Topic Sentiment Trend Detection and Prediction for Social Media

A Thesis

in

Computer Science

Presented to the Faculty of the University of
Missouri–Kansas City in partial fulfillment of the requirements for
the degree
MASTER OF SCIENCE

by

Aashish Thota

B. Tech., Jawaharlal Nehru Technological University, Telangana, India, 2018

Kansas City, Missouri

2020

Topic Sentiment Trend Detection and Prediction for Social Media

Aashish Thota, Candidate for the Master of Science Degree
University of Missouri-Kansas City, 2020

ABSTRACT

Social media often plays a crucial role in disseminating information to warn the public about health concerns. Opioid addiction has become of the significant outbreaks in the United States. Studying opioid issues in social media has the potential to reveal patterns of opioid abuse and understand people's opinions on this issue. On the other hand, social media forums like Twitter allow for open discussions among the Public and popular information exchanges. The question arises if the trends of such concerns like opioid abuse can be automatically detected and predicted for a better understanding of people's attitude changes in a specific direction.

In this thesis, we developed a novel framework for topic sentiment trend detection and prediction in social media. The proposed framework copes with the following tasks: topic trend detection, sentiment analysis, and topic prediction. The VADER-based time series sentiment analysis and the KATE-based topic modeling methods were applied to analyze the social media data from the public, social media news, and newspapers. We have further extended the framework for the successful prediction of topic trends for given the current issues. For the topic trend prediction model, the deep neural network model called Long Short-Term Memory (LSTM) used along with topic embedding techniques. The two-step communication model was used to evaluate how different types of media are active in the emergence and escalation dissemination and how effective it is in the pacification and prevention of concerns related to epidemics, like opioids. The proposed framework is also applied to the New York Times articles on opioids over ten years, from 2010 to 2019. The results in this study have shown some exciting findings from topic sentiment detection and high accuracies from the topic trend prediction.

APPROVAL PAGE

The staff recorded beneath, designated by the Dean of the School of Computing and Engineering, have inspected a proposal titled "Topic Sentiment Trend Detection and Prediction for Social Media," presented by Aashish Thota, candidate for the Master of Science degree, and confirm that as they would see it is deserving of acknowledgment.

Supervisory Committee

Yugyung Lee, Ph.D. (Committee Chair)
Department of Computer Science & Electrical Engineering

Ye Wang, Ph.D.
Department of Communication Studies

Petri, Alexis Nicolle, Ed.D.
Department of Psychology

Content

ILLUSTRATIONS

TABLES

ACKNOWLEDGMENTS

I would like to thank Dr. Yugyung Lee for her valuable guidance and immense support throughout the research work as my advisor. Data analytics growing very fast, Dr. Lee always keeps herself up to date with the latest research and encourages her students to work towards cut edge technologies. I am amazed by her positive energy and patience. Her vast experience, unparalleled knowledge, agile and prompt feedback coupled with smart ideas have helped me immensely in putting up the whole work. She is very patient in listening to all the new ideas, pragmatic in giving suggestions, and always helps me in doing the reality check.

I would also like to thank Saria Goudarzvand, Ph.D. student, for her guidance for my thesis. Her skills in machine learning and deep learning are amazing. Her fantastic energy and enthusiasm always motivate me to go the extra mile. It has been an honor to work with her on many projects besides the thesis.

I wish to thank the members of my dissertation committee: Dr. Yugyung Lee, Dr. Ye Wang, and Alexis Petri for generously offering their time, support, guidance, and goodwill throughout the preparation and review of this document.

I would like to thank the University of Missouri-Kansas City for providing me with a platform to plan and execute my research work. It provided me with many opportunities to support myself and world-class facilities to research with the machines available from August 2019 until May 2020.

Finally, I would like to thank my parents for their enormous support and spiritual guidance in every walk of life. They are the backbone of all my academic progress to date.

CHAPTER 1
INTRODUCTION

In social media, millions of active users express their opinions and interact with each other daily. Such users' content in the form of posts or tweets provides a vast amount of useful information if analyzed carefully. Therefore, the data streamed from social media such as Twitter, Facebook, or Instagram is so precious for researchers to perceive the users' social behavior by applying sentiment analysis on it. Twitter is one of the most significant ones among all the micro-blogging services. A massive amount of user-generated online content is freely available to the real-time monitoring of public sentiment.

Opioid epidemic has become a crisis in the United States, seeing a considerable rise in the number of opioid-related deaths in recent years. With the dramatic increase of opioid overdose deaths, the opioid crisis has been declared as a national public health emergency by President Trump. In Figure 1, we see that according to the National Institute on Drug Addiction, the number of deaths due to drug overdose increased each year gradually. In 2017 there were more than 49,000 deaths, which made opioid a national crisis that year.

The fusion of topic and sentiment has been used for Topic Trend Detection and Sentiment Analysis on Opioid data. Topic modeling is an unsupervised statistical machine learning technique. The purpose of the topic modeling is to discover the abstract topics from the collection of documents. It is different from the rule-based approach, where we use the Dictionary or lexicon to search keywords. The topic modeling is used to extract and find a group of words called "topics" in substantial text clusters. There are several approaches available for finding out topics from text corpus, namely Term Frequency-Inverse Document Frequency and Non-Negative Matrix

Factorization technique. Latent Dirichlet Allocation (LDA) [2] is the most famous example of the topic model. It is used to classify text in a document which is assumed to be a mixture of topics, to a particular topic.

Topic modeling is used to extract the topics from the unstructured twitter dataset to find the opinion of the public and their trend. We use KATE [1], a novel auto-encoder based approach, as a topic modeling method. Because of the opposition between the neurons in the shrouded layer, every neuron gets had practical experience in perceiving explicit information examples, and in general, the model can learn significant portrayals of literary information.

It is difficult to find the contextual sentiment of a text. Sentiment analysis is one of the critical issues today. The primary job is to fast-pace the process of opinion extraction from the given subject. The subject here can be an excerpt from the written text, debate, or day to day conversation. In sentiment analysis, we also evaluate the positive and negative intensities of symbols and words. Sentiment analysis helps to improve customer services, Political planning Policies, and manufacturing quality products.

Online platforms like social media and blogs are widely used by public and mass media to express their opinions during the crisis. Moreover, sentiment analysis is also performed on those opinions to better understand the emotion attached to those opinions. Topic modeling on Social media gives us better insights into public view during an epidemic like the opioid crisis. Notably, in social media, people are widely expressing their problems related to a drug overdose. Moreover, the analysis of the data is useful to decipher the change in opinions and trends of people.

Analyzing opioid-related Social Media like Twitter, Reddit communities can bring better insights on how it will affect the quality of life of the public. Due to potential social stigma, users using opioids may not be willing to discuss their concerns openly; this could lead to insufficient access to healthcare and negatively impact patients' education and employment. Social media platforms allow increased self-disclosure for users to discuss otherwise sensitive topics.

Many research works have been conducted for Opioid epidemic analysis; some of them include Discovering opioid use patterns [10], Analysing psychological trends in opioid, prevention of opioid prescriptions, etc. But no research work tries to get a public opinion at the time of such epidemic and to get insight into how public opinion is changing.

So, as a crisis remains a significant health problem, it is necessary for us to be aware of what public concerns are throughout this health hazard and their emotions attached to it and how they are changing. There is also a need for a framework to see how news media is trying to influence public concerns and up to which extent.

So, this research work has two objectives:

i. To apply Topic Trend Detection and Sentiment analysis to get insights on public opinion and news media over the period 2010 to 2019.

ii. Topic Prediction to understand what future concerns might be.

CHAPTER 2
BACKGROUND AND RELATED WORK

This chapter gives background information of various components used in the thesis and provides an overview of related work that will help in understanding this work better.

2.1. Related Work

2.1.1. Topic Detection

Twitter is a blogging site that attracted many researchers for its scope in data. Many research works were done in Topic Trend Detection. We can categorize that research works in Statistical Methods and Machine learning methods.

One such statistical method used was High Utility Pattern Mining [8], which collects tweets and gets the frequency of words in the tweets and utility, Growth rate of frequency of the word. For a chunk of tweets by time-based windowing on the Twitter stream, we define the utility of words based on the growth rate in frequency and find groups of words with high frequency[8]. After collecting utility of each word minimum threshold for the utility was set. Words with more than threshold utility were selected to generate candidate topics. Candidate topics are top n results from the patterns, the co-occurrence of words, and their support. For Post-processing to extract actual topic patterns from candidate topic patterns generated by HUPM, an efficient data structure called Topic-tree (TP-Tree) is also proposed. There were some limitations, the work does not consider the context of the words, and there is no customization of models.

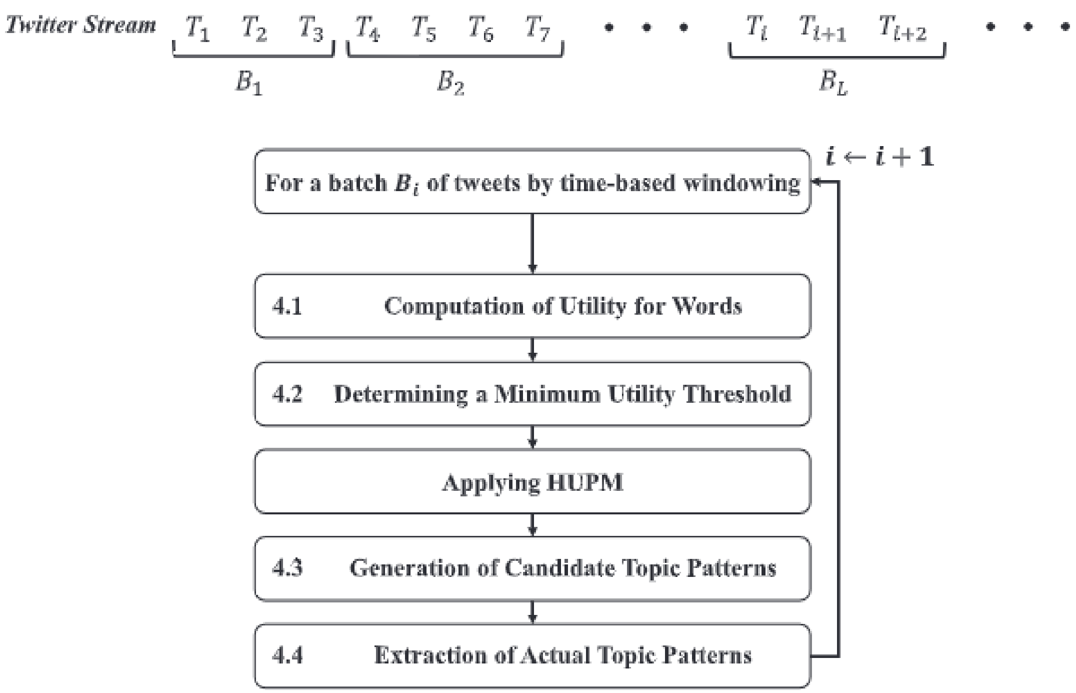

Twitter Stream
T_1 T_2 T_3 T_4 T_5 T_6 T_7 • • • T_i T_{i+1} T_{i+2} • • •
B_1
B_2
B_L
For a batch B_i of tweets by time-based windowing
$i \leftarrow i+1$
4.1 Computation of Utility for Words
4.2 Determining a Minimum Utility Threshold
Applying HUPM
4.3 Generation of Candidate Topic Patterns
4.4 Extraction of Actual Topic Patterns

Figure 1 High Utility Pattern Mining

In this paper [9], the author proposed an association analysis and ensemble forecasting to automatically discover topics from a set of text documents and forecast their evolving trend in the near future. To find meaningful topics, they collected publications from a particular research area, data mining, and machine learning, as our data domain. An association analysis process is applied to the collected data first to identify a set of topics, followed by temporal correlation analysis to help discover correlations between topics, and identify a network of topics and communities. After that, an ensemble forecasting approach was proposed to predict the popularity of research topics in the future. The limitation of this work is, association rules were selected randomly, and at last, forecasting was aggregated.

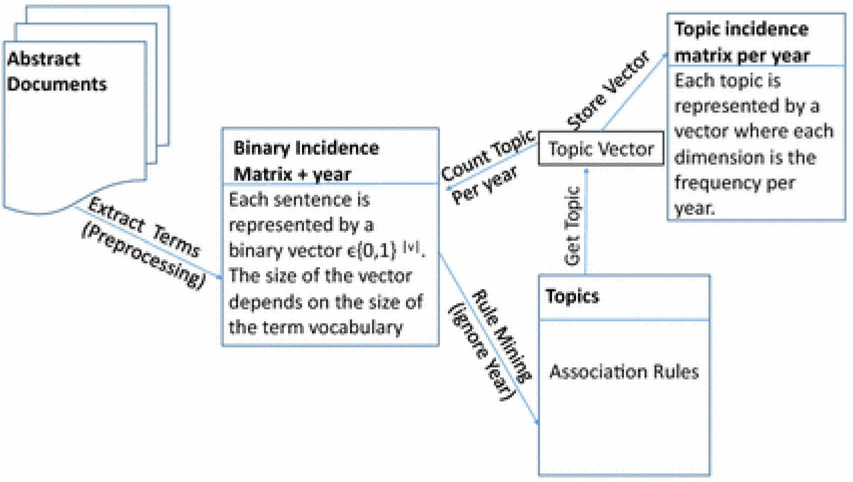

Abstract Documents
Extract Terms (Preprocessing)
Binary Incidence Matrix + year
Each sentence is represented by a binary vector ϵ{0,1} |v|. The size of the vector depends on the size of the term vocabulary
Count Topic Per year
Topic Vector
Store Vector
Topic incidence matrix per year
Each topic is represented by a vector where each dimension is the frequency per year.
Get Topic
Rule Mining (ignore Year)
Topics
Association Rules

Figure 2 Association rule-based learning

In this paper [13], we propose an efficient method to pick out the keywords frequently employed on Twitter that are mostly related to events of interest, like protests. The quantity of those keywords is tracked in real-time to spot the events of interest in a very binary classification scheme. We use keywords within word pairs to capture the context. The proposed method is to binarize vectors of daily counts for every word-pair by applying a spike detection temporal filter. Then the strategy uses the Jaccard metric to live the similarity of the binary vector for every word-pair with the binary vector describing event occurrence. The highest n-word pairs are used as features to classify any day to be an incident or non-event day. The chosen features are tested using multiple classifiers like Naive Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbour (KNN), and decision trees.

2.1.2. Opioid Analysis

Opioid abuse has become an epidemic in the United States. There has been some research work done for topic identification on opioid data. In this paper [3], the author talks about efficiently identifying the topics of opioid-related discussions from social media may provide valuable information such as users behaviors, preference, and utilization patterns. Prior research for topic identification is primarily based on probabilistic graphical models. It does not directly capture word co-occurrences information that is primary to preserve topic coherence.

