645/mac.v10i2.5150.

Bruce, P., Bruce, A. and Gedeck, P. (2020) *Practical Statistics for Data Science*. O'Reilly Media, Inc. Available at: https://learning.oreilly.com/library/view/practical-statistics-for/9781492072935/ch03.html (Accessed: 12 September 2022).

Cambria, E. (2016) *Affective Computing and Sentiment Analysis; Affective Computing and Sentiment Analysis*. Available at: https://doi.org/10.1109/MIS.2016.31.

Chen, C. (2006) 'CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature', *Journal of the American Society for Information Science and Technology*, 57(3), pp. 359–377. Available at: https://doi.org/10.1002/asi.20317.

Chepurnaya, A. (2021) 'Modeling public perception in times of crisis: discursive strategies in Trump's COVID-19 discourse', *Critical Discourse Studies*, 0(0), pp. 1–18. Available at: https://doi.org/10.1080/17405904.2021.1990780.

Cherry, K. (2020) 'How Does the James-Lange Theory Account for Emotions?', *Verywell Mind*. Available at: https://www.verywellmind.com/what-is-the-james-lange-theory-of-emotion-2795305 (Accessed: 11 September 2022).

Chiril, P. *et al.* (2022) 'Emotionally Informed Hate Speech Detection: A Multi-target Perspective', *Cognitive Computation*, 14(1), pp. 322–352. Available at: https://doi.org/10.1007/s12559-021-09862-5.

Creswell, J. (2018) *Research Design*. SAGE Publications,Inc.

Daily, S.B. *et al.* (2017) 'Chapter 9 - Affective Computing: Historical Foundations, Current Applications, and Future Trends', in M. Jeon (ed.) *Emotions and Affect in Human Factors and Human-Computer Interaction*. San Diego: Academic Press, pp. 213–231. Available at: https://doi.org/10.1016/B978-0-12-801851-4.00009-4.

Ekman, P. (2009) 'Darwin's contributions to our understanding of emotional expressions', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3449–3451. Available at: https://doi.org/10.1098/rstb.2009.0189.

Fehle, J., Schmidt, T. and Wolff, C. (2021) 'Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques', p. 18.

Feng, F. *et al.* (2021) 'Time horizon-aware modeling of financial texts for stock price prediction', *Proceedings of the Second ACM International Conference on AI in Finance*,

pp. 1–8. Available at: https://doi.org/10.1145/3490354.3494416.

Finlay, J.M. *et al.* (2021) 'Coping During the COVID-19 Pandemic: A Qualitative Study of Older Adults Across the United States', *Frontiers in Public Health*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.643807 (Accessed: 11 September 2022).

Guidi, J. *et al.* (2021) 'Allostatic Load and Its Impact on Health: A Systematic Review', *Psychotherapy and Psychosomatics*, 90(1), pp. 11–27. Available at: https://doi.org/10.1159/000510696.

Gupta, I. and Joshi, N. (2018) 'Tweet normalization: A knowledge based approach', *2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017*, 2018-January, pp. 157–162. Available at: https://doi.org/10.1109/ICTUS.2017.8285996.

Hipson, W.E. and Mohammad, S.M. (2021) 'Emotion dynamics in movie dialogues', *PLOS ONE*, 16(9), p. e0256153. Available at: https://doi.org/10.1371/journal.pone.0256153.

Hussain, A. *et al.* (2021) 'Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study', *Journal of Medical Internet Research*, 23(4), p. e26627. Available at: https://doi.org/10.2196/26627.

Lin, H.Y. and Moh, T.S. (2021) 'Sentiment analysis on COVID tweets using COVID-Twitter-BERT with auxiliary sentence approach', in. *Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference*. Available at: https://doi.org/10.1145/3409334.3452074.

Liu, Q. *et al.* (2020) 'Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach', *Journal of Medical Internet Research*, 22(4), p. e19118. Available at: https://doi.org/10.2196/19118.

Luscombe, A., Dick, K. and Walby, K. (2022) 'Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences', *Quality & Quantity*, 56(3), pp. 1023–1044. Available at: https://doi.org/10.1007/s11135-021-01164-0.

Lwin, M.O. *et al.* (2020) 'Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends'. Available at: https://doi.org/10.2196/19447.

MacFarland, T.W. and Yates, J.M. (2016) 'Spearman's Rank-Difference Coefficient of Correlation', in T.W. MacFarland and J.M. Yates (eds) *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Cham: Springer International Publishing, pp. 249–297. Available at: https://doi.org/10.1007/978-3-319-30634-6_8.

Medhat, W., Hassan, A. and Korashy, H. (2014) 'Sentiment analysis algorithms and

applications: A survey', *Ain Shams Engineering Journal*, 5(4), pp. 1093–1113. Available at: https://doi.org/10.1016/J.ASEJ.2014.04.011.

Mohammad, S. and Turney, P. (2010) 'Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon', in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, CA: Association for Computational Linguistics, pp. 26–34. Available at: https://aclanthology.org/W10-0204 (Accessed: 30 August 2022).

Nandwani, P. and Verma, R. (2021) 'A review on sentiment analysis and emotion detection from text', *Social Network Analysis and Mining*, 11(1), pp. 81–81. Available at: https://doi.org/10.1007/S13278-021-00776-6.

