Literature Review for Covid19 Tracking

Description:
In this document, I will keep a record of papers I have reviewed. For each paper, I will record the Title and Author for reference; as well as the key takeaways that a) I think are interesting, b) I plan on implementing for my own research or c) I think I might be adding in the future. Some of the points introduced here are copy pasted from the original article. Some are paraphrased to add additional context. In any case, it should be obvious whether something is copy pasted or reworded.

Note: In most(all?) articles, I am most interested in *methodology and data processing* over actual results

1. Design and analysis of a large-scale COVID-19 tweets dataset

Author(s): Rabindra Lamsal

Keywords: Social computing, crisis computing, sentiment analysis, network analysis, twitter data

Summary/Takeaways:
This paper basically ensures that we are on the right track at least with what we have so far. It goes over cov19 tweet dataset being big, usefulness of geotweet dataset, proportion of tweets with geo tag, and necessary compute power to perform processing. Some important points are as follows:

Twitter data is good for situational awareness during crises. It is an always-on data source, so we can collect data for immediate response.

As crisis unfolds, social media platforms such as Facebook and Twitter become an active source of information because these platforms break the news faster than official news channels and emergency response agencies

The article says that in Twitter people: “share their safety status, querying about their loved ones’ safety status, and report ground level scenarios of the event”. This is important because when we filter out covid tweets from other events (e.g.: US presidential election), we know we can try keywords that involve covid terms, queries of loved ones, and safety keywords. Example: #lockdown, #mother, #safe or something like this…

social media data can be analyzed and processed to extract situational information that can be further used to derive actionable intelligence for an effective response to the crisis

Twitter sets limits on the number of requests that can be made to its API. Its filtered stream endpoint has a rate limit of 450 requests/15-minutes per app., which is why the maximum number of tweets that can be fetched in 24 hours is just above 4 million.

Section 2.2 Sentiment analysis has some useful links to other papers about covid19 and mental health/emotions.

Infrastructure: “The collection of tweets is a small portion of the dataset design. The other tasks include filtration of geo-tagged tweets and computation of sentiment score for each captured tweet, all that in real-time.” Our study does this too, but not in real time.

Computation: The computation of sentiment score for each captured tweet requires the VM to constitute powerful enough CPUs to avoid a bottleneck scenario. Every information gathered to this point needs to be stored on a database, which necessitates a disk with excellent performance. Summing up, a cloud-based VM is required to automate all these tasks. I'm not sure how well Compute Canada does all of this; I assume if used properly it should run much faster than locally; it also transfers the computation load from local to remote.

Keywords for filtering: corona, #corona, coronavirus, #coronavirus, covid, #covid, covid19, #covid19, covid-19, #covid-19, sarscov2, #sarscov2, sars cov2, sars cov 2, covid 19, #covid 19, #ncov, ncov, #ncov2019, ncov2019, 2019-ncov, #2019-ncov, #2019ncov, 2019ncov, pandemic, #pandemic, quarantine, #quarantine, flatten the curve, flattening the curve, #flatteningthecurve, #flattenthecurve, hand sanitizer, #handsanitizer, #lockdown, lockdown, social distancing, #socialdistancing, work from home, #workfromhome, working from home, #workingfromhome, ppe, n95, #ppe, #n95

Preprocessing before Filtering: Hash symbol (#), mention symbol (@), URLs, extra spaces, and paragraph breaks are cleaned. Punctuations, emojis, and numbers are included. Advance-level preprocessing, such as (i) correction of incorrectly spelt words, (ii) conversion of abbreviations to their original forms, are bypassed to avoid analysis bottleneck.

Out of 310 million tweets, 141k tweets (0.045%) were found to be geo-tagged.