To evaluate the performance of the topic identification model, the author collected data from the "r/opiates" community on Reddit between January 1, 2017, and January 1, 2018. The dataset contains 27,290 main posts and 173,552 comments, including 668,057 sentences in total.

This work aims to obtain the attention weights of words in the sentence for sentence representation learning. Firstly, each word h in the vocabulary is associated with a representation vector. Then, suppose that we focus on k topics and provide some seed words of each topic. We encode the prior knowledge into a background topic representation matrix with an average of embedding of seed words so that the topic representation and word representation are on the same embedding space.

Explicitly, for each word t in a sentence, we compute the cosine similarity between its vector and each topic background representation in the embedding space. The larger the cosine distance is, the more likely the word belongs to that topic, we use the maximum similarity of each name to measure its attention weight when modeling sentence embedding. Then sentence embedding is done using the average of weighted word embedding. Then they measure the topic distribution of sentence to decide to which topic sentence belong to and then reconstruct the sentence from the embedding. The limitation of this work was the method needs to provide seed words for topics externally to be able to give prior knowledge to the model to identify topics.

In this paper [4], the goal of the current study is to identify the public's reactions to the opioid epidemic by identifying the most popular topics tweeted by users.
This study used textual analytics to identify topics and extract meanings contained in unstructured textual data, and followed the same procedures used in past studies identifying prevalent themes among tweets. 17-18 Twitter messages were captured during a period beginning August 15 and ending October 15, 2016. The following specific keywords or phrases were used to obtain relevant twitter messages: "turnthetide," "#turnthetide," "turnthetiderx," "#turnthetiderx," "turn the tide rx," "actonopioids," "opioid," "opioids," "opium," and "opiate." The keywords returned 226,711 messages.

The unstructured data content was analyzed using SAS Text Miner 12.1. SAS Text Miner is an algorithmic-driven statistical software used to uncover and understand textual data. SAS Text Miner provides the ability to parse and extract information from text, filter and store the data, and assemble it into related topics for introspection and insights by the researchers. After the tweets and retweets were separated, the initial step was to extract, clean, and create a dictionary of words using a natural language processor (NLP) for each data set. With the inclusion criteria set, a Text Topic node was used to combine terms into ten to fifteen topic groups. This clustering divided the document collection into groups based on the presence of similar themes using expectation maximization (EM) clustering.

In this paper [12], the author analyzed the data to understand the psychological categories of the posts and performed topic modeling to reveal the essential topics of interest. It also characterized the extent of social support received from comments and scores by each post. Lastly, they analyzed a statistically significant difference in the posts between anonymous and non-anonymous users.

All the work in Opioid-related field Concentrated on topics spoken by people in one year but did not concentrate on the trend of public opinion; at the same time, none of the work considered the effect of News media on public opinion.

2.1.3. Topic Prediction

In this paper [14], the author designed a preliminary attention model that can detect the trending topics in text streams. Attention models are encoder-decoder models that support the decoder to pick only the encoded inputs that are important for each step of the decoding process.

The model is composed of three main layers, namely, encoder, attention, and decoder. In the first layer, the encoder encodes the entire input sequence, for each time step, into a fixed-length vector.

Then, the softmax operation normalizes the probabilities of the encoded vectors by multiplying each encoded vector by its weight to obtain a time-dependent input encoding, which is fed to each step of the decoder.

Afterward, the model calculates the context vector. This vector summarizes the importance of the different encoded values. The decoder steps through the output time series while reading from the context vector.

CHAPTER 3
METHODOLOGY

The proposed framework is divided into two parts, topic trend detection and sentiment analysis, which is a combination of sentiment analysis using VADER Sentiment Analysis and topic modeling using KATE topic modeling. Topic Prediction, which is a Combination of KATE Topic modeling and Long-short term Memory model. Figure 3 shows the overall architecture of the proposed framework.

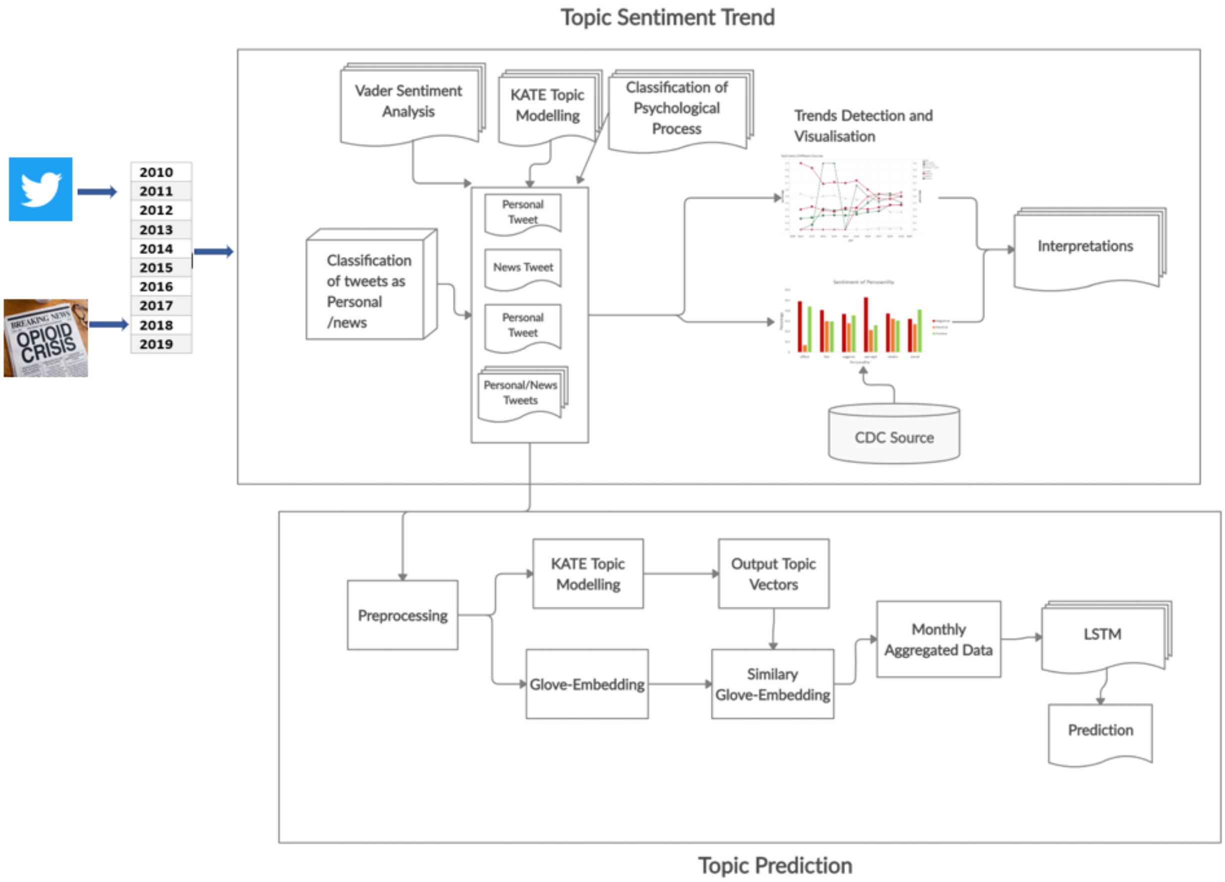


Figure 3 Overall Framework

3.1 Data Collection and Pre-processing

Twitter provides a large scale of text data, which is nothing but public opinions and news. So, we scrapped twitter data from 2010 until 2019 with hashtags opioid, opioids and opioid crisis. We also want to see how Traditional News is working at the time of the epidemic, so we scrapped New York Times articles on opioids from 2010 to 2019. CDC website data on Drug overdose was also considered. After collecting the data from three sources, we applied some pre-processing techniques like

- Scrapped Twitter and Newspaper Articles for ten years(2010 - 2019): Scarped tweets from twitter advance search from January 2010 until December 2019.
- Remove all stop words and punctuation: Removed all Stop words like is, the, of, was, etc. and also removed punctuation like ',",! ,. ,

  Example:

  Original Text: "Hello! how are you doing?"

  After Removing Punctuation: Hello how are you doing
- Remove URLs: Some tweets may contain URLs in them. So, this step removes all the URLs from tweets.

  Example:

  Original text: " Hey Mayur, We have successfully credited the cashback amount of INR 600 in your CITRUS wallet linked to mobile - 9130977755. Please download and rate our apps here - http://smarturl.it/ixigoapps Happy traveling! Team ixigo"

After Removing URLs: "Hey Mayur, We have successfully credited the cashback amount of INR 600 in your CITRUS wallet linked to mobile - 9130977755. Please download and rate our apps here. Happy traveling! Team ixigo"

- Remove hashtags: We collected tweets using some hashtags like #opioid, #opioids, and #opioidcrisis. Some tweets may also contain unnecessary hashtags. So, we remove the hashtags.

  Example:

  Original Text: "Hello! how are you doing?" #GOODMORNING

  After Removing Hashtags: Hello how are you doing

- Remove UTF characters: Remove Characters like emoticons.

  Example:

  Original Text: Dukes are looking to defend their title while Bison are looking to reclaim it NDSU last won in   were eliminated in Ã¢Â€Â˜   by JMU

  After Removing UTF characters: dukes defend title bison reclaim ndsu eliminated jmu

- Convert all the text to lowercase: Convert all the tweets and text to lower case.

  Example:

  Original Text: Hello, How are you today?

  After Converting to Lower Case: hello, how are you doing today.

3.2 Topic Trend Detection and Sentiment Analysis

3.2.1 Tweet Classification

In this section, we classify the tweets into News and Personal tweets. For this, we use a clue-based tweet classification. The clue-based classifier parses each tweet into a set of tokens and matches them with a corpus of Personal clues. There is no available corpus of clues for Personal versus News classification. We used a subjective corpus Multi-Perspective Question Answering MPQA.

The MPQA corpus contains words, including adjectives, adverbs, any position words, nouns, and verbs. As for the sentiment polarity, among all words, 4912 are negatives, 570 are neutrals, 2718 are positives, and 21 will be both negative and positive. Concerning the strength of subjectivity, among all words, 5569 are strongly subjective words, and the other 2652 are weakly subjective words.

Twitter users tend to specific their personal opinions more casually compared with other documents, like News, online reviews, and article comments. It's expected that the existence of any profanity might cause the conclusion that the tweet could be a Personal tweet. We added a group of 247 selected profanity words (Ji 2014a) to the corpus described within the previous paragraph. USA law, enforced by the Federal Communication Commission, prohibits the utilization of a shortlist of profanity words in TV and radio broadcasts (Federal Communications Committee 2014). Thus, any word from this list in a very tweet indicates that the tweet is not a News item. We counted the amount of strongly subjective term, and the number of weakly subjective terms and checked for the presence of profanity words in each tweet, and experimented with different thresholds. A tweet is labeled as Personal if its count of subjective words surpasses the chosen threshold; otherwise, it's labeled as a News tweet

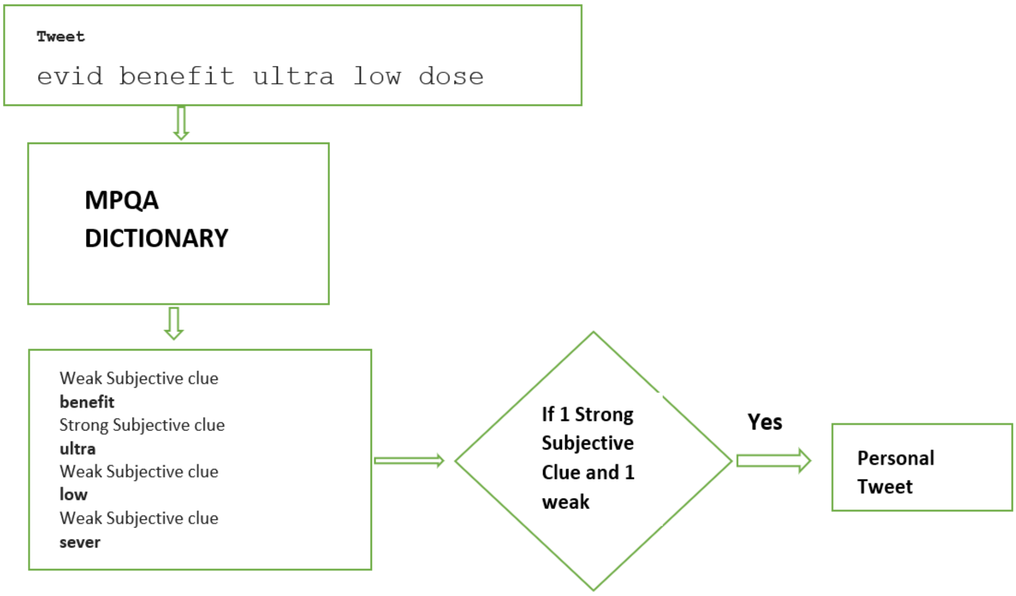


Figure 4 Dataflow of the MPQA Dictionary

3.2.2 Sentiment Analysis

Sentiment Analysis, or Opinion Mining, maybe a sub-field of Natural Language Processing (NLP) that tries to spot and extract opinions within a given text. Sentiment analysis aims to measure the attitude, sentiments, evaluations, attitudes, and emotions of a speaker/writer supported the computational treatment of subjectivity in an exceeding text.

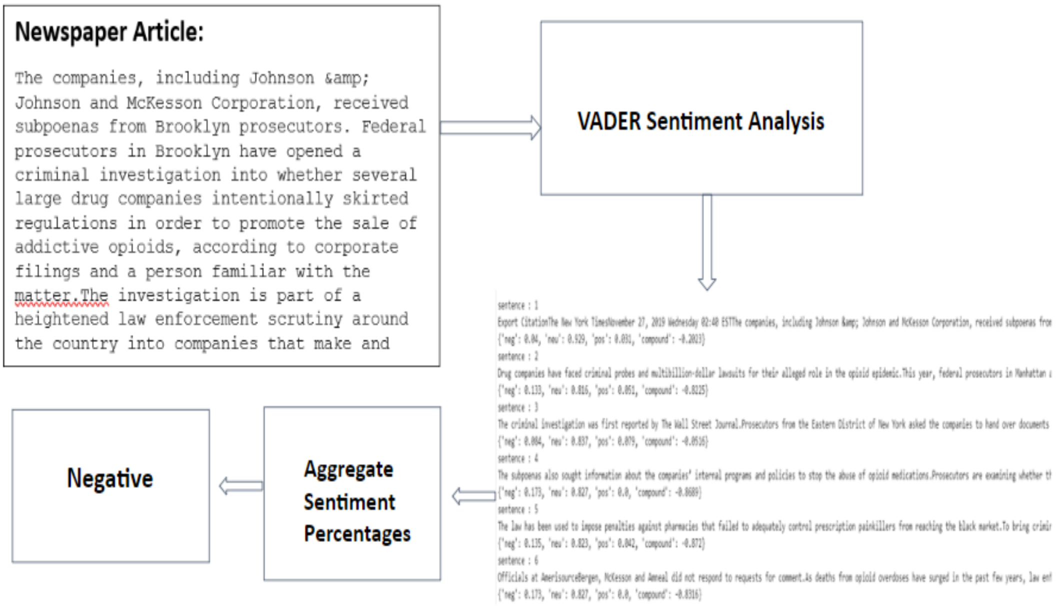
To get emotion related to opinion, i.e., Personal Tweets and journalism like News Tweets and Traditional Newspaper articles, we applied the VADER Sentiment Analysis Technique.

