AN APPROACH TO QUANTIFY INFORMATION IN TWEETS

By

RUCHISHYA RAMINENI

Bachelor of Computer Science

Jawaharlal Nehru Technological University

Hyderabad, Telangana, IN

2016

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
MAY 2019

AN APPROACH TO QUANTIFY

INFORMATION IN TWEETS

Thesis Approved:

Dr. K.M. George

Thesis Advisor

Dr. Johnson P Thomas

Dr. Esra Akbas

ACKNOWLEDGMENTS

The Master degree from Computer Science department, Oklahoma State University has given me an immense experience and knowledge in my fields of interest, I would take interest to thank my Thesis Advisor and Head of Computer Science Department, Oklahoma State University, Dr. K. M. George for his continuous assistance and encouragement to learn new technologies.

I would like to express my gratitude to committee members, Dr. Johnson Thomas and Dr. Esra Akbas for their guidance and support.

Finally, I would express my profound gratitude to my family and friends who supported me throughout my years of study.

Name: RUCHISHYA RAMINENI

Date of Degree: MAY 2019

Title of Study: AN APPROACH TO QUANTIFY INFORMATION IN TWEETS

Major Field: COMPUTER SCIENCE

Abstract: Microblogs such as Twitter play an important role in online social communications. Unlike traditional media, hot topics and emerging news will become much more popular in a short span with the help of information spreading platforms like Twitter. Nowadays Twitter is widely used in many professions to analyze data. For example, sentiment analysis is the popular approach to opinion mining where the sentiment values of the tweets are classified into weighted classes positive, negative or neutral. These signed weights may not be the best approach for analysis in all cases. Information diffusion is an alternative method to analyze the information defined as information passing through person to person where the research mostly focuses on graph based models. The edges of the network graph are constructed based on either retweet status or hashtags, and information flow is modelled as transmission from node to node where nodes are users.

Generally speaking, analysis of tweets quantify information inherent in tweets. In this research, a new approach is proposed to quantify information in tweets as unsigned weights. This approach is suitable to analyze problems if tweets can be interpreted to convey unsigned weight contribution to the problem. The weight computation method presented in this thesis extract keywords called tokens from tweets. Then weights are associated with tokens. The weights are interpreted as quantification of information. To identify tokens two methods are used, one approach uses a technique in Topic Modeling LDA (Latent Dirichlet allocation) to determine tokens and their weights. The second approach is iterative which starts with some anchor words (keywords set) and with similarity measure between anchor word set and the words in tweets. More words are added based on some threshold value of similarity. To associate weights to tokens NMF (Non-numeric Matrix Factorization) is used. To compute weight contribution of a tweet, a formula for its potential is used.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Twitter users and Social media users continue to increase steadily as shown in figure 1a and 1b. Users of these platforms simultaneously generate and consume information. Nowadays the internet is replacing the traditional media (as shown in figure 1.3 (Statista)). These data provide the justification for searching the information from social media. Micro-blogging sites like Twitter can be viewed as a social network or information network. It has become the source of information where people post their real-time experiences and their opinions on various day-to-day issues which can be used to predict and analyze the data. This information can be either explicit or implicit in social media sites. There are numerous papers, example O'Connor et al. (2010), Ribeiro et al. (2016), Yang and Leskovec (2010), that analyze Tweets and other text data by extracting information. Nowadays, Twitter is the most common platform for Big Data analysis. Due to the size, speed, and variety of these tweets and posts, they fit the characterization of big data, and hence, big data-related environments and tools are used in data collection and analysis.

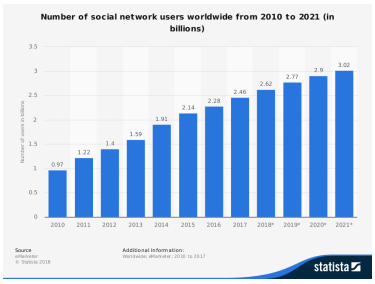

Figure 1.1: Twitter user growth

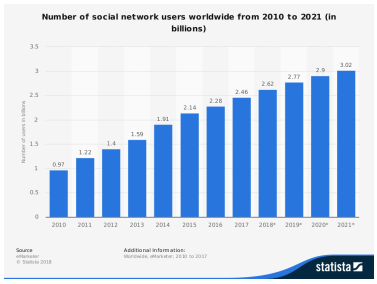

Figure 1.2: Social Media user growth

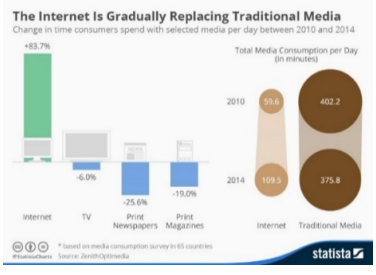

Figure 1.3: Internet vs traditional media

This thesis deals with a new approach for quantifying and computing information in tweets. We combine and expand ideas, concepts, and formulae gleaned from different papers to reach our