Common terms during periods with high negative sentiment:

Table 3 Trending unigrams and bigrams

Date	score[a]	Unigrams[b]	Bigrams
May 28, 2020	−0.03	deaths, people, trump, pandemic, cases, world, US, virus, health, UK, death, government, china, police	nursing_homes, covid_deaths, bad_gift, tested_positive, gift_china, death_rate, supreme_court, new_york, real_virus, covid_racism
June 01, 2020	−0.05	people, US, health, protests, care, cases, pandemic, home, testing, trump, black, virus, please, masks, curfew, tests	covid_testing, stay_home, testing_centers, impose_curfew, eight_pm, curfue_impose, fighting_covid, peaceful_protests, health_care, enough_masks, masks_ppe
June 14, 2020	−0.11	pandemic, people, children, cases, virus, staff, US, deaths, killed, worst, disease, beat, unbelievable	covid_blacks, latinx_children, unbelievable_asians, systematically_killed, exposed_corona, going_missing, staff_sitting, recovered_covid, worst_disease
June 21, 2020	−0.02	trump, people, pandemic, masks, rally, tulsa, cases, social, distancing, lockdown, died, hospital, mask, call,	wearing_masks, social_distancing, wake_call, mother_died, still_arguing, tested_positive, trump_campaign, tulsa_rally, trump_rally
June 24, 2020	−0.01	pandemic, people, trump, cases, US, testing, lockdown, positive, lindsay, world, social, masks, president	covid_cases, social_distancing, last_year, drunk_driving, lindsay_richardson, tested_positive, wear_mask, america_recovering
July 06, 2020	−0.02	pandemic, people, trump, cases, lockdown, positive, US, virus, wear, social, distancing, mask	social_distancing, got_covid, severe_respiratory, respiratory_cardiovascular, wear_mask, kimberly_guilfoyle, donald_trump
July 10, 2020	−0.01	andemic, coronavirus, people, cases, trump, control, lockdown, US, schools, students, deaths, masks, virus, home, government	control_covid, covid_cases, covid_schools, social_distancing, shake_hands, kneel_bow, hands_hug, vs_right, left_vs

[a] the lowest average sentiment reached on the particular date, [b]excluding the significantly dominating unigrams: COVID, corona, coronavirus and other terms, such as SARS, nCoV, SARS-CoV-2, etc

Machine learning models can be trained on large-scale tweets corpus for classifying the tweets into multiple informational categories, including a separate class for "queries." Even after the automatic classification, each category still contains hundreds of thousands of tweets conversations, which require further in depth analysis.

2. Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends

Author(s): May Oo Lwin1 , PhD; Jiahui Lu2 , PhD; Anita Sheldenkar1 , MSc; Peter Johannes Schulz3 , PhD; Wonsun Shin4 , PhD; Raj Gupta5 , PhD; Yinping Yang5 , PhD

Keywords: COVID-19; Twitter; pandemic; social sentiments; emotions; infodemic

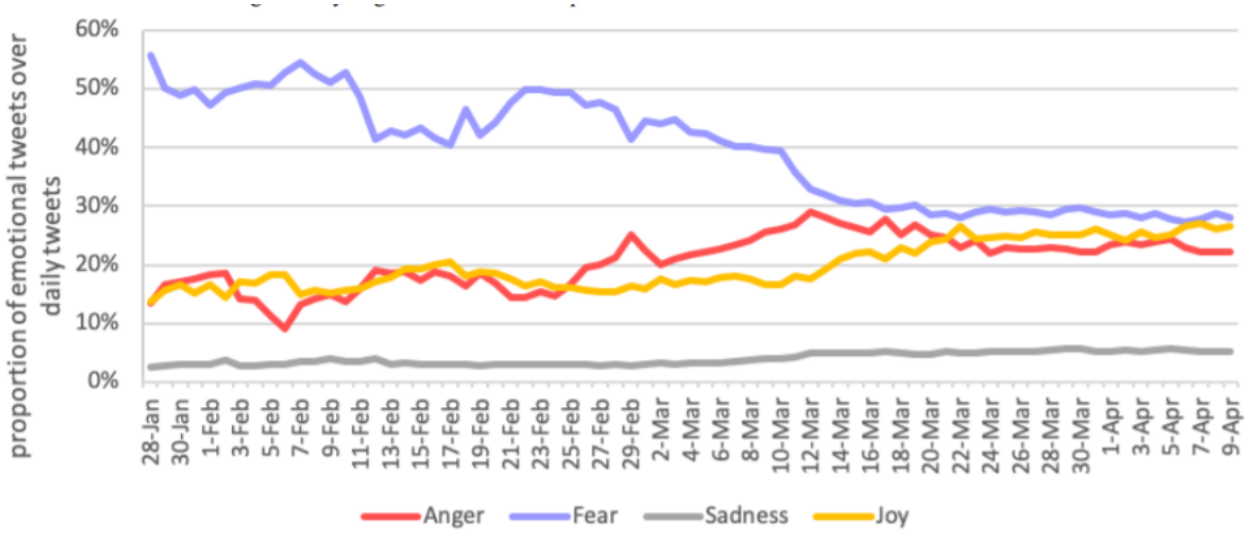
Summary/Takeaways:
Objective: This study aimed to examine worldwide trends of four emotions—fear, anger, sadness, and joy—and the narratives underlying those emotions during the COVID-19 pandemic.