VADER (Valence Aware Dictionary and Sentiment Reasoner) could be a lexicon and rule-based sentiment analysis tool accustomed to getting sentiment related to each sentence. VADER uses a mixture of A sentiment lexicon may be a list of lexical features (e.g., words), which are generally labeled consistent with their semantic orientation as either positive or negative.

VADER was victorious when addressing social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

It is fully open-sourced under the MIT License. The developers of VADER have used Amazon's Mechanical Turk to induce most of their ratings; you'll find complete details on their GitHub Page.

This tool is specifically attuned to Social media texts like tweets and New York Times editorials. This tool showed better performance on single sentences than paragraphs. While Newspaper articles are more than one paragraph, we divided the paragraphs into sentences and then applied VADER on each sentence and aggregated the result to get the overall sentiment. Figure 5 shows the architecture of how newspaper articles are processed through VADER.



Newspaper Article:

The companies, including Johnson &amp; Johnson and McKesson Corporation, received subpoenas from Brooklyn prosecutors. Federal prosecutors in Brooklyn have opened a criminal investigation into whether several large drug companies intentionally skirted regulations in order to promote the sale of addictive opioids, according to corporate filings and a person familiar with the matter.The investigation is part of a heightened law enforcement scrutiny around the country into companies that make and

```
sentence : 1
Export CitationThe New York TimesNovember 27, 2019 Wednesday 02:40 ESTThe companies, including Johnson &amp; Johnson and McKesson Corporation, received subpoenas fro
{'neg': 0.04, 'neu': 0.929, 'pos': 0.031, 'compound': -0.2023}
sentence : 2
Drug companies have faced criminal probes and multibillion-dollar lawsuits for their alleged role in the opioid epidemic.This year, federal prosecutors in Manhattan
{'neg': 0.133, 'neu': 0.816, 'pos': 0.051, 'compound': -0.8225}
sentence : 3
The criminal investigation was first reported by The Wall Street Journal.Prosecutors from the Eastern District of New York asked the companies to hand over documents
{'neg': 0.084, 'neu': 0.837, 'pos': 0.079, 'compound': -0.0516}
sentence : 4
The subpoenas also sought information about the companies' internal programs and policies to stop the abuse of opioid medications.Prosecutors are examining whether t
{'neg': 0.173, 'neu': 0.827, 'pos': 0.0, 'compound': -0.8689}
sentence : 5
The law has been used to impose penalties against pharmacies that failed to adequately control prescription painkillers from reaching the black market.To bring crimi
{'neg': 0.135, 'neu': 0.823, 'pos': 0.042, 'compound': -0.872}
sentence : 6
Officials at AmerisourceBergen, McKesson and Amneal did not respond to requests for comment.As deaths from opioid overdoses have surged in the past few years, law en
{'neg': 0.173, 'neu': 0.827, 'pos': 0.0, 'compound': -0.8316}
```

Figure 5 Newspaper articles sentiment analysis

3.2.3 Topic Modelling

In simple words, Topic Modelling is used for Dimensionality reduction. LDA is one of the probability-based Topic modeling techniques. The method has an assumption that a document is a mixture of topics, and topics are a mixture of words. Another deep learning-based Topic modeling technique is KATE. We used the KATE model to get topics for personal tweets, News tweets, and Newspaper articles.

The objective of an autoencoder is to reduce the reconstruction error, and our goal is to extract essential features from data. Compared to image data, textual data is more difficult for autoencoders since it's typically high-dimensional, sparse, and has power-law word distributions. While examining the features extracted by an autoencoder, we observed that they weren't distinct from each other.

To overcome this, our approach guides the autoencoder to specialize in essential patterns within the data by adding constraints within the training phase via mutual competition. In competitive learning, neurons compete for the proper to retort to a subset of the input file, and as a result, the specialization of every neuron within the network is increased. Note that the specialty of neurons is what we wish for an autoencoder, mainly when applied to textual data. By introducing competition into an autoencoder, we expect each neuron within the hidden layer to recognize different patterns within the input file.

In recent years KATE has been successfully used to extract features from documents. There are many traditional auto-encoders which tend to learn possibly trivial representations of text documents, but we used K-competitive autoencoder as it can determine better representation than traditional autoencoders. Contrary to conventional shallow autoencoders, a K-competitive autoencoder introduces a competition layer among the hidden neurons. Figure 6 shows the architecture of KATE.

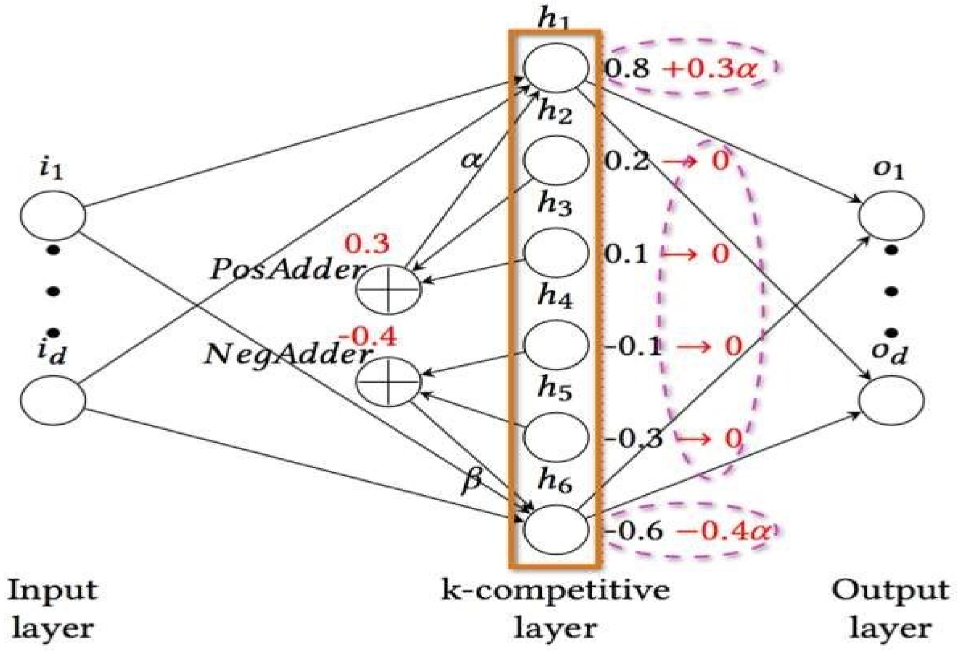
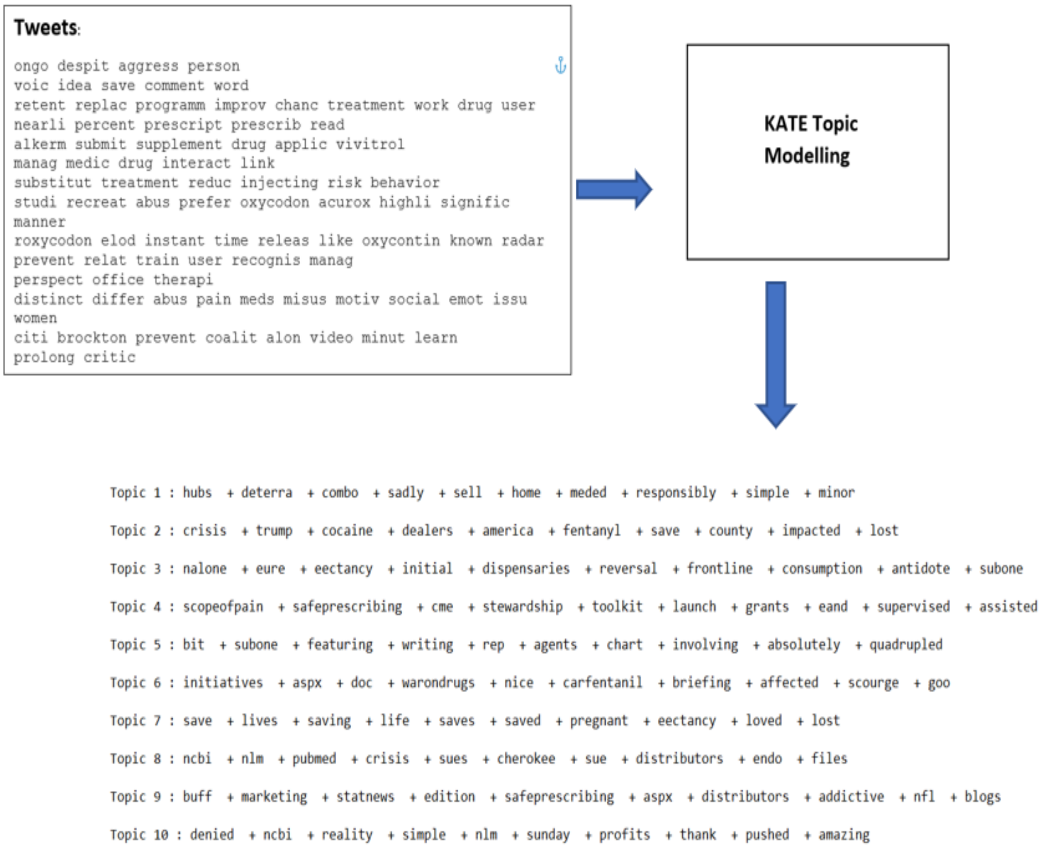


Figure 6 KATE Architecture

KATE uses tanh activation function for the k-competitive hidden layer. We divide these neurons into positive and negative neurons that supported their activations. The foremost competitive k neurons are those who have the first significant absolute activation values. However, we select the ⌈k/2⌉ most vital positive activations because the positive winners and reallocate the energy of the remaining positive loser neurons among the winners using an α amplification connection, where α could be a hyperparameter. Finally, we set the activations of all losers to zero. Similarly, the lowest negative activations are the negative winners, and that they get the amplified energy from the negative loser neurons. We argue that the α amplification connections are a critical component within the k competitive layer.

We applied topic modeling for each year individually for personal tweets, news tweets, and newspaper articles. For each year, we got ten topics. And ten words for each topic. After applying topic modeling, we analyzed the topics for each year and selected ten topics for which we visualized the trend for ten years. Then we found the sentiment of the topic for each year to see how Personal tweets are getting influenced by news media. What kind of environment is news media trying to set?



```
Tweets:

ongo despit aggress person
voic idea save comment word
retent replac programm improv chanc treatment work drug user
nearli percent prescript prescrib read
alkerm submit supplement drug applic vivitrol
manag medic drug interact link
substitut treatment reduc injecting risk behavior
studi recreat abus prefer oxycodon acurox highli signific
manner
roxycodon elod instant time releas like oxycontin known radar
prevent relat train user recognis manag
perspect office therapi
distinct differ abus pain meds misus motiv social emot issu
women
citi brockton prevent coalit alon video minut learn
prolong critic
```

```
Topic 1 : hubs  + deterra  + combo  + sadly  + sell  + home  + meded  + responsibly  + simple  + minor
Topic 2 : crisis  + trump  + cocaine  + dealers  + america  + fentanyl  + save  + county  + impacted  + lost
Topic 3 : nalone  + eure  + eectancy  + initial  + dispensaries  + reversal  + frontline  + consumption  + antidote  + subone
Topic 4 : scopeofpain  + safeprescribing  + cme  + stewardship  + toolkit  + launch  + grants  + eand  + supervised  + assisted
Topic 5 : bit  + subone  + featuring  + writing  + rep  + agents  + chart  + involving  + absolutely  + quadrupled
Topic 6 : initiatives  + aspx  + doc  + warondrugs  + nice  + carfentanil  + briefing  + affected  + scourge  + goo
Topic 7 : save  + lives  + saving  + life  + saves  + saved  + pregnant  + eectancy  + loved  + lost
Topic 8 : ncbi  + nlm  + pubmed  + crisis  + sues  + cherokee  + sue  + distributors  + endo  + files
Topic 9 : buff  + marketing  + statnews  + edition  + safeprescribing  + aspx  + distributors  + addictive  + nfl  + blogs
Topic 10 : denied  + ncbi  + reality  + simple  + nlm  + sunday  + profits  + thank  + pushed  + amazing
```

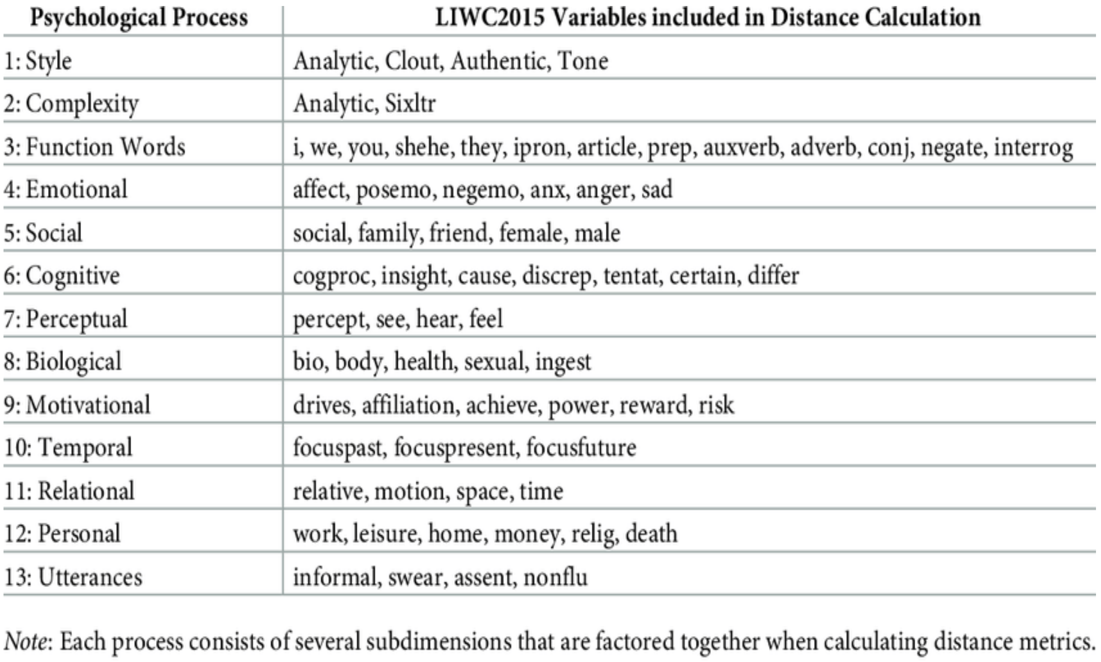
Figure 7 Dataflow through KATE Topic Modelling

3.2.4. Psychological Categorisation

We also want to see what different types of Psychological categories do public and news portray. For this, we have used the LIWC dictionary to classify each tweet of public and news to different psychological categories.