Park, S. *et al.* (2021) 'Customer sentiment analysis with more sensibility', *Engineering Applications of Artificial Intelligence*, 104, p. 104356. Available at: https://doi.org/10.1016/j.engappai.2021.104356.

Pastor, C.K. (2020) 'Sentiment Analysis of Filipinos and Effects of Extreme Community Quarantine Due to Coronavirus (COVID-19) Pandemic'. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.3574385.

Plutchik, R. (2001) 'The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice', *American Scientist*, 89(4), pp. 344–350.

Price, P.C., Chiang, I.-C.A. and Jhangiani, R. (2014) *Research Methods in Psychology*. BCcampus, BC Open Textbook Project.

Qanita Bani Baker *et al.* (2019) 'Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets'.

Raamkumar, A.S., Tan, S.G. and Wee, H.L. (2020) 'Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook During the COVID-19 Pandemic in Early 2020: Cross-Country Comparison', *Journal of Medical Internet Research*, 22(5), p. e19334. Available at: https://doi.org/10.2196/19334.

Rai, B., Kasturi, M. and Huang, C. yu (2018) 'Analyzing stock market movements using news, tweets, stock prices and transactions volume data for APPLE (AAPL), GOOGLE (GOOG) and SONY (SNE)', *ACM International Conference Proceeding Series*, pp. 109–112. Available at: https://doi.org/10.1145/3243250.3243263.

Ravi, K. and Ravi, V. (2015) 'A survey on opinion mining and sentiment analysis: Tasks, approaches and applications', *Knowledge-Based Systems*, 89, pp. 14–46. Available at: https://doi.org/10.1016/j.knosys.2015.06.015.

Roe, C. *et al.* (2021) 'Public Perception of SARS-CoV-2 Vaccinations on Social Media: Questionnaire and Sentiment Analysis', *International Journal of Environmental Research and Public Health*, 18(24), p. 13028. Available at: https://doi.org/10.3390/ijerph182413028.

Samonte, M.J.C. (2018) 'Polarity Analysis of Editorial Articles towards Fake News Detection', in *Proceedings of the 2018 International Conference on Internet and e-Business - ICIEB '18*. *the 2018 International Conference*, Singapore, Singapore: ACM Press, pp. 108–112. Available at: https://doi.org/10.1145/3230348.3230354.

Samuel, J. *et al.* (2020) 'COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification', *Information*, 11(6), p. 314. Available at: https://doi.org/10.3390/info11060314.

Saunders, M. *et al.* (2019) '"Research Methods for Business Students" Chapter 4: Understanding research philosophy and approaches to theory development', in, pp. 128–171.

Snyder, P.J. *et al.* (2010) 'Charles Darwin's Emotional Expression "Experiment" and His Contribution to Modern Neuropharmacology', *Journal of the History of the Neurosciences*, 19(2), pp. 158–170. Available at: https://doi.org/10.1080/09647040903506679.

Steven Bird, Evan Klein, and Edward Loper (2009) *Natural Language Processing with Python*. Available at: https://learning.oreilly.com/library/view/natural-language-processing/9780596803346/ (Accessed: 9 September 2022).

Strongman, K.T. (2003) *The Psychology of Emotion*. Fifth. John Wiley & Sons Ltd.

Taj, S., Shaikh, B.B. and Fatemah Meghji, A. (2019) 'Sentiment analysis of news articles: A lexicon based approach', *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, iCoMET 2019* [Preprint]. Available at: https://doi.org/10.1109/ICOMET.2019.8673428.

Talpur, B.A. and O'Sullivan, D. (2020) 'Cyberbullying severity detection: A machine learning approach', *PLOS ONE*, 15(10), p. e0240924. Available at: https://doi.org/10.1371/journal.pone.0240924.

Tshimula, J.M., Chikhaoui, B. and Wang*, S. (2022) 'COVID-19: Detecting depression signals during stay-at-home period', *Health Informatics Journal*, 28(2), p. 14604582221094932. Available at: https://doi.org/10.1177/14604582221094931.

Viegas, F. *et al.* (2020) 'Exploiting semantic relationships for unsupervised expansion of sentiment lexicons', *Information Systems*, 94, p. 101606. Available at: https://doi.org/10.1016/j.is.2020.101606.

Xue, J. *et al.* (2020) 'Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter', *PLOS ONE*, 15(9), p. e0239441. Available at: https://doi.org/10.1371/journal.pone.0239441.

Yu, X. *et al.* (2020) 'Sentiment Analysis for News and Social Media in COVID-19 ACM Reference format', *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Emergency Management using GIS* [Preprint]. Available at: https://doi.org/10.1145/3423333.

Zad, S., Jimenez, J. and Finlayson, M. (2021) 'Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon', in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). ACL-IJCNLP-WOAH 2021*, Online: Association for Computational Linguistics, pp. 102–113. Available at: https://doi.org/10.18653/v1/2021.woah-1.11.

APPENDIX A: ETHICAL APPROVAL

LETTER OF CONFIRMATION

Applicant: Ms. Shivangi Dubey

Project Title: Sentiment and Emotional Analysis of News Articles using a lexicon-based approach with a subsequent correlation analysis with the new Covid- 19 cases data collected per day in the UK.