Result: Public emotions shifted strongly from fear to anger over the course of the pandemic, while sadness and joy also surfaced. Findings from word clouds suggest that fears around shortages of COVID-19 tests and medical supplies became increasingly widespread discussion points. Anger shifted from xenophobia at the beginning of the pandemic to discourse around the stay-at-home notices. Sadness was highlighted by the topics of losing friends and family members, while topics related to joy included words of gratitude and good health.

Conclusion: The steady rise of societal concerns indicated by negative emotions needs to be monitored and controlled by complementing regular crisis communication with strategic public health communication that aims to balance public psychological wellbeing.

specific emotions have been found to be more closely linked to psychological processes and behaviors than the overall positive and negative valences [4]. Therefore, we postulate that distinct emotions emerging from social media and their underlying narratives are highly relevant to the current COVID-19 crisis

Investigating the evolution of these four basic emotions can demonstrate the changing dynamics of the public's experience to the crisis.



Method:
"wuhan," "corona," "nCov," and "covid" as search keywords. These keywords were selected because they were widely used during the early assessment of the COVID-19 situation.
The underlying emotions of tweets were analyzed using the algorithm **CrystalFeel,** a sentiment analytic technology whose accuracy had been demonstrated

Word clouds were generated for each of the four emotions based on the top frequent unigrams and bigrams.

Keyword Identifiers:
fear: related to the emergence of COVID-19 and its unknown nature, causing uncertainty about containment and spread, indicated by words such as “first case” and “outbreak.”
as pandemic escalated = fear about shortages of COVID-19 tests and medical supplies indicated by words such as “test shortages” and “uncounted.”
anger: word clouds suggest xenophobia at the beginning of the pandemic when the disease was predominantly localized to China and Asia, indicated by words such as “racist” and “Chinese people.”
shifted to discourse around isolation fatigue, indicated by words such as “stay home” and several swear words.
sadness: topics of losing friends and family members are surfacing, with words relating to “loved one” and “passed away,”
joy: hope, gratitude, and human resilience with words such as “Thank,” “good news,” and “feel good.”

Our findings reveal that negative emotions are dominant during the COVID-19 pandemic, supporting the recent call for action to maintain the public’s mental wellbeing for this unprecedented crisis

Future studies should further investigate sentiments by examining specific countries and expanding the scope to include other media platforms such as Facebook and Weibo. Well… we are looking at the US only, so maybe our research (if successful) can be viewed as an extension of this.

3. Examining the Impact of COVID-19 Lockdown in Wuhan and Lombardy: A Psycholinguistic Analysis on Weibo and Twitter

Author: Yue Su 1,2, Jia Xue 3,*, Xiaoqian Liu 1,*, Peijing Wu 1,2,4, Junxiang Chen 5 , Chen Chen 6 , Tianli Liu 7 , Weigang Gong 8 and Tingshao Zhu 1,2

Keywords: impact of COVID-19 lockdown; public health emergencies; psycholinguistic analysis; psychological states

Summary/Takeaways:
study aims to examine and compare the impact of COVID-19 lockdown on individuals’ psychological states in China and Italy.

(1) sampling Weibo users (geo-location = Wuhan, China) and Twitter users (geo-location = Lombardy, Italy); (2) fetching all the users’ published posts two weeks before and after the lockdown in each region (e.g., the lockdown date of Wuhan was 23 January 2020); (3) extracting the psycholinguistic features of these posts using the Simplified Chinese and Italian version of **Language Inquiry and Word Count (LIWC)** dictionary; and (4) conducting **Wilcoxon tests** to examine the changes in the psycholinguistic characteristics of the posts before and after the lockdown in Wuhan and Lombardy, respectively.