LIWC (Linguistic Inquiry and Word Count) is a software application that provides a useful tool for studying the emotional, cognitive, and structural components contained in language on a word-by-word basis. Early approaches to psycholinguistic concerns involved almost exclusively qualitative philosophical analyses. In this field, more modern researches have provided empirical evidence on the relation between language and the state of mind of subjects or even their mental health. LIWC was developed for providing an efficient method for studying these psycholinguistic concerns thanks to corpus analysis, and it has been considerably improved since its first version.

The LIWC2007 Dictionary is composed of 2,290 words and word stems. Each word or word-stem defines one or more-word categories or sub dictionaries. For



Psychological Process	LIWC2015 Variables included in Distance Calculation
1: Style	Analytic, Clout, Authentic, Tone
2: Complexity	Analytic, Sixltr
3: Function Words	i, we, you, shehe, they, ipron, article, prep, auxverb, adverb, conj, negate, interrog
4: Emotional	affect, posemo, negemo, anx, anger, sad
5: Social	social, family, friend, female, male
6: Cognitive	cogproc, insight, cause, discrep, tentat, certain, differ
7: Perceptual	percept, see, hear, feel
8: Biological	bio, body, health, sexual, ingest
9: Motivational	drives, affiliation, achieve, power, reward, risk
10: Temporal	focuspast, focuspresent, focusfuture
11: Relational	relative, motion, space, time
12: Personal	work, leisure, home, money, relig, death
13: Utterances	informal, swear, assent, nonflu

Note: Each process consists of several subdimensions that are factored together when calculating distance metrics.

Figure 8 LIWC Psychological Process

example, the word 'crying' is part of three-word categories: sadness, negative emotion, and overall affect. Hence, if it is found in the target text, each of these three sub dictionary scale scores will be incremented.

Each Tweet is passed through this tool to get the percentage of tweets belonging to each psychological processes. Figure 9 shows the dataflow of how a tweet is classified into different psychological processes.

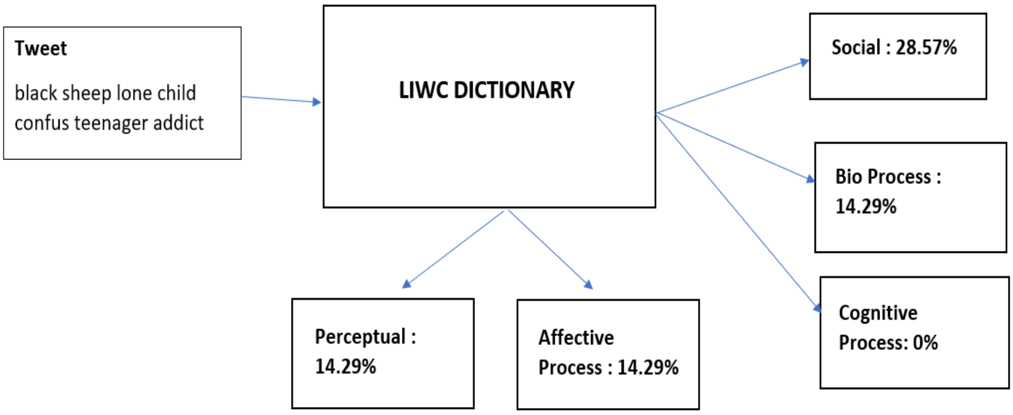


Figure 9 LIWC Dataflow

3.3 Topic Prediction

3.3.1 Glove Embedding

Embedding is employed to represent a document in a vector format. Word embeddings are a sort of word representation that enables words with similar assuming to have the same representation. There are distributed representations for the text that's perhaps one among the critical breakthroughs for the deep learning methods on Natural Language processing problems.

The Global Vectors for Word Representation is an extension to the word2vec method for efficiently learning word vectors. Classical vector space model representations of words were developed using matrix factorization techniques like Latent Semantic Analysis (LSA). LSA did a superb job of using global text statistics but isn't nearly as good because of the learned methods like word2vec at capturing meaning and demonstrating it on tasks like calculating analogies.

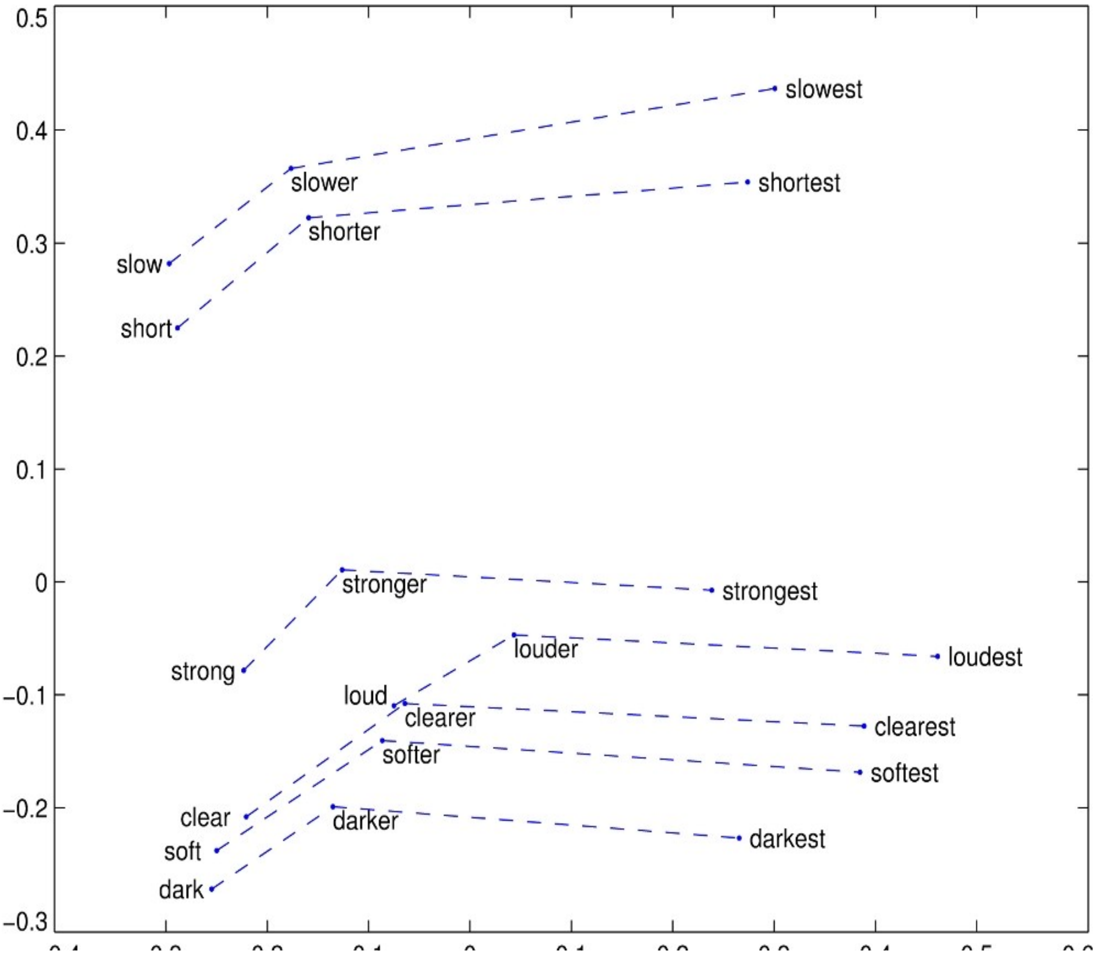


Figure 10 GloVe Embedding

GloVe is an approach to marry both the worldwide statistics of matrix factorization techniques like LSA with the local context-based learning in word2vec.

Instead of employing a window to define local context, GloVe constructs a certain word-context or word co-occurrence matrix using statistics across the entire text corpus. The result's a learning model that will lead to generally better word embeddings. We can see in Figure 10 that the words having semantic meaning are place together.

Before passing data to any deep learning model, we convert the text data to a numerical format. So, after getting topic vectors as output from KATE for each tweet, we give a similar GloVe embedding vector for that tweet using Cosine Similarity.

Cosine similarity is used to measure the similarity between two non-zero vectors. It is an inner product space that measures the cosine angle between them. Figure 7 shows the formula for the Cosine Similarity.

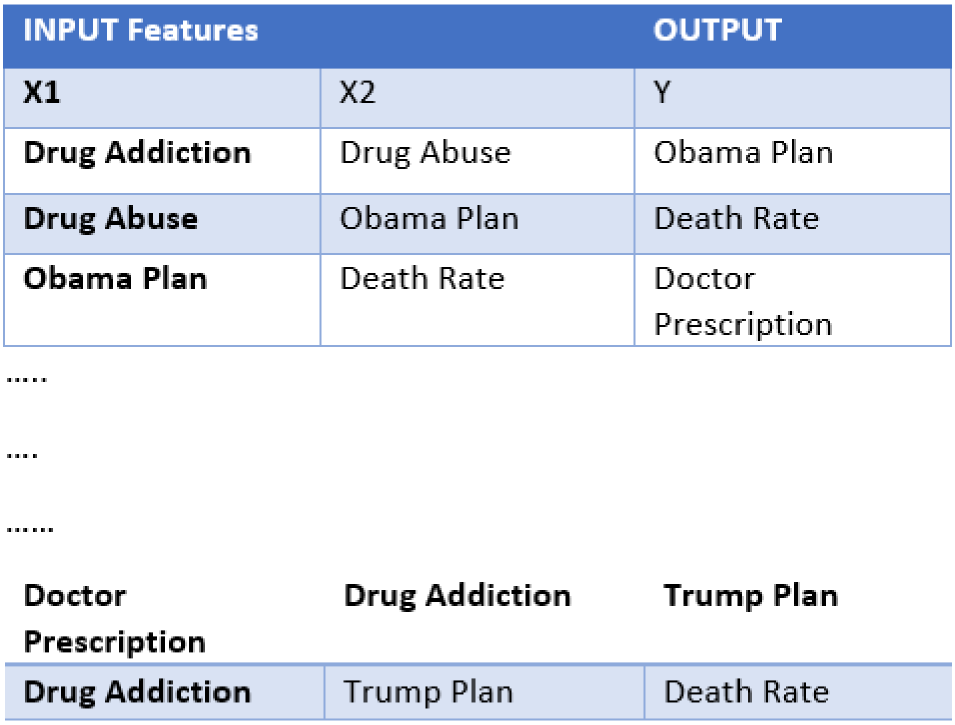
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 11 Cosine Similarity Formula

GloVe embedding is trained on twitter data to get a vector representation of each topic. The vector dimension was set to 10. Each Topic vector of KATE output was compared to the topic representation provided by GloVe. The topic vector, similar to GloVe embedding, is replaced with GloVe embedding.

3.3.2 Dataset Preparation

Before passing tweets to LSTM, we need to arrange our data in such a format like two months topics should be the input time steps to the model, and the model should be able to predict third-month topics. For this, we aggregated all the Glove embeddings for each month. And with time step as two, we formed a data structure with two months vectors as input features and third month as output feature. Figure 8 shows the Data structure.



INPUT Features		OUTPUT
X1	X2	Y
Drug Addiction	Drug Abuse	Obama Plan
Drug Abuse	Obama Plan	Death Rate
Obama Plan	Death Rate	Doctor Prescription
.....		
....		
......		
Doctor Prescription	**Drug Addiction**	**Trump Plan**
Drug Addiction	Trump Plan	Death Rate

Figure 12 Data structure prepared for the LSTM model

3.3.3 LSTM Model

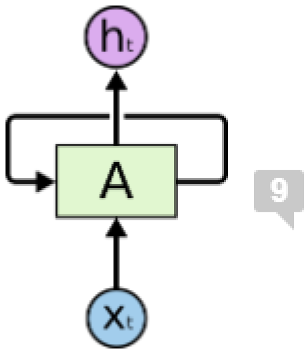
Humans don't start their thinking from scratch every second. As you read this essay, you understand each word supported your understanding of previous words.

You do not throw everything away and begin thinking from scratch again. Your thoughts have persistence.

Traditional neural networks can't do this, and it looks like a significant shortcoming. As an example, imagine you would like to classify what event is going on at every point in a movie. It's unclear how a standard neural network could use its reasoning about previous events within the film to tell later ones.

Recurrent neural networks address this issue. They're networks with loops in them, allowing information to persist.

Figure 13 RNN Neural Network



Recurrent Neural Networks have loops.

Sometimes, we only must observe recent information to perform the current task. For instance, consider a language model is trying to predict the following word supported the previous ones. If we are attempting to predict the last word in "the clouds are within the sky," we do not need from now on context – it's pretty apparent the following word goes to be the sky. In cases where the gap between the relevant information and the place that it's needed is little, RNNs can learn to use past information.

Long memory networks – usually just called "LSTMs" – are a special quite RNN, capable of learning long-term dependencies.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the shape of a sequence of repeating modules of the neural network. In standard RNNs, this repeating module will have a simple structure, like one tanh layer.

LSTMs even have this chain-like structure, but the repeating module contains a different structure. Rather than having one neural network layer, there are four interacting in an unprecedented way.

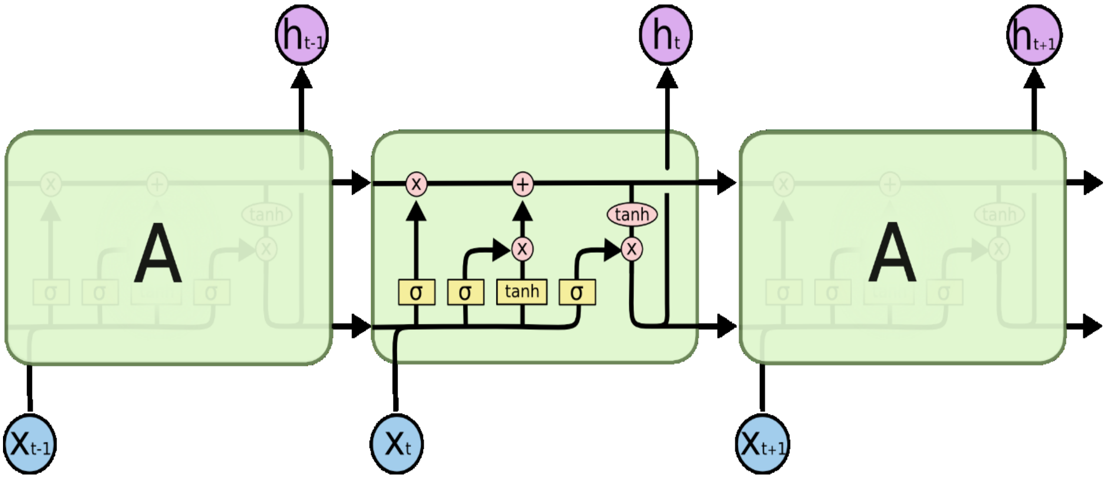


Figure 14 The repeating module in an LSTM contains four interacting layers.