Reference: 37758-NER-Jun/2022- 40079-1

Dear Ms. Shivangi Dubey

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Good luck with your research!

Kind regards,

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

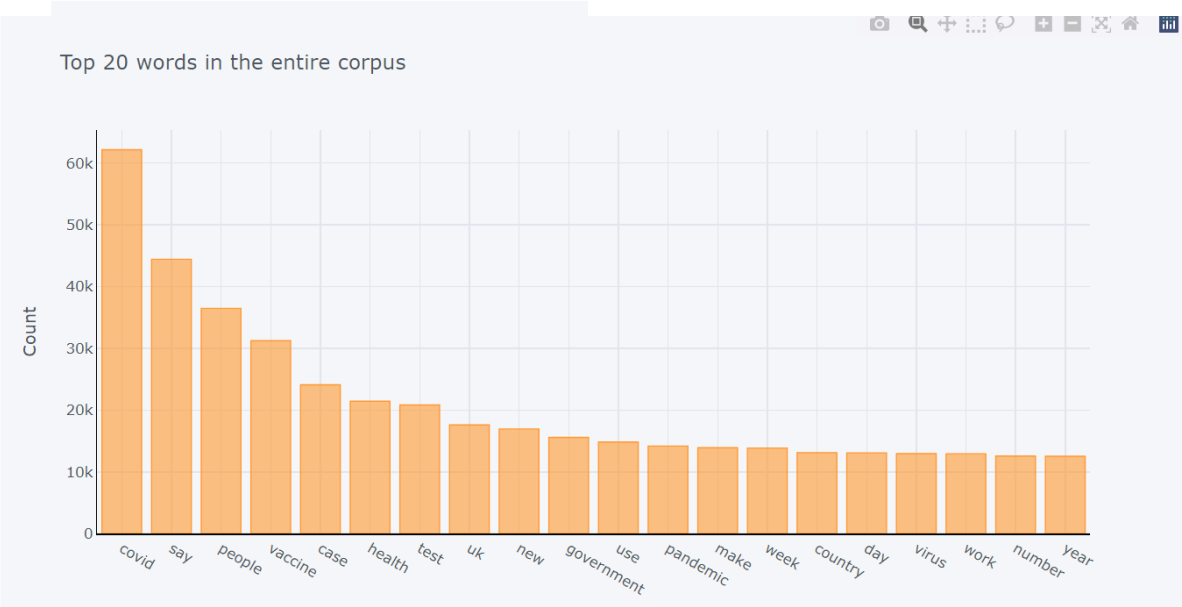
Brunel University London

APPENDIX B

To analyse the words in the texts, word clouds were plotted. For all the texts combined, the result is included below.

The top words are

```
Frequent Words in entire text
covid 62155
say 44435
people 36479
vaccine 31249
case 24096
health 21442
test 20842
uk 17606
new 16948
government 15585
use 14830
pandemic 14176
make 13918
week 13867
country 13117
day 13074
virus 12965
work 12935
number 12571
year 12526
```

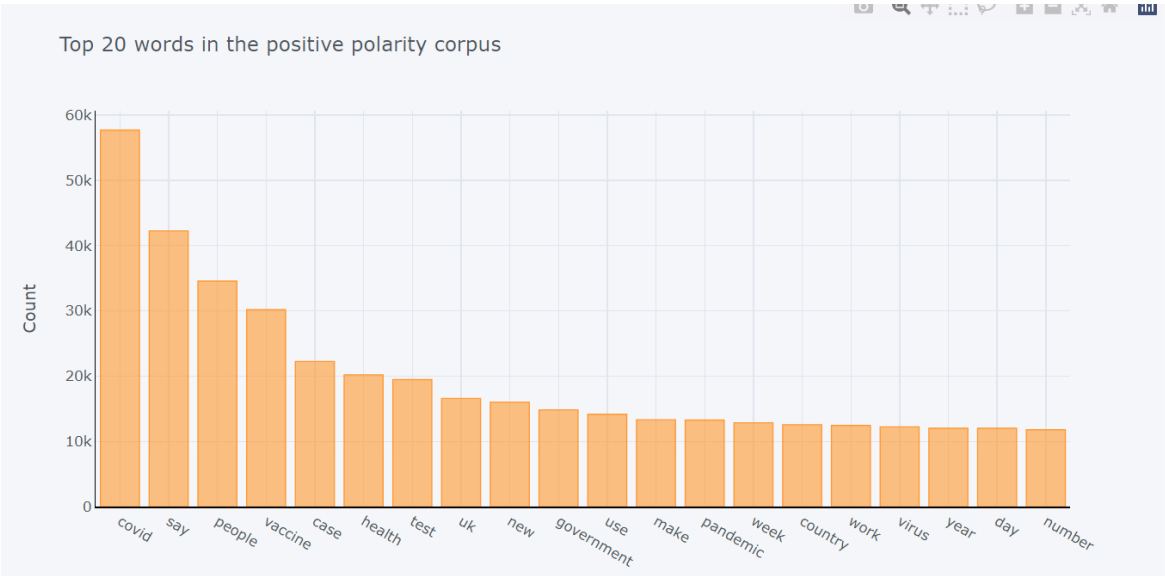


The texts were initially classified as positive and negative by the text blob package and later by the NRC lexicon.
The results of the text blob are:

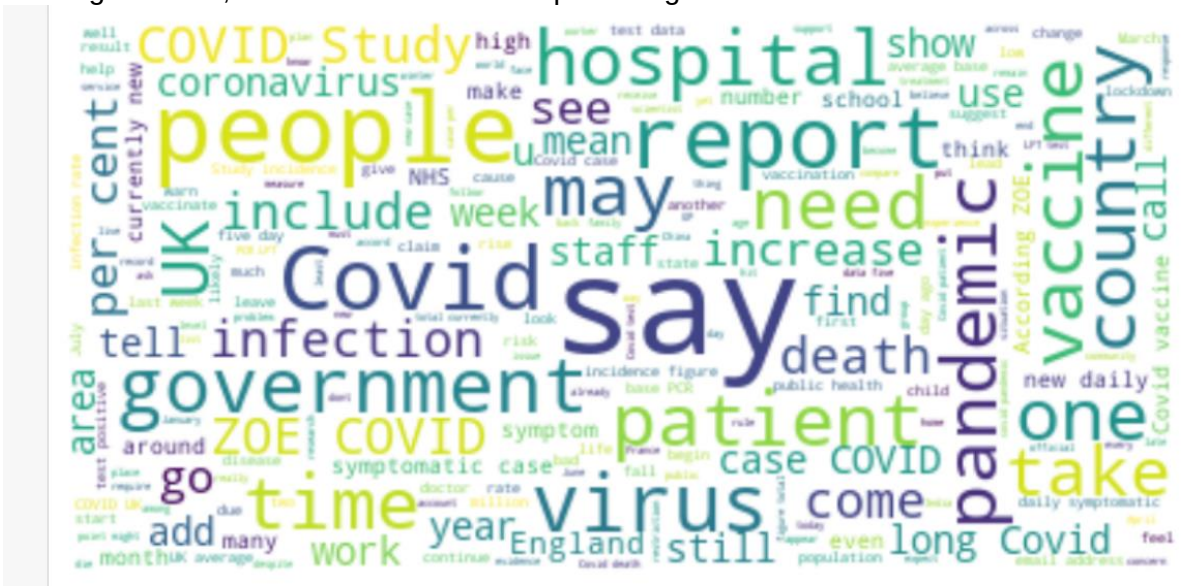
Negative texts: 466
Neutral Texts: 10
Positive Texts: 7315

The top frequent unigrams as per the polarity are:
For positive polarity, the word clouds and frequent unigrams are:

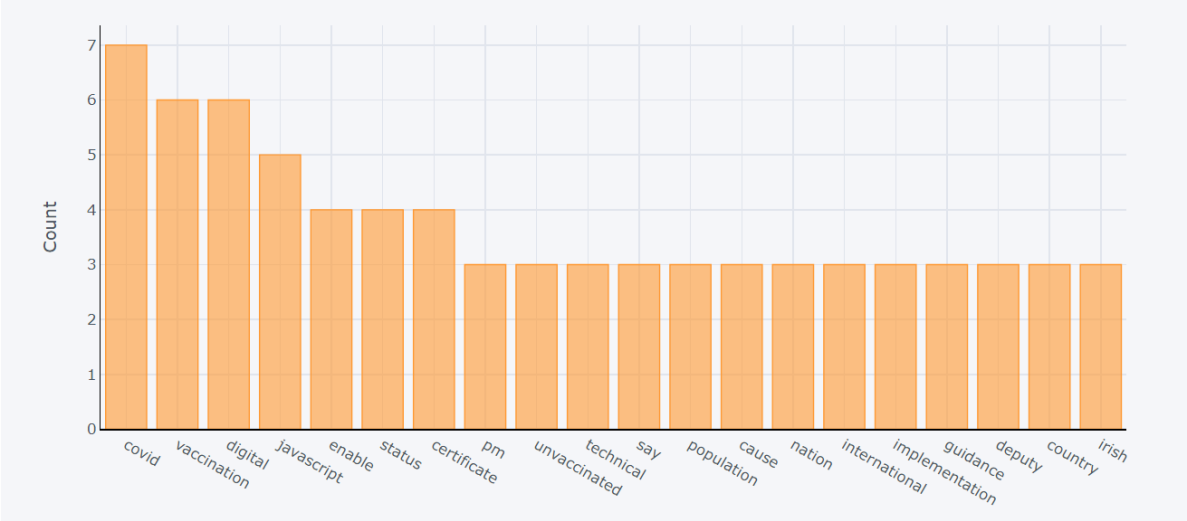
Frequent Words in positive polarity text
covid 57690
say 42255
people 34572
vaccine 30183
case 22257
health 20206
test 19479
uk 16587
new 16030
government 14846
use 14164
make 13322
pandemic 13278
week 12872
country 12564
work 12460
virus 12235
year 12032
day 12027
number 11794