Existing studies have widely used the **Language Inquiry and Word Count (LIWC)** and confirmed it as a valid tool for psychometric analysis. The LIWC dictionary includes many word categories of linguistic features that are related to mental processes and human behaviors [22]. For example, the word category of personal pronouns reflects attentional allocation [22]

4. Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems

Author: Carlos de las Heras-Pedrosa, Pablo Sánchez-Núñez and José Ignacio Peláez

Keywords: coronavirus; COVID-19; SARS-CoV-2; public health; health communication; sentiment analysis; emotion understanding; opinion mining; risk communication; social media

Summary/Key Takeaways:

One of the most usual problems that we must deal with when using information from digital ecosystems is detecting spammers, fake information generated by bots, which tries to influence or modify the perceived opinion on existing information. To detect and discard this type of information we have implemented different types of algorithms based on Support Vector Machine (SVM) techniques which can detect the patterns of this kind of communications, such as the age of the account (in days), the number of comments from the account, follower/following ratio, and the ratio of messages containing URLs. To prevent the effect of spammers, in this work we implemented and applied filters previously defined and tested in other scientific works [27,28].

The emotion information from each communication was extracted employing the natural language analysis tools provided by the **IBM Watson Analytics service** [29]. The emotional intensity was measured in a 0 to 1 scale, where 0 represents the complete absence of this emotion; and 1 represents an absolute high intensity of the emotion. In total, this study measured the emotional intensity of four primal emotions—anger, fear, disgust, and sadness

In this work, we made use of the Natural Language Understanding service from the IBM Watson platform which, given an input text, provides an analysis of syntactic characteristics as well as information on categories, concepts, emotions, entities, keywords, metadata, relationships, and semantic roles

The reliability of the resultant emotion information was tested using the **Interval Majority Aggregation Operator (ISMA-OWA)** [36], which is designed for Decision Making in Social Media with Consistent Data, leveraged by the combination of computational intelligence and Big Data techniques [37]

When people express opinions in communications, they do not do so in numerical value with a fixed scale, they use natural language expressions such as “this is great” or “this is not so good”, so we employed the intervalar representation proposed in [36,38] instead of a numerical scale. The main advantage of this approach is that intervals represent the information within communication in a way that is more similar to the way people express themselves in digital ecosystems, thus reducing the loss of information associated with forcing linguistic data to a hard-numerical scale.

Another advantage of the usage of an intervalar representation of digital ecosystem data is the availability of consistency indices that can be applied to the matrices obtained from communications to detect inconsistencies derived from uninformed opinions. For this purpose, in this work, we employed the CI+ index defined in [39].

The frequency of the words comprising the sample of communications was calculated using a natural language processing algorithm implemented in Python 3, using the Natural Language Toolkit (NLTK) [40].

Moreover, the emotion polarity (positive or negative) was measured using a multilayer perceptron model, trained to classify the emotional weight of written communications [38,41].
For the analysis of the messages emitted by the Spanish government, a content analysis of all press releases during the period of study was carried out. Messages were classified as positive, neutral, or negative by selecting the most significant words from them. The frequency of repetition of these words was another objective of this content analysis. The result has been shown through a word cloud representative of the emotions and feelings expressed by the government in its press releases.

5. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis

Author: Angela Leis, PsyM; Francesco Ronzano, PhD; Miguel A Mayer, MD, PhD; Laura I Furlong, PhD; Ferran Sanz, Prof Dr

Keywords: depression; social media; mental health; text mining

Summary/Takeaways:
Keyword indicators of depression: overwhelmed, exhausted, distressed, anxiety, anxious, tired, low, depression, discouraged, desperate, demotivated, insomnia, cry, nervous, worried, lonely, sad, empty, suicide, antidepressant(s), hopeless,

All the profile descriptions, including 1 or more occurrences of the word “depr” and all the possible derivations related to the word depression in Spanish, such as “depre,” “depresión,” “depresivo,” “depresiva,” “deprimido,” and “deprimida,” were considered.