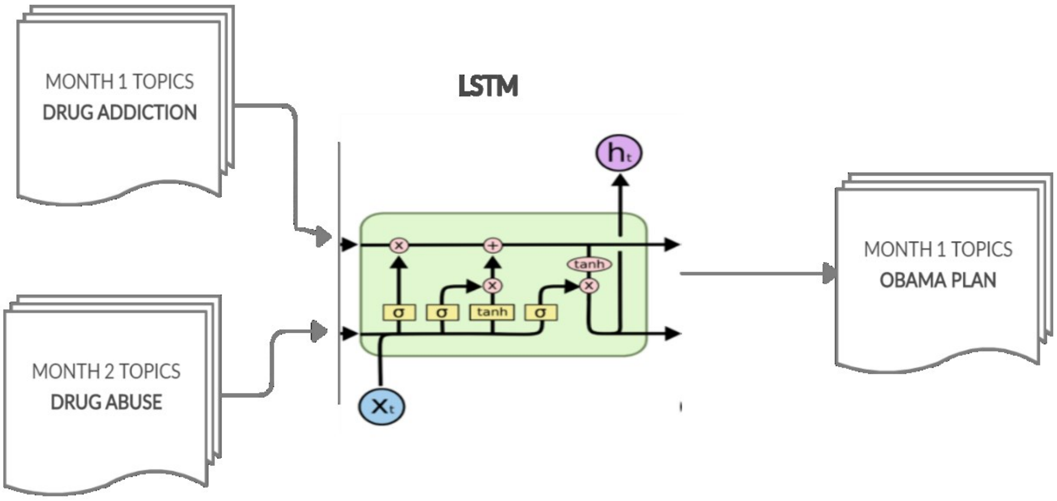


Figure 15 LSTM MODEL

So, we used the LSTM Model for time series topic forecasting. The data we created has two months as time steps. We designed a neural network model with the first layer as LSTM, with 125 LSTM cells. Second Layer as Dropout to avoid overfitting the model. The last layer is having 10 Neurons, which is used to produced vector with ten dimensions. The resulting vector is compared with GloVe to know which topic is predicted. Figure 14 shows the architecture of our Prediction Model. In Figure 15, we can see the overall use of LSTM in our framework.

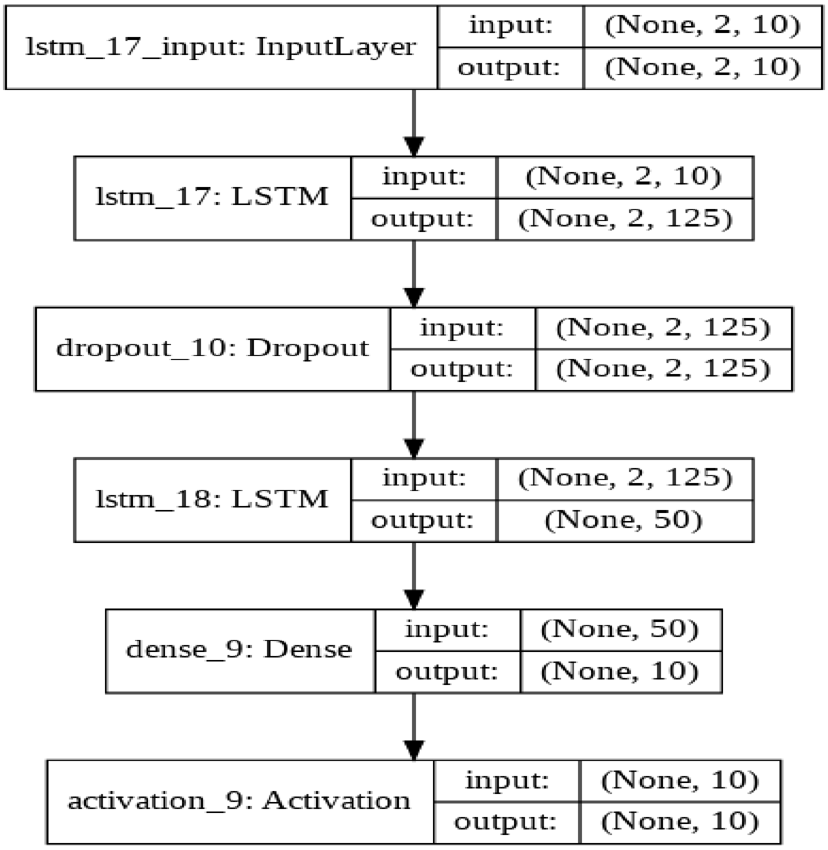


Figure 16 LSTM Model Architecture

CHAPTER 4
RESULTS AND EVALUATIONS

4.1 Data Collection and Classification

We collected a total of 133,107 tweets. All the tweets were then classified into personal and news tweets using the MPQA dictionary. Different settings were used to classify the tweets as if the tweet has two strong subjective clues, and two weak subjective clues are said to be Personal Tweet else a news tweet. Among different settings, we found one strong subjective clue and one weak subjective clue to be the best threshold that segregates Personal and News tweets. Table 1 shows the number of Personal and News Tweets every year. And Table 2 shows the number of Newspaper articles every year. As CDC is also a source of data, we got a number of deaths each year due to Drug overdose from CDC. Table death shows the number of deaths every year. We see a strong correlation between death counts, number of tweets, and the number of articles. In 2017 the number of deaths increased so as to the number of tweets and articles. We applied the Pearson Coefficient to see if there is a correlation between them. The value was 0.9, which says that there is a strong correlation between the number of deaths and tweets every year.

Table 1: Number of Tweets from different sources

	Number of tweets		
Year	Total	Personal	News
2010	159	0	159
2011	708	4	704
2012	2238	4	2234
2013	3676	14	3662
2014	5610	35	5575
2015	11189	14	11175
2016	27298	56	27242
2017	45841	441	45400
2018	33094	2105	30989
2019	32947	1955	30992

Table 2: Number of Newspaper Articles

Year	Number of Articles
2010	6
2011	14
2012	16
2013	21
2014	40
2015	46
2016	251
2017	680
2018	942
2019	892

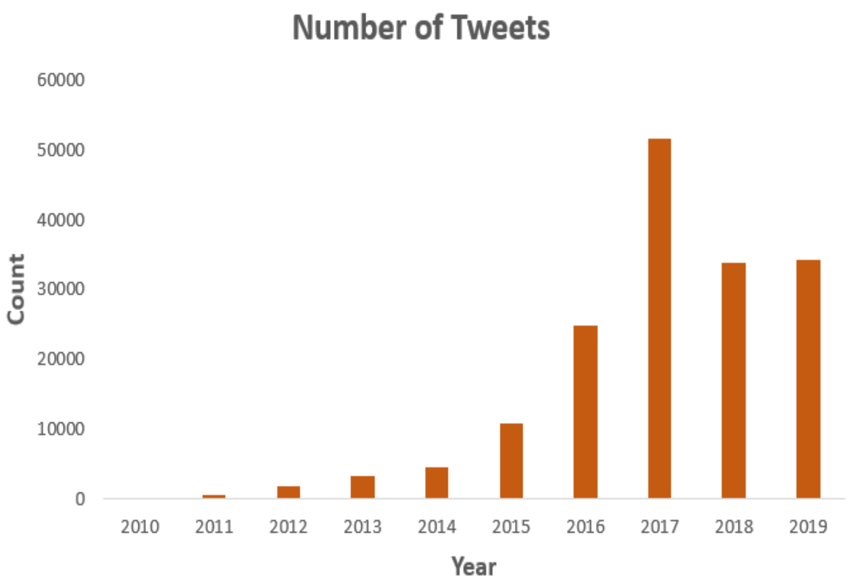


Figure 17 Number of Tweets

Figure 18 shows the graph of the number of tweets in every year from 2010 to 2019. Y-axis is marked with the number of tweets and X-axis with the year. We see that there is a drastic increase in the number of tweets in 2017. By this, we can say that public concern has increased in 2017, which might be due to the number of deaths.

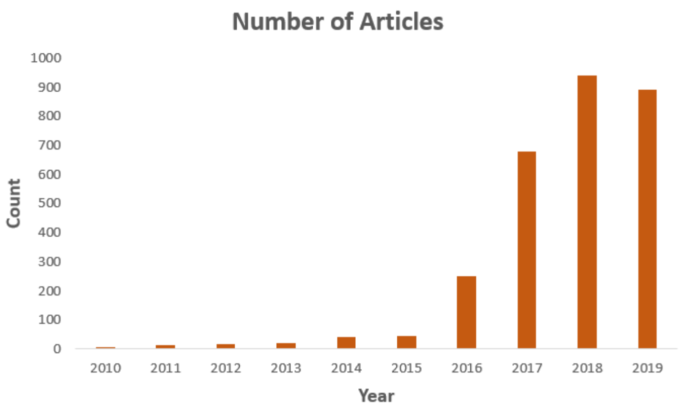


Figure 18 Number of Articles

Figure 17 shows the graph for the number of New York Times articles in each year from 2010 to 2019. New York Times here is considered as Traditional News Source. So, Traditional News Source also has an increase in the number of articles from 2017. In 2016, the number of articles was between 200 and 300, but in 2017 it increased to 700.

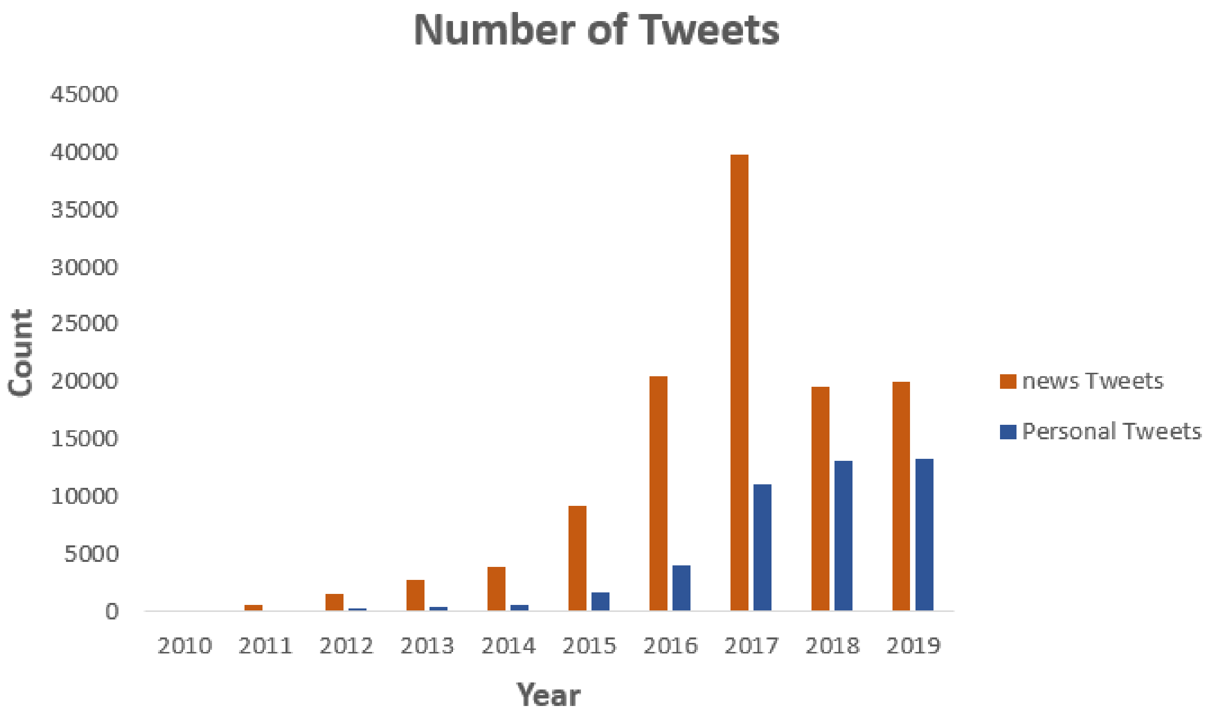


Figure 19 Number of Personal and News Tweets After Classification

Figure 19 shows the graph of the number of personal tweets and news tweets after the classification of twitter data using the MPQA dataset. We see that the number of Personal tweets is less compared to News tweets; this is one of the limitation to this work. In the future, we plan to collect more personal tweets. We also see that the number of Personal and news tweets increased in 2017.

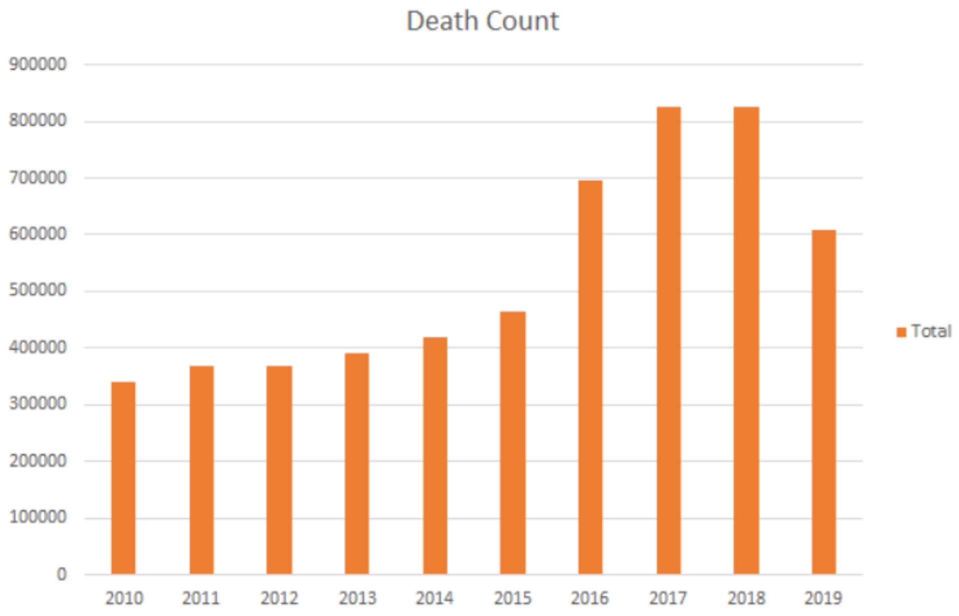


Figure 20 Number of Deaths from CDC

Figure 20 shows the number of deaths every year from 2010 to 2019 due to drug overdose. The X-axis gives the death count each year, and Y-axis is marked with the years 2010 to 2019. We see that death due to drug overdose has increased from 2016. This might be the reason why opioids became a hot topic in 2017, and the number of tweets and articles have a peak in 2017. To evaluate this, we conducted Pearson Correlation. The result from the Pearson correlation was 0.9, which says that the number of deaths due to drug overdose is highly correlated with the number of tweets tweeted and articles published. By this, we can interpret that the public concern on the opioid epidemic has increased when the death rate increased.