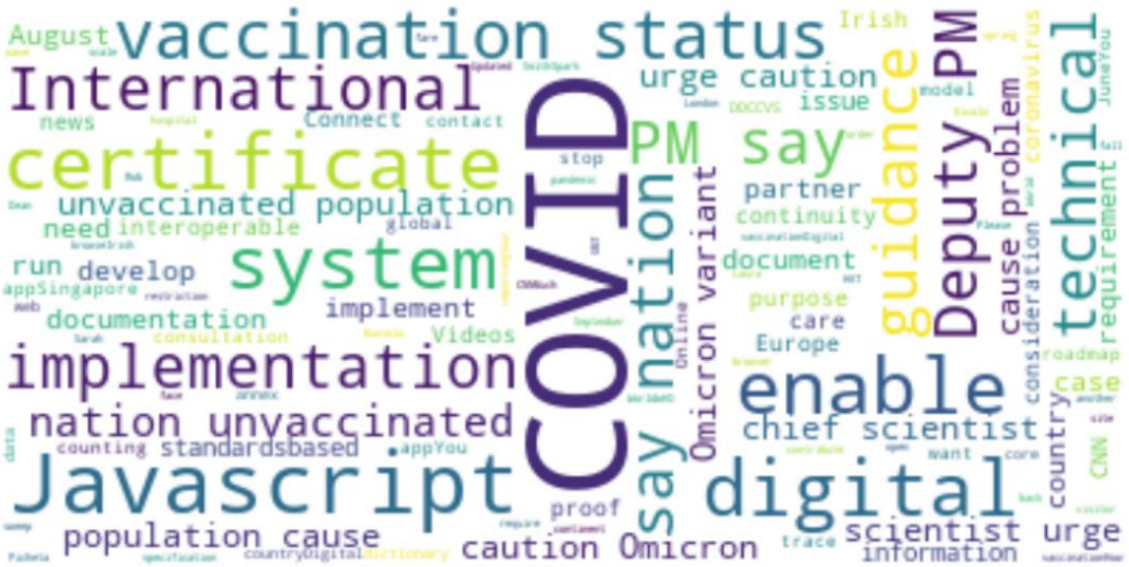
For negative text, the word clouds and frequent unigrams are:



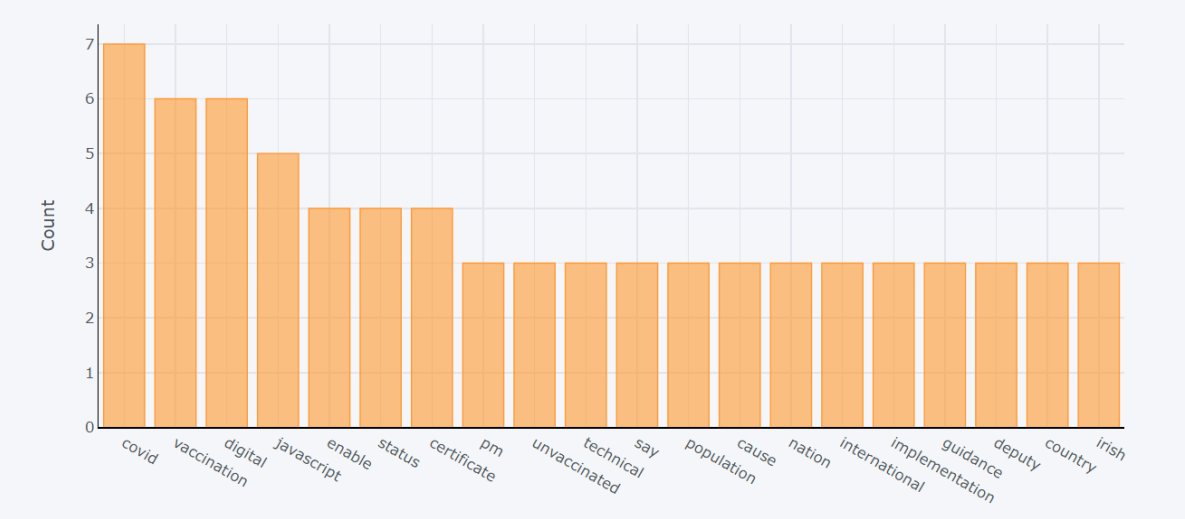
```
Frequent Words in negative polarity text
covid 4458
say 2177
people 1907
case 1837
test 1363
health 1236
vaccine 1066
day 1047
uk 1019
week 995
data 970
new 918
death 910
pandemic 897
infection 826
rate 785
number 777
report 743
government 739
virus 730
```



For neutral text, the word clouds and frequent unigrams are:



Frequent Words in neutral polarity text
covid 7
digital 6
vaccination 6
javascript 5
enable 4
certificate 4
status 4
country 3
guidance 3
technical 3
implementation 3
international 3
irish 3
deputy 3
pm 3
say 3
nation 3
unvaccinated 3
population 3
cause 3