The textual content of each tweet was analyzed by means of the following sequence of steps:
• Tokenization performed by means of a custom Twitter tokenizer included in the Natural Language Toolkit [43].
• Part-of-Speech (POS) tagging performed by means of the **Freeling Natural Language Processing tool** in order to analyse the usage patterns of grammatical categories (eg, adjectives, nouns, or pronouns) in the text of tweets [44]
• Identification of negations performed by relying on a custom list of Spanish negation expressions, such as nada (nothing), nadie (nobody), no (no), nunca (never), and alike.
• Identification of occurrences of positive and negative words inside the text of each tweet by means of 2 Spanish polarity lexicons: the Spanish Sentiment Lexicon and the Spanish SentiCon Lexicon [45,46]. We exploited 2 lexicons to consider and compare 2 approaches of modeling polarity in Spanish texts, thus reducing any language modeling bias that the use of a single language resource could introduce.
• Identification of words and expressions associated with the basic emotions [47] by using the Spanish Emotion Lexicon [48]. Such emotions are alegría (happiness), enojo (anger), miedo (fear), repulsión (disgust), sorpresa (surprise), and tristeza (sadness).

6. Random Blog Post

Title: Detecting Depression in Social Media Via Twitter Usage

Author: Tulasi ram Ponaganti

Summary/Takeaways:
Tweets were retrieved using the Twitter scraping tool TWINT. The scraped Tweets were manually checked for relevance (for example, Tweets indicating emotional rather than economic or atmospheric

depression) and Tweets were cleaned and processed. Tweets were collected by searching for terms specifically related to depression, specifically to lexical terms as identified in the unigram.

VADER was also utilized for general sentiment analysis of Tweets. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media for general sentiment analysis that is specifically attuned to sentiment in microblog-like contexts. It allows for not only the classification of sentiment but also the associated sentiment intensity measures.

7. Identifying Depression on Twitter

Author: Moin Nadeem1 , Mike Horn. 1 , Glen Coppersmith2 , Johns Hopkins University and Dr. Sandip Sen3 , PhD, University of Tulsa

Keywords: Depression, Machine Learning, Social Media, Twitter

Summary/Takeaways:
Approaches that utilize objective information, such as log data about an individual's activities to predict depression have been studied recently. Resnik et al. has formulated a method for identifying depression in individuals through analyzing textual data written by these individuals. They obtained topics from the essays written by college students by applying **latent Dirichlet allocation (LDA),** a popular topic extraction model within Machine Learning [20]. Through using these discovered topics from a statistical model, they were able to estimate depression and neuroticism in college students with an r value of .45, thus discovering a slight correlation between neuroticism, depression, and academic works by college attendees. Resnik et al. becomes relevant for their novel use of topic modeling; otherwise these academic works are often a poor dataset to derive diagnoses from [20].

Using a simple regressive analysis, **Tsugawa et al**. discovered that frequencies of word usage are useful as features towards identifying depression on Twitter, and therefore is notable for furthering the search of which features to utilize in order to estimate the severity of depression [22].

De Choudhury et al. discovered that the onset of depression through social media may have been able to be characterized through a decrease in social activity, raised negative effect, a highly clustered egonetwork (ie. highly clustered social groups, as opposed to an open-graph model), heightened relational and medicinal concerns, and a greater expression of religious involvement.

Coppersmith et al. developed a Shared Task for the Computational Linguistic and Clinical Psychology (CLPsych) conference. Through this shared task, Coppersmith et al. distributed a standardized dataset of depressed, PostTraumatic Stress Disorder (PTSD), and control users to all competitors in order to normalize fundamental computational technologies which often were at play [24]

UMD utilized a supervised topic model approach to discover groupings of words that provided maximal impact to differentiate between the three provided classes for each user. Furthermore, rather than treat each tweet as its own document, or treat each user as one collective document, they chose to sensibly concatenate all tweets from a given week as a single document [20].

The **WWBP team** utilized straightforward regression models with a wide variety of features, including inferring topics automatically, and binary unigram vectors (ie. "did this user ever tweet this word?"). These topic models provided varying interpretations on which groups of words belonged together, thus providing insight as to which approach best expresses mental health-related signals [26].

The team from Duluth took a powerful approach to this by decoupling the power of an open-vocabulary approach Figure 1 shows the ROC curves for each particular model for each combination of classes. In particular, University of Maryland's approach consistently outperformed competitors. 4 to simple, raw language features. Quite importantly, this open vocabulary approach might have been simplistic in nature but achieved an average precision in the range of .70 - .76, while complex machine learning or complex weighting schemes performed just as well [27].