4.2 Topic Trends and Sentiment Analysis

4.2.1 Sentiment Analysis

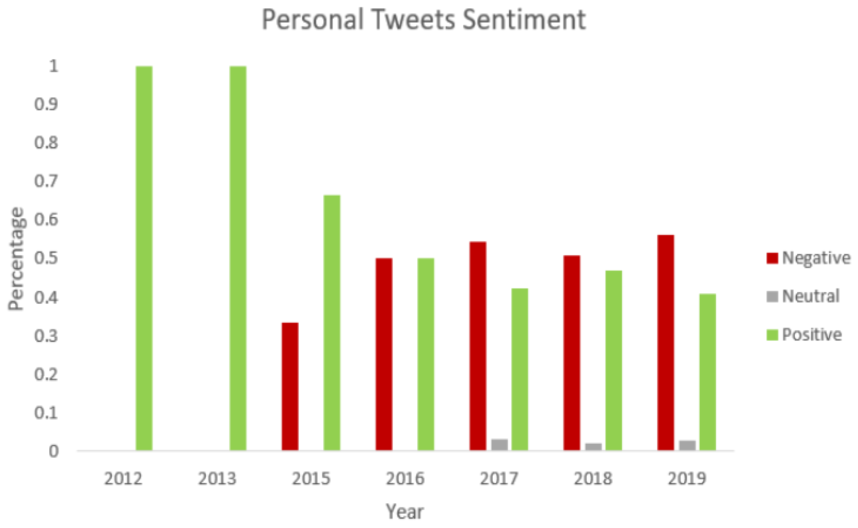


Figure 21 Sentiment of Personal Tweets

Figure 22 shows the sentiment graph for Personal tweets after VADER sentiment analysis. We see that the public showed negative emotion towards opioids from 2015 and increased rapidly.

Y-axis shows the percentage of tweets showing different emotions. The X-axis is marked with years from 2010 to 2019. The reason we do not see any sentiment for the years 2010 and 2011 is that the number of personal tweets is negligible in 2010 and 2011. The green color bar in the graph shows positive sentiment attached to tweets, and Red color shows negative sentiment attached to tweets.

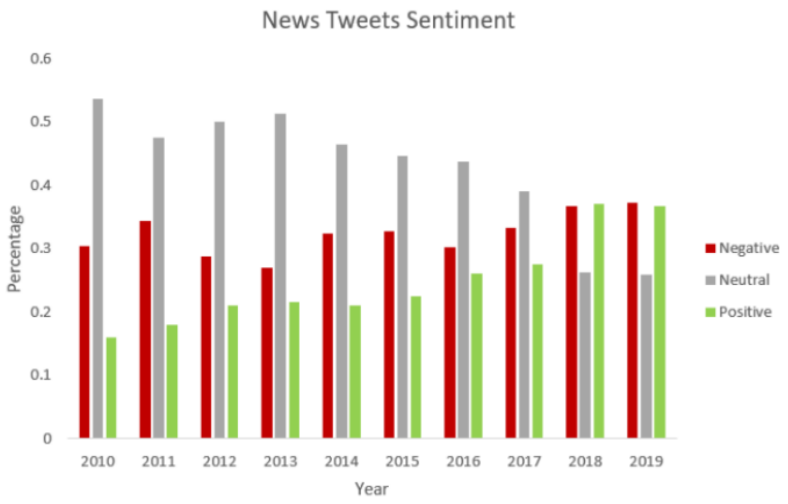


Figure 22 Sentiment of News Tweets

Figure 21 shows the sentiment graph for News tweets. We see that the News tweets does not have any biased sentiment. The percentage of positive tweets every year is almost similar to the percentage of negative tweets.

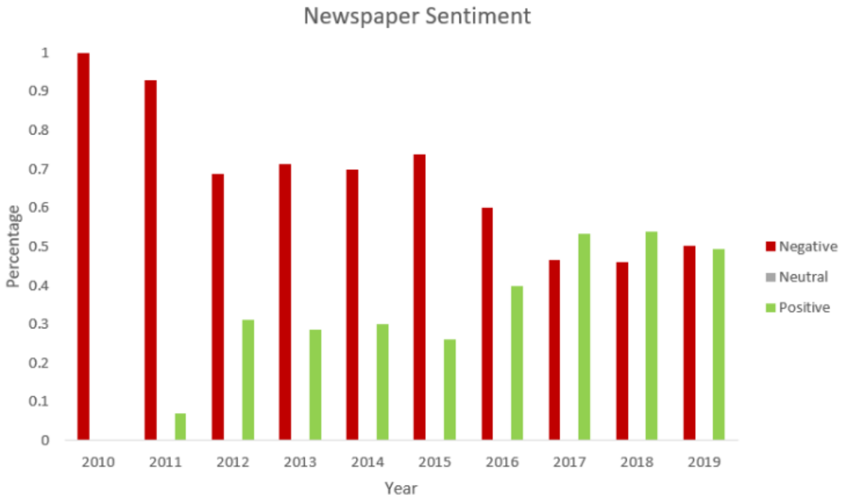


Figure 23 Sentiment of Newspaper Tweets

Figure 23 shows the graph of sentiment for Newspaper articles from the year 2010 to 2019. We see that Newspaper articles show the most negative impact of opioids during the Obama period .i.e. from 2010 to 2015. From 2016 Newspaper articles tried to maintain a positive environment. Even though the number of deaths increased in 2017, newspaper articles tried to show a positive environment.

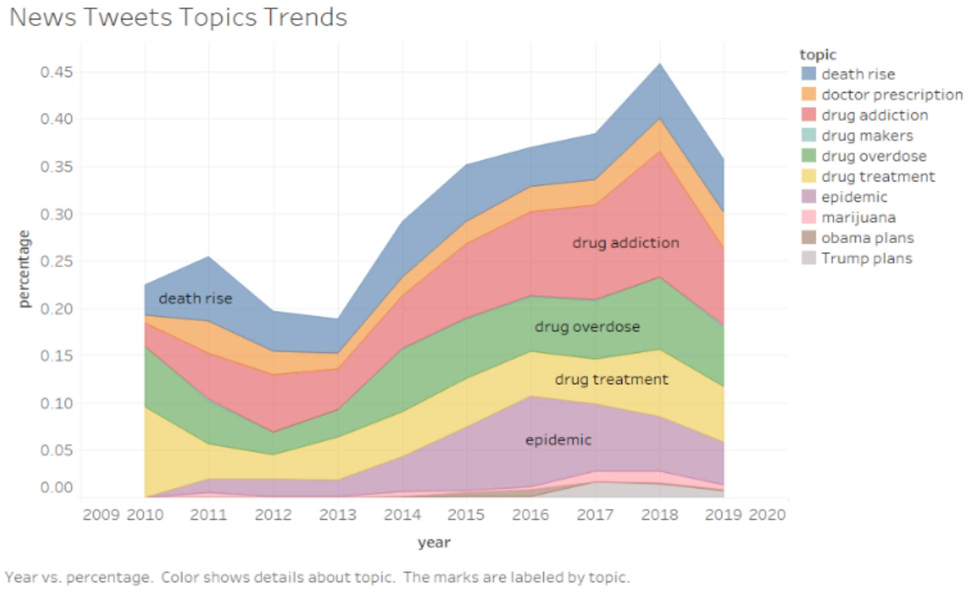
4.2.2 Topic Modelling

Topic Name	Topic Terms
Drug Addiction	dare, addiction, problem, monkey, someone, suffer, acknowledge, time, quit, humili
Drug Treatment	safety, care, opioid, plan, educ, safe, primary, approve, treatment, company
Drug Overdose	death, deter, abuse, advice, gener, increase, drug, overdose, best, tell
Doctor prescription	the drug, doctor, prescript, monitor, overprescribe, Canada, Educ, prescribe, painkiller, target
Obama plan	Obama, handle, plan, crisis, released
Trump plan	the program, fight, little, maintain, history, critic, trump, hous, already, decent
Drugmakers	fight, crisis, drugmakers, opioid limit, limit
Epidemic	epidemic, pain, crisis, address, task, guideline, forc, tune, tackle, combat

Death rates	patients, overdose, health, years, million, percent, deaths, addiction, products, likely
Marijuana	pharma, drug, epidemic, legal, fight, company, abus, state, marijuana, declare

Table 3: Topics from KATE Topic Modelling

KATE Topic Modelling was assigned with ten topics as setting, so; KATE generated ten topics with ten words for each topic. By seeing the topic words, each topic is assigned a name. Topics are evaluated qualitatively using [4]. Table 3 shows the topic words and names assigned to each topic. Using those topics, we found the frequency of topics every year in all the sources of data and generated trends of those topics in different sources.


News Tweets Topics Trends

Year vs. percentage. Color shows details about topic. The marks are labeled by topic.

Figure 24 News Tweets Topic Trend

Figure 24 shows the area graph for different topics identified by KATE from News tweets. Using this area graph, we can see the trend for each topic. Drug Addiction and Drug Overdose was mostly covered in News tweets. We see that the mention of death rise kept increasing, which is similar to the death count trend.

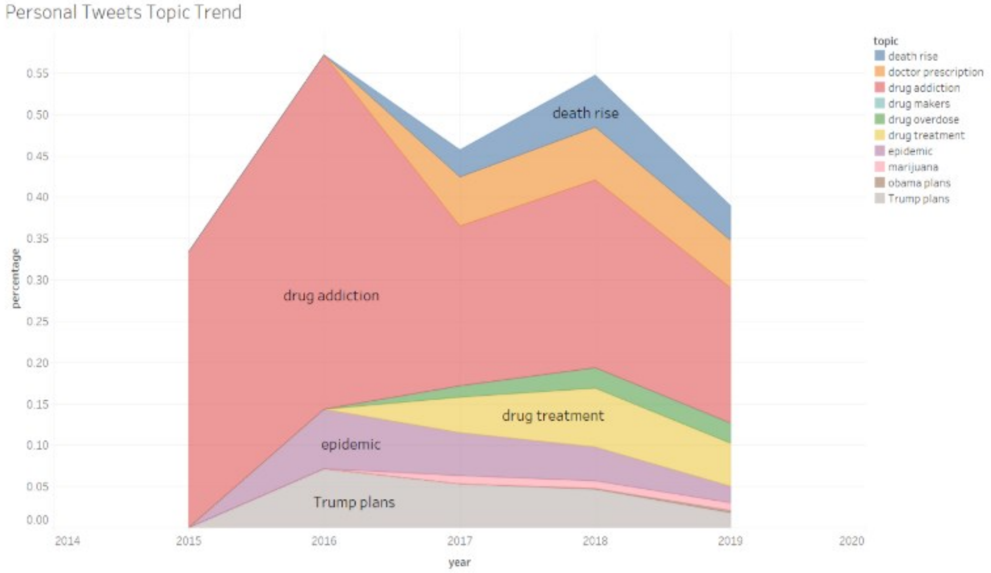


Figure 25 Personal Tweets Topic Trend

Figure 25 shows the area graph for topics identified by KATE from Personal tweets source. We see that the public mention most about Drug Addiction. Trump's plan was also covered in Personal tweets. We also see that the mention of the death rate is also increasing from 2017. We can interpret that the public is following the CDC website, and as the death count is increasing, the number of tweets is increasing and also the mention of the death rate.

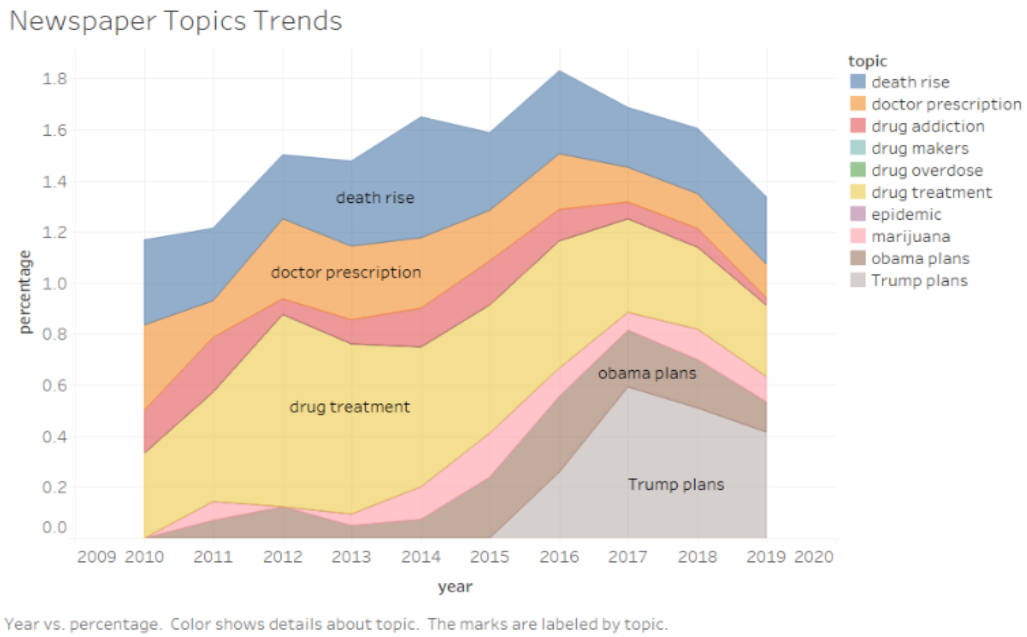


Figure 26 Newspaper Topic Trend

Figure 26 shows the trends of topics from Traditional Source. We see that Trump Plan and Drug Treatment was mostly covered in Newspaper Articles. Trump's plan was the most covered topic in the year 2017.