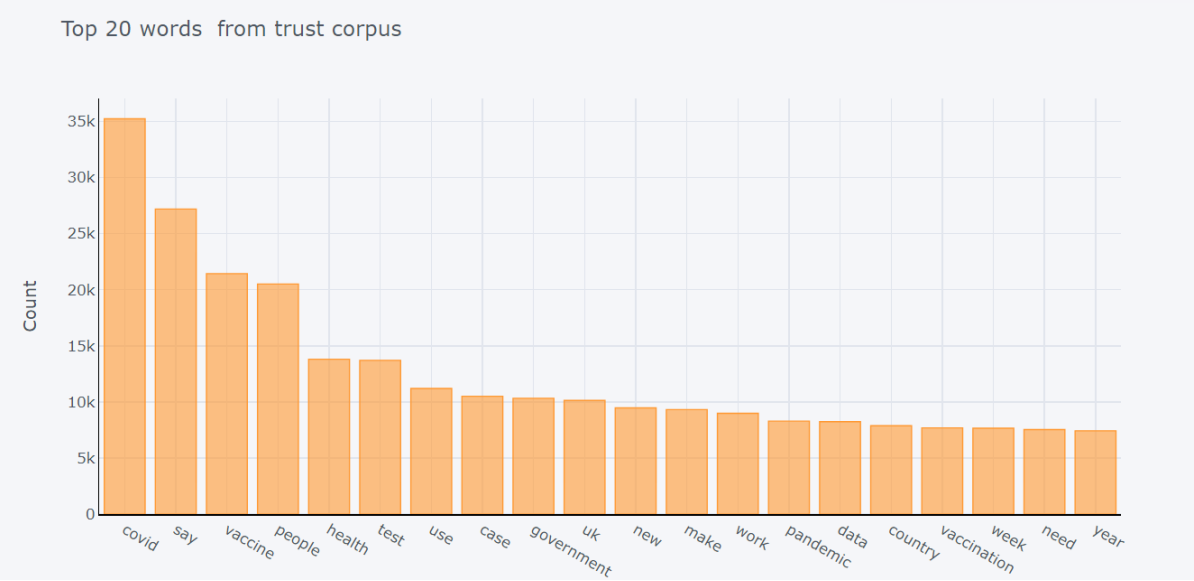


Post emotion analysis, the overall_emo frequencies of the texts are:

```
avgLexVal_anger 15
avgLexVal_anti 803
avgLexVal_disgust 4
avgLexVal_fear 1920
avgLexVal_joy 7
avgLexVal_sad 205
avgLexVal_trust 4837
```

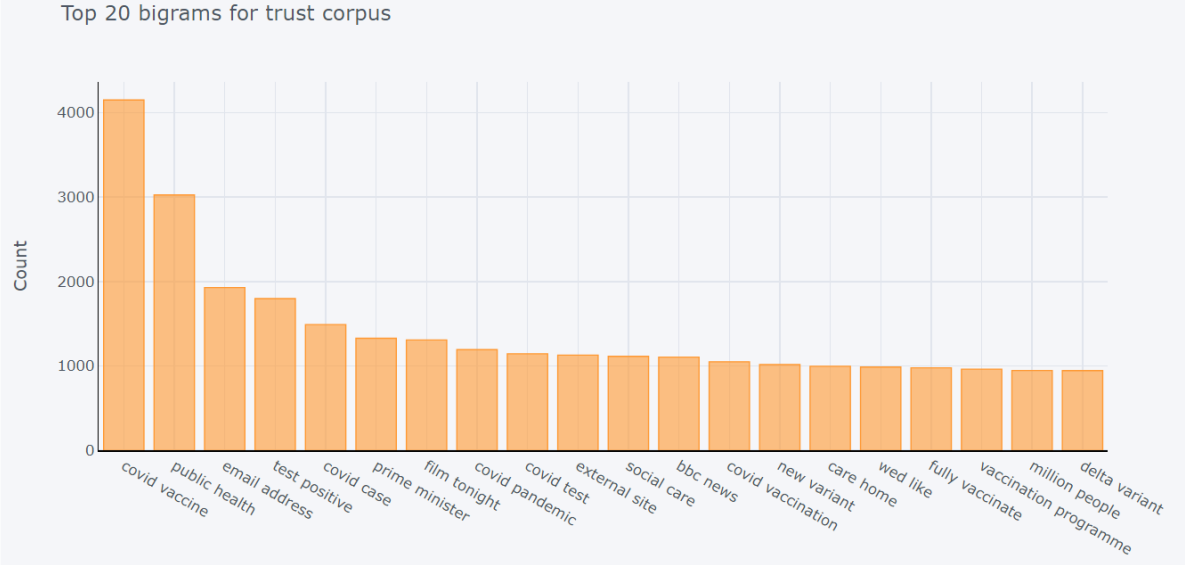
Since the trust and fear emotions are dominant, we analyse those texts.

```
Frequent Words trust text
covid 35231
say 27199
vaccine 21438
people 20515
health 13815
test 13715
use 11212
case 10517
government 10334
uk 10138
new 9476
make 9328
work 8991
pandemic 8298
data 8248
country 7883
vaccination 7698
week 7670
need 7540
year 7435
```