The Microsoft-IHMC-Qntfy joint team utilized a character language model (CLMs) to determine how likely a given sequence of characters is to be generated by each classification class, and provided a score for each string. The beauty of this approach lied within scoring extremely short text, capturing information for creative spellings, abbreviations, and other textual phenomena which derives from Twitter's unique 140-character limit [24].

We employ four different types of binary classifiers in order to estimate the likelihood of depression within users. For each classifier, we utilize Scikit-Learn from Pedregosa et al. to implement the learning algorithm [32]. We chose to evaluate Linear, Non-Linear, and Tree-based approaches in order to shallowly explore foundational learning models against our dataset. Ultimately, we decided upon Decision Trees, a Linear Support Vector Classifier, a Logistic Regressive approach, as well as a Naïve Bayes algorithm.

In implementation, we used employed a CountVectorizer with default settings from Scikit-Learn developed by Pedregosa et al [32].

8. Depression Status Based on Twitter Posts Using Machine Learning

Author: Chempaka Seri binti Abdul Razak1 , Muhammad Ameer bin Zulkarnain1 , Siti Hafizah binti Ab Hamid1 , Badrul Anuar bin Juma'at1 , Mohd Zalisham Jali2 , Hasni bin Meon3

Keywords: Depression, Twitter, Machine Learning, Emotion

Summary/Takeaways:

Naïve Bayes Classifier

The Naïve Bayes classifier is the simplest and most commonly used classifier. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$P(label|features)$ = P(label)$*P(features|label)$ / P(features) (1)

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features|label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred. Given the Naïve Bayes assumption which states that all features are independent, the equation could be rewritten as follows:

$P(label|features)$ = P(label)$*$P($f1|label$)$*… …*P(fn|label) / P(features) (2)

Natural Language Processing (NLP)

Natural Language Processing (NLP) techniques are sometimes used with the lexiconbased approach to find the syntactical structure and help in finding the semantic relations. Moreo and Romero [9] have used NLP techniques as preprocessing stage before they used their proposed lexicon- based SA algorithm.

Their proposed system consists of an automatic focus detection module and a sentiment analysis module capable of assessing user opinions of topics in news items which use a taxonomy- lexicon that is specifically designed for news analysis. Their results were promising in scenarios where colloquial language predominates.

The approach for SA presented by Caro and Grella [10] was based on a deep NLP analysis of the sentences, using a dependency parsing as a pre-processing step. Their SA algorithm relied on the concept of Sentiment Propagation, which assumed that each linguistic element like a noun, a verb, etc. can have an intrinsic value of sentiment that is propagated through the syntactic structure of the parsed sentence. They presented a set of syntactic-based rules that aimed to cover a significant part of the sentiment salience expressed by a text. They proposed a data visualization system in which they needed to filter out some data objects or to contextualize the data so that only the information relevant to a user query is shown to the user. In order to accomplish that, they presented a context- based method to visualize opinions by measuring the distance, in the textual appraisals, between the query and the polarity of the words contained in the texts themselves. They extended their algorithm by computing the context-based polarity scores. Their approach approved high efficiency after applying it on a manual corpus of 100 restaurants reviews.

Min and Park [11] have used NLP from a different perspective. They used NLP techniques to identify tense and time expressions along with mining techniques and a ranking algorithm. Their proposed metric has two parameters that capture time expressions related to the use of products and product entities over different purchasing time periods. They identified important linguistic clues for the parameters through an experiment with crawled review data, with the aid of NLP techniques.

Feedforward Neural Network

A feedforward neural network is a type of artificial neural network composed of a collection of linked computational units, often referred to as nodes or neurons, that are 5 arranged in multiple layers, where information in the resulting network of units is propagated forward from one layer to another, from an initial input layer, to one or more hidden layers, and then finally to an output layer, in a non-cyclic fashion. In this study, the initial layer consisted of the inputs, represented as a “bag of words”, or a vector of zeros and ones, with each value in the vector representing the presence or absence of a unique token. These inputs are then multiplied by a vector of weights, which are the parameters for the network that are trained through a learning process. The output produced by a hidden layer, or a layer between the input and output layers, is the result of the application of an activation function to the values from the previous layer.