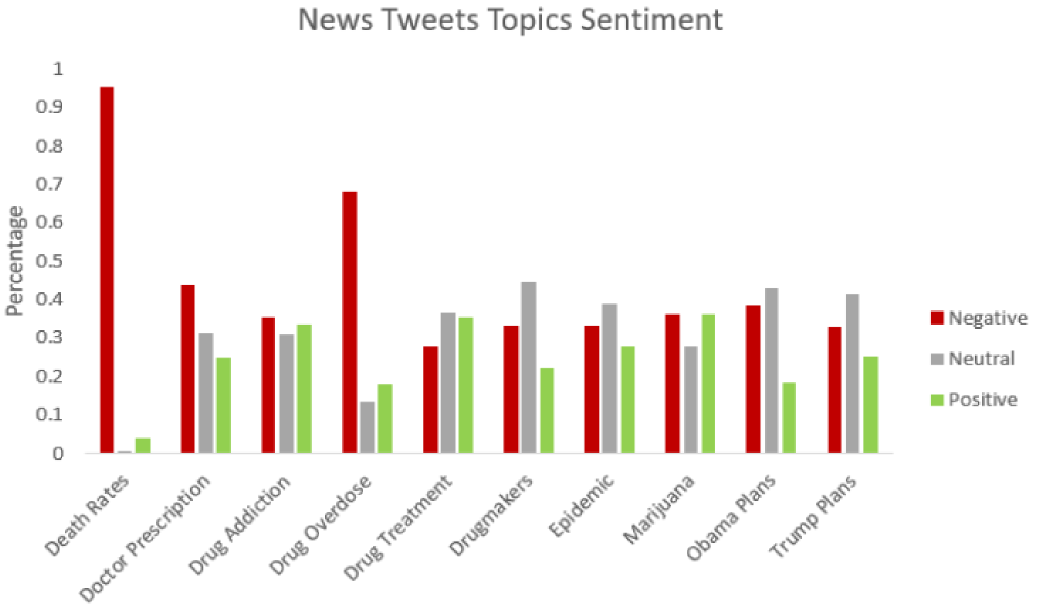


Figure 27 News Tweets Topic Sentiment

Figure 27 shows the percentage of sentiment for Different topics from News tweets. We see that News tweets show negative sentiment towards Drug overdose and Death rates. The X-axis is marked with different topics, and Y-axis is marked with a percentage of tweets.

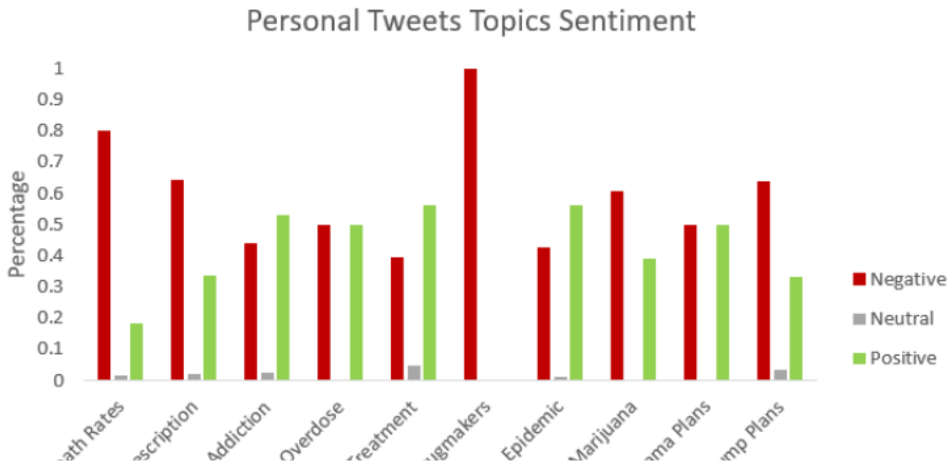


Figure 28 Personal Tweets Topic Sentiment

Figure 28 shows the sentiment graph for topics from Personal tweets. Personal tweets show negative sentiment towards trump plans, which were most mention topics in Personal tweets and also negative sentiment towards death rates. We also see that most of the topics related to opioids have a negative sentiment. We can interpret that the public was not happy with trump's plans.

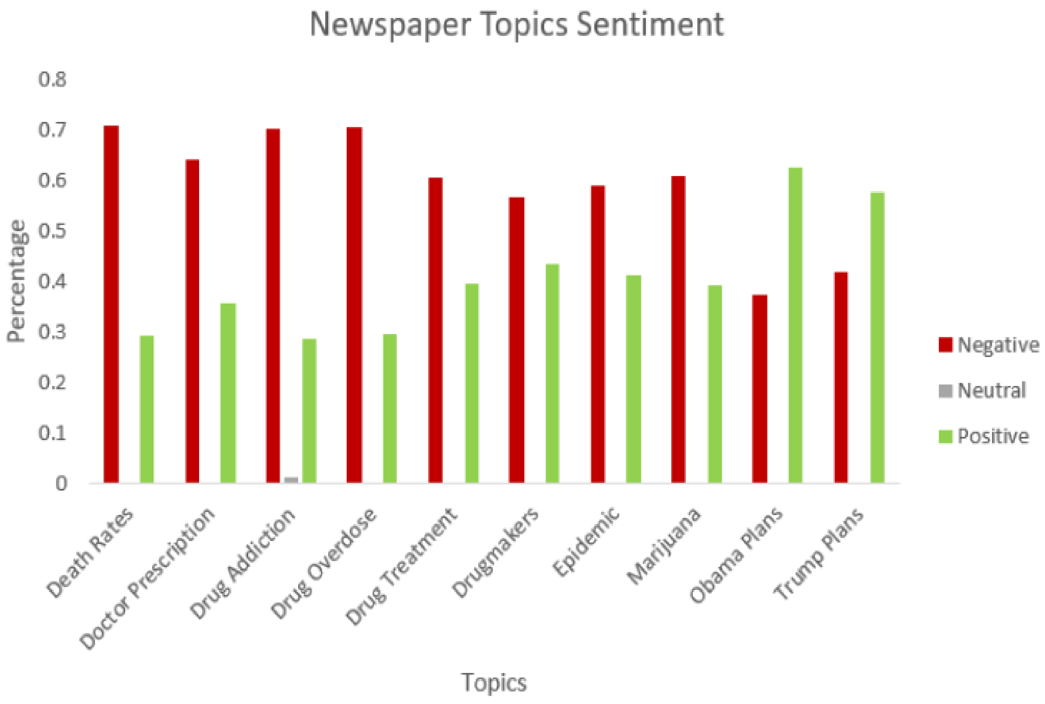


Figure 29 Newspaper Topic Sentiment

Figure 29 shows the sentiment graph for topics from Newspaper articles. Newspaper articles show positive sentiment towards trump and Obama's plans. We see that trump's plan was mostly covered in the year 2017, and Overall, we have positive sentiment in the year 2017, including trump's plan. By this, we can interpret that Traditional News is trying to set a positive environment to support Trump Plans even though the death rate was increasing in the year 2017.

To support this statement, we went through the articles in the year 2017. Below are the articles which support our interpretation.

- Trump Administration Announces $1.8 Billion in Funding to states to continue combating the opioid crisis.
- Trump declares the opioid crisis a national health crisis.

4.2.3 Different Opioid Names

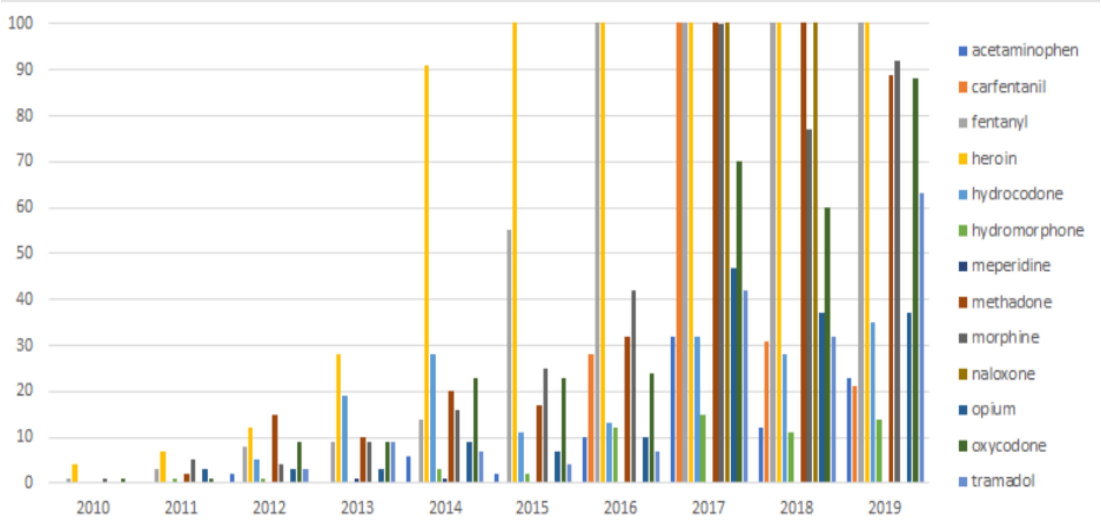


Figure 30 Mention of Different Opioid names

Figure 30 shows the mention of different types of opioids over the year 2010 to 2019 in personal tweets. We see that mention of heroin was increasing from 2014. The X-axis shows the years from 2010 to 2019, and Y-axis shows the percentage of tweets. Among all the opioids listed, Morphine and Heroine were most mentioned in Personal Tweets.

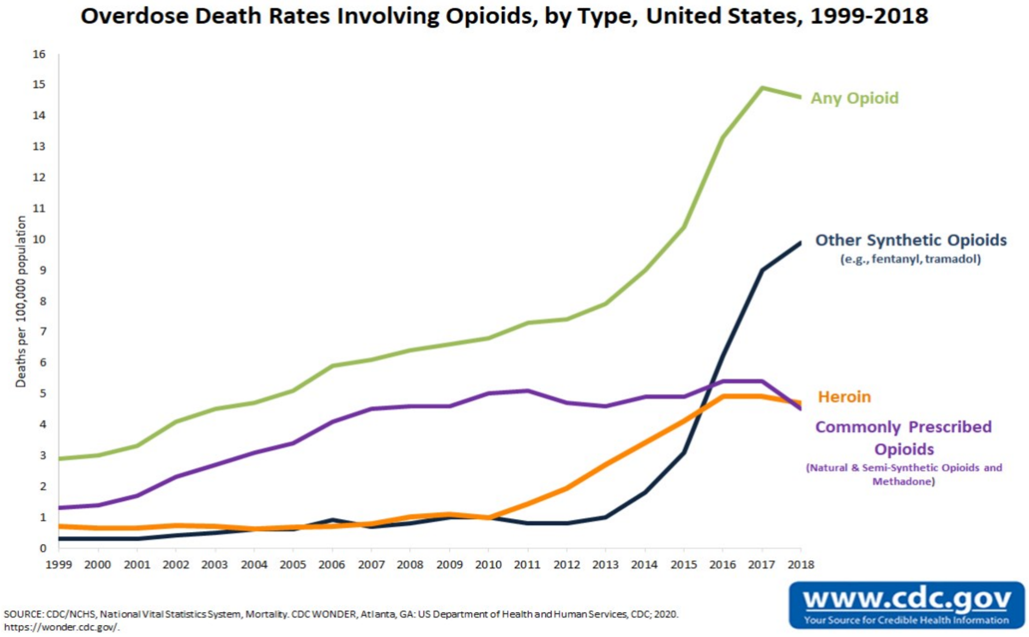


Figure 31 Deaths due to different drugs from the CDC

Figure 31 shows the number of deaths in different years due to different types of a drug overdose. This data was collected from the CDC website. From this figure, we see that deaths due to heroin increased from 2014. We also see that mention of heroin increased from 2014 in Personal tweets.

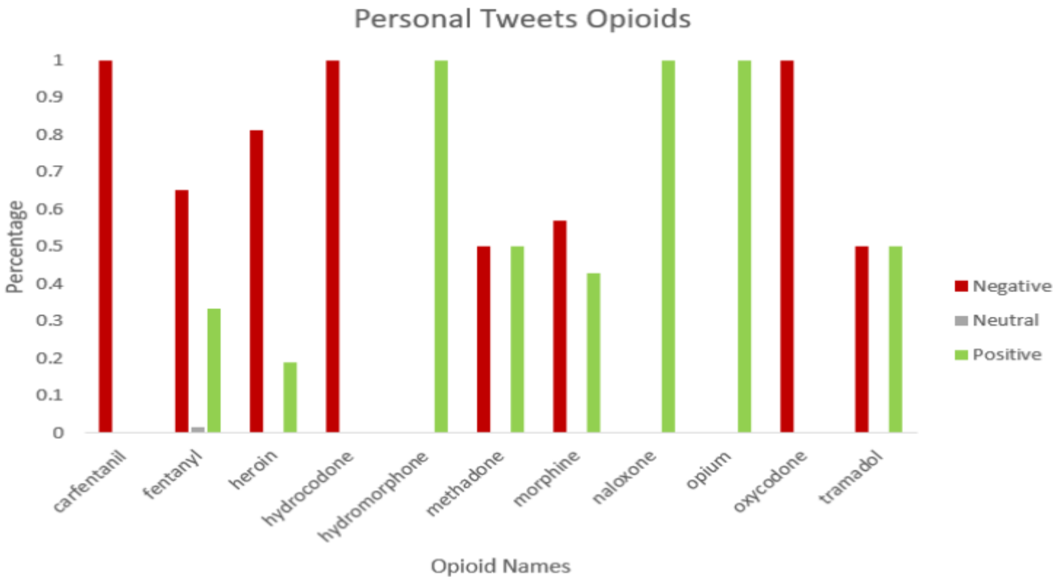


Figure 32 Sentiment of Different opioid from Personal Tweets

Figure 32 shows the percentage of the different sentiment of different drugs. We see that heroin had a substantial negative impact on the public. So by combining Figures 29, 30, and 31, we can say that as the number of deaths due to heroin increased from 2014 public had a negative impact of this and the mention of heroin increased in twitter from 2014 with negative emotion.

4.2.4 Psychological Categorisation

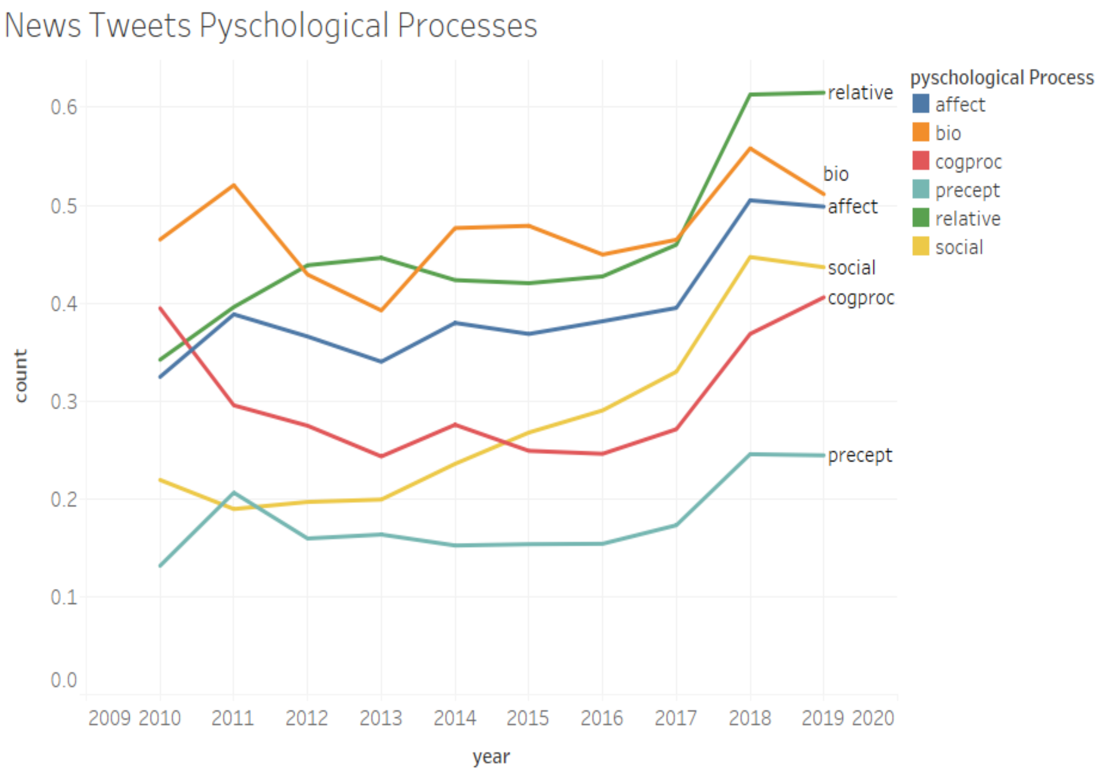


Figure 33 News Tweets Psychological Processes

Figure 33 shows the trend of Different Psychological Processes identified by LIWC in the News tweets. We see that news tweets mostly show the relative psychological process. It means that the news tweets are trying to give news about opioids using related words like related to public or politics etc.