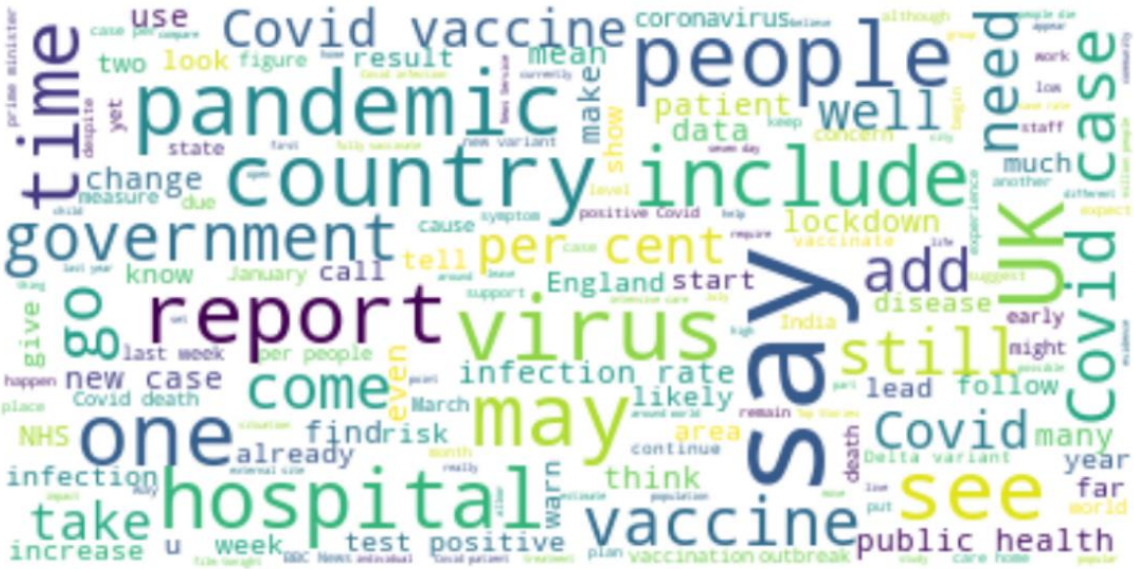


From the trust texts, the most frequent bigrams are:

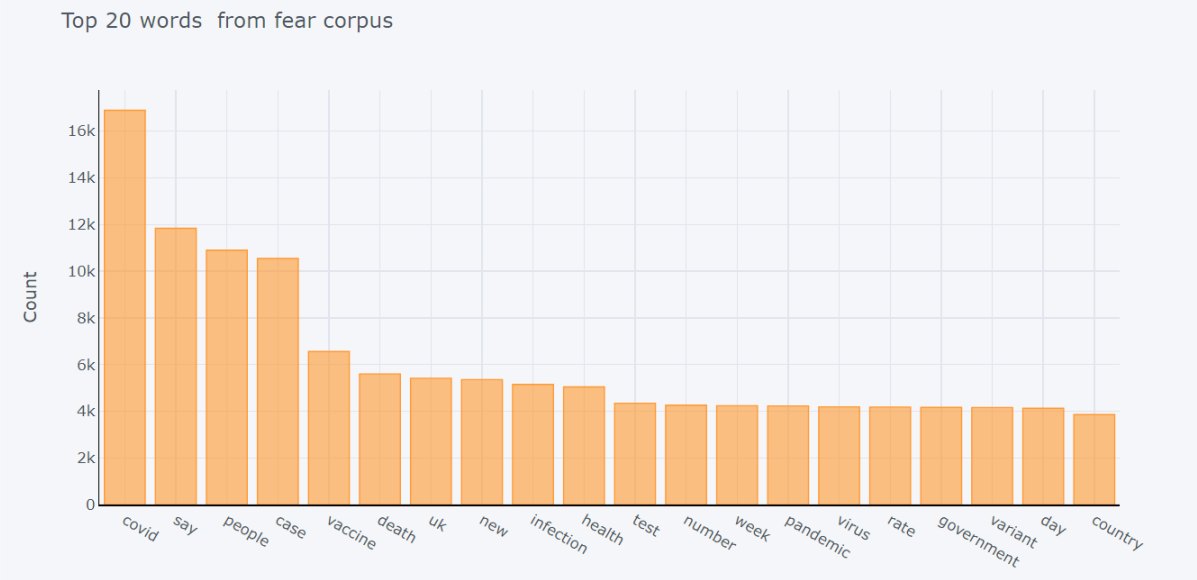
```
Frequent Bigrams for trust corpus
covid vaccine 4148
public health 3024
email address 1929
test positive 1798
covid case 1491
prime minister 1329
film tonight 1309
covid pandemic 1194
covid test 1145
external site 1130
social care 1115
bbc news 1106
covid vaccination 1051
new variant 1017
care home 998
wed like 989
fully vaccinate 979
vaccination programme 964
million people 948
delta variant 946
```



For fear, the most frequent words are:

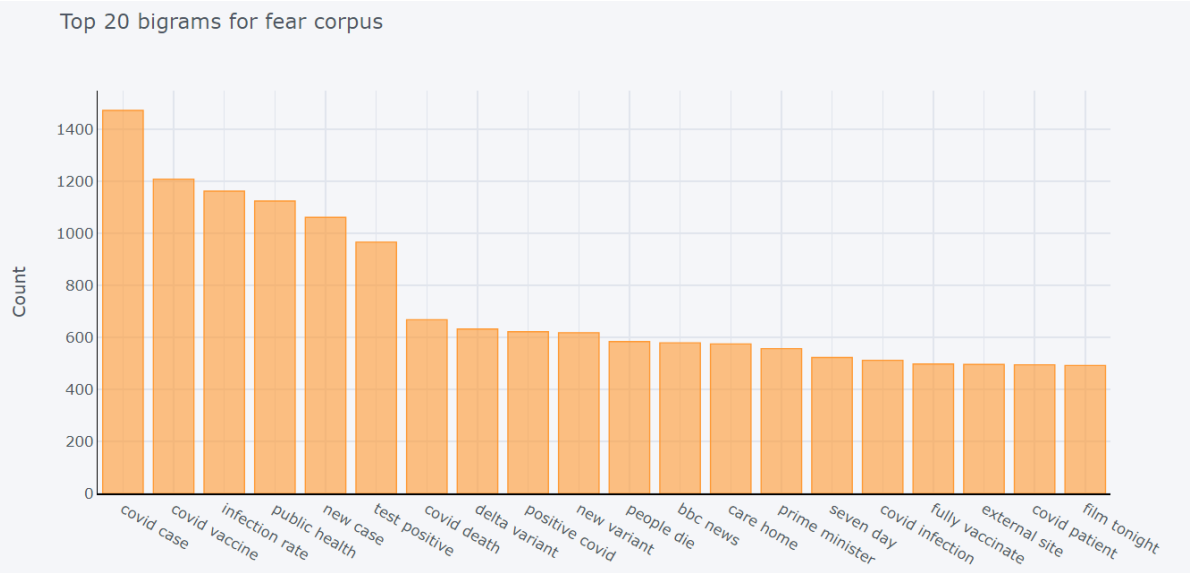


Frequent Words fear text
covid 16888
say 11834
people 10899
case 10543
vaccine 6559
death 5600
uk 5415
new 5359
infection 5149
health 5044
test 4343
number 4262
week 4238
pandemic 4228
virus 4191
rate 4183
government 4166
variant 4164
day 4130
country 3858



The most frequent bigrams from the texts of fear are:

```
most frequent words for fear corpus
covid case 1472
covid vaccine 1208
infection rate 1162
public health 1124
new case 1061
test positive 966
covid death 668
delta variant 632
positive covid 622
new variant 618
people die 584
bbc news 579
care home 575
prime minister 557
seven day 523
covid infection 512
fully vaccinate 498
external site 496
covid patient 495
film tonight 492
```

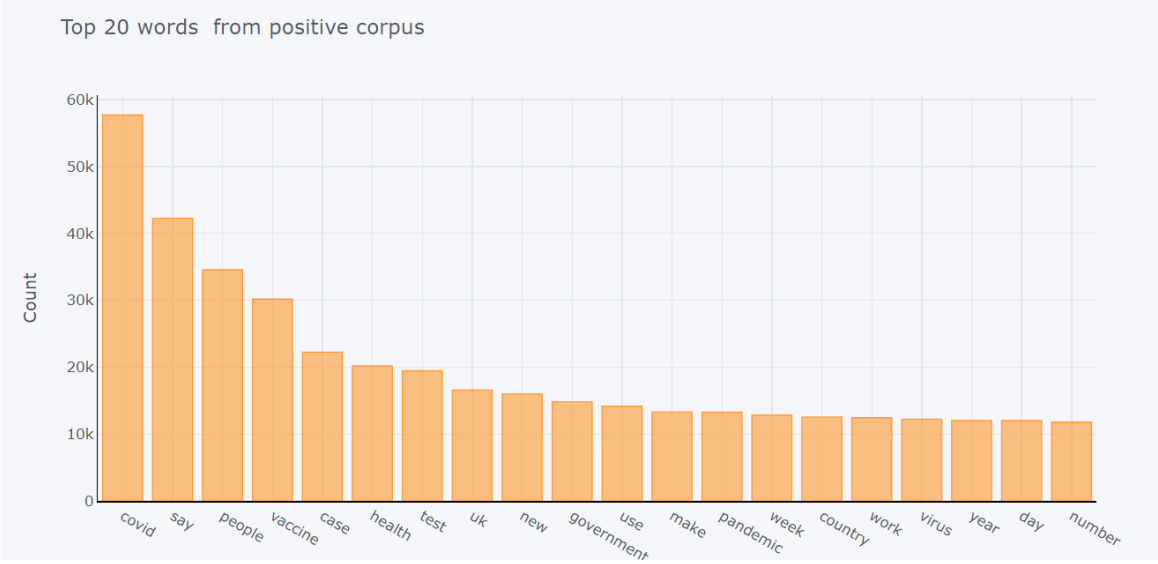


The results from the NRC Emotion Dynamics implementation are:

```
avgLexVal_neg 1642
avgLexVal_pos 6149
```

The frequent words from positive texts are

```
Frequent Words positive text
covid 57690
say 42255
people 34572
vaccine 30183
case 22257
health 20206
test 19479
uk 16587
new 16030
government 14846
use 14164
make 13322
pandemic 13278
week 12872
country 12564
work 12460
virus 12235
year 12032
day 12027
number 11794
```



The negative texts' frequent words are:

```
Frequent Words negative text
covid 4458
say 2177
people 1907
case 1837
test 1363
health 1236
vaccine 1066
day 1047
uk 1019
week 995
data 970
new 918
death 910
pandemic 897
infection 826
rate 785
number 777
report 743
government 739
virus 730
```