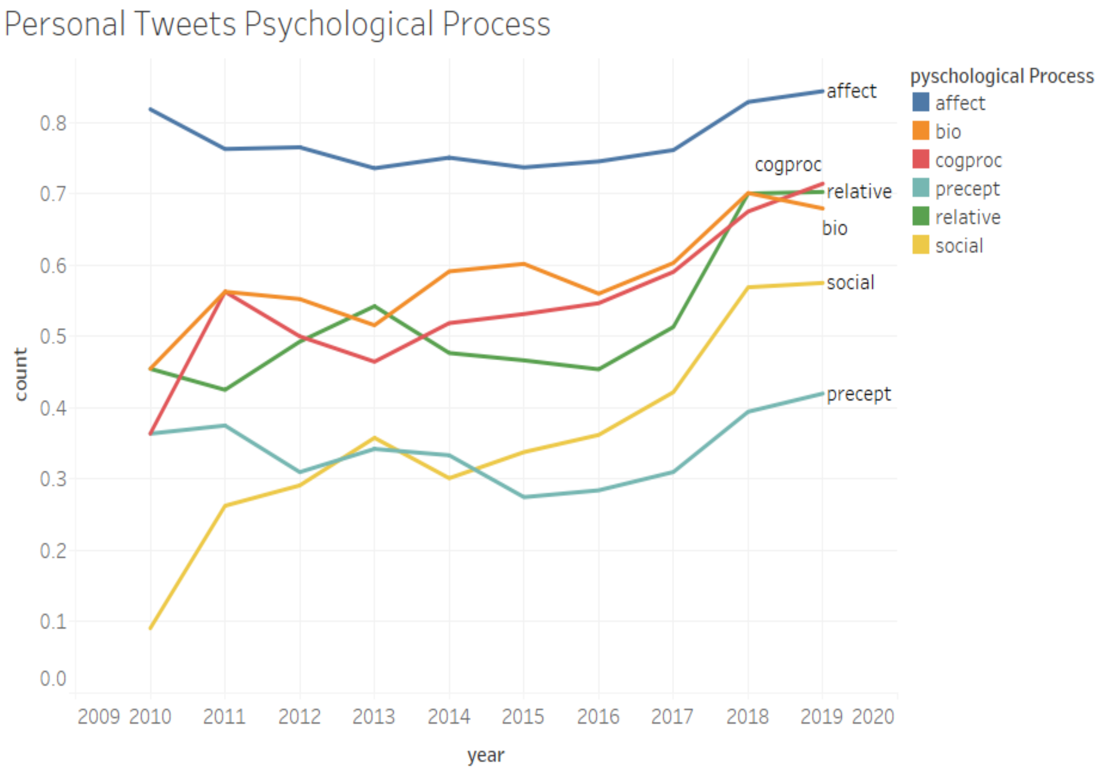


Figure 34 Personal Tweets Psychological Processes

Figure 36 shows the trend of different psychological processes in personal tweets. We see that affective processes were found by LIWC to be most used in personal tweets. It means that personal tweets are trying to give their opinions using affective statements, which says how they are affected by this epidemic.

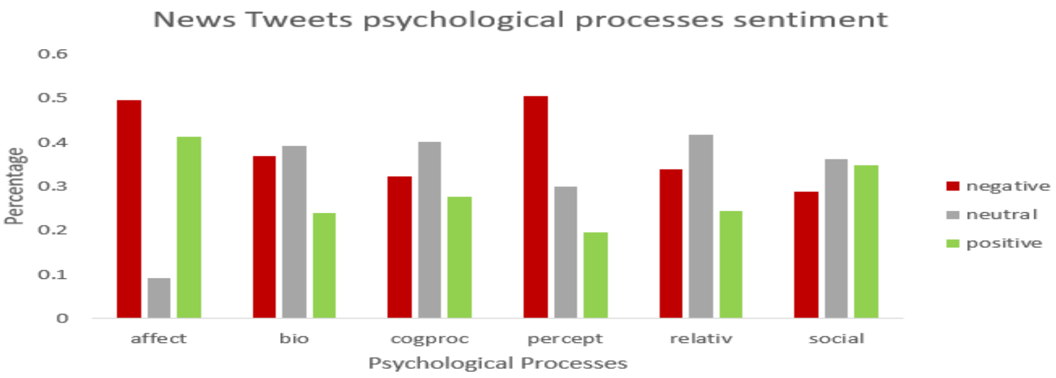


Figure 35 Sentiment for News Tweets Psychological Processes

Figure 34 shows the sentiment of different psychological processes in news tweets. We see that relative processes show negative sentiment. So by this, we can say that news tweets are trying to portray the news relative to the public and with negative emotion.

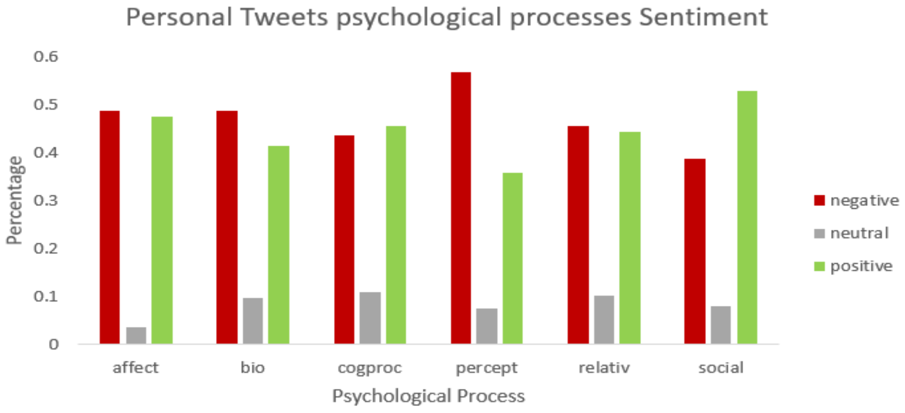


Figure 36 Sentiment for Personal Tweets Psychological Processes

Figure 35 shows the sentiment of different psychological processes in personal tweets. We see that affective processes have negative emotions in personal tweets. By this, we can say that the public is portraying their opinion to show the negative impact of opioids on public life.

4.3 Topic Prediction

After applying KATE topic modeling on Personal Tweets, we got topics and also topic vectors for each tweet. We also trained the GloVe model on Personal Tweets, and then we got GloVe embedding for each tweet. We found the cosine similarity of each topic vector with GloVe embedding of topic names. Then GloVe embedding of the topic with less cosine similarity was assigned to each tweet. We aggregated the embedding for each month and passed it to LSTM Model. The data was arranged as two months as an input feature and the third month as output. Overall, we had 120 rows to train LSTM Model. Figure 9 model shows the architecture of the LSTM model we used to train.

For test data, we passed January 2011 and February 2011 with topics drug addiction and drug overdose respectively as input to the model, and the model predicted that Obama's plan would be the topic to be discussed about the most in March 2011.

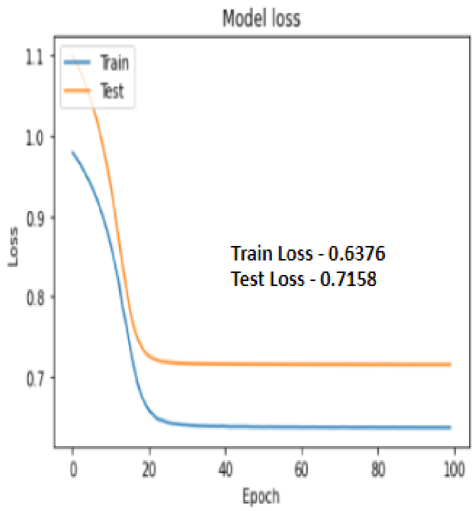


Figure 38 Loss graph for LSTM

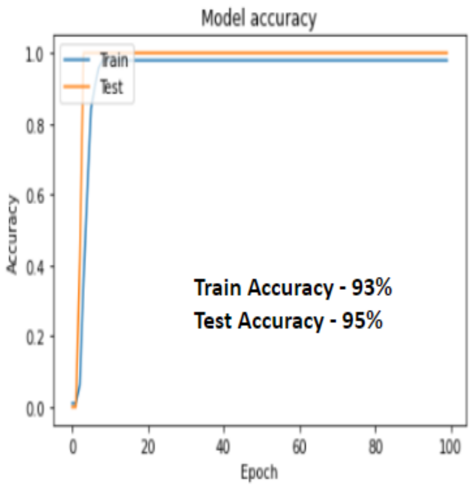


Figure 37 Accuracy graph for LSTM model

Figures 37 and 38 show the accuracy and loss graph of our Prediction model. We got training accuracy as 93% and test accuracy as 95%. We can say that our model is working well. But when we see the Loss graph, we have a high loss for the test. We can say that our model is slightly overtrained. This is due to when we converted our topics to monthly topics for ten years; we got over all of 120 rows, which is less amount of data to train the model. So in the future, we plan to collect more data and arrange the topic in a weekly manner to get more rows and better results.

CHAPTER 5
CONCLUSION

The opioid epidemic has become a major national public health problem that demands better understanding, which is limited by the lack of large-scale data. Social media Twitter, which comes with community curated content, provides a candid medium that makes large scale data available to understand the topics, emotions, psychology, and opinions of the public. Our initial study using twitter and Newspaper demonstrates that information provided an essential and useful resource to understand the opioid epidemic from users' discussions on a large scale. This provides the first step for more in-depth studies to provide insights that may help to combat the opioid crisis. And this study says that the topics spoken on Twitter and published in Newspaper articles are the same but have different trends and attitudes. Public opinion mostly change based on Newspaper articles and Government Data Sources like CDC. The prediction model is also able to detect the future concerns of the public. The Future scope for this work would be to compare our model with other baseline forecasting models, utilizing other datasets such as Reuters, USA Today, The Wall Street Journal, and Include different Case Studies.

BIBLIOGRAPHY

1. Y. Chen and M. Zaki. "KATE: K-competitive autoencoder for text," in *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2017, pp. 85–94. https://doi.org/10.1145/3097983.3098017.
2. Z. Tong and H. Zhang. "A text mining research based on LDA topic modelling," *Computer Science & Information Technology*, vol. 6, pp. 153-167, May 2016. [Online} doi: 10.5121/csit.2016.60617 .
3. R.Yao et al. "A prior knowledge-based neural attention model for opioid topic identification." In: *IEEE International Conference on Intelligence and Security Informatics(ISI)*,2019 IEEE. Doi: 10.1109/isi.2019.8823280.
4. E.M. Glowacki, J.S. Glowacki, G.B. Wilcox. A text-mining analysis of the public's reactions to the opioid crisis. *Subst Abus*.2018;vol.39,no. 2:129-133. DOI:10.1080/08897077.2017.1356795
5. J. A. Lossio-Ventura and J. Bian, "An inside look at the opioid crisis over Twitter," *2018 IEEE International Conference on Bioinformatics and Biomedicine* (BIBM), Madrid, Spain, 2018, pp. 1496-1499, DOI: 10.1109/BIBM.2018.8621101.
6. Y. Fan, Y. Zhang, Y. Ye, X. li, and W. Zheng. 2017. Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from Twitter and case studies. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (CIKM '17). Association for Computing Machinery, New York, NY, USA, 1259–1267. DOI:https://doi.org/10.1145/3132847.3132857
7. T.K. Mackey, J.Kalyanam, T. Katsuki, and G. Lanckriet. "Machine learning to detect prescription opioid abuse promotion and access via Twitter.*" American Journal of Public Health* vol.107, no. 12: pp. 1910-1915, 2017.

8. H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert Systems with Applications*, vol. 115, pp. 27–36, 2019. DOI:10.1016/j.eswa.2018.07.051.
9. J. Hurtado, S. Huang, and X. Zhu, "Topic discovery and future trend prediction using association analysis and ensemble forecasting," *2015 IEEE International Conference on Information Reuse and Integration*, 2015. DOI:10.1109/IRI.2015.40
10. S. Bradshaw & R.Hewett & F. Jin. Discovering opioid use patterns from social media for relapse prevention. arXiv: 1912.001122, 2019
11. Y. Wu, P. Skums, A. Zelikovsky, D. C. Rendon, and X. Liao, "Predicting opioid epidemic by using Twitter data," *Bioinformatics Research and Applications Lecture Notes in Computer Science*, pp. 314–318, 2018. DOI:10.1007/978-3-319-94968-0_30.
12. S. Pandrekar, X. Chen, and G. Gopalkrishna, "Social media based analysis of opioid epidemic using Reddit," *AMIA Annual Symposium proceedings*., Dec-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371364/.
13. A. H. Hossny and L. Mitchell, "Event Detection in Twitter: A keyword volume approach," *2018 IEEE International Conference on Data Mining Workshops* (ICDMW), 2018. DOI:10.1109/ICDMW.2018.00172.
14. A. Saleh and A. Scherp, "Attend2trend: Attention model for real-time detecting and forecasting of trending topics," *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Singapore, Singapore, 2018, pp. 1509-1510, DOI: 10.1109/ICDMW.2018.00222.
15. Center for Disease Control and Prevention. "Opioid data analysis and resources" [online]. Available: http://www.cdc.gov/drugoverdose/data/analysis.html.

VITA

Aashish Thota completed his bachelor's degree in Computer Science Engineering from Jawaharlal Nehru Technological University in Hyderabad, INDIA. He started his master's in computer science at the University of Missouri-Kansas City (UMKC) in August 2018, with an emphasis on Data Sciences and graduates in May 2020. While he was studying at UMKC, he worked as a Graduate Assistant for Python/Deep Learning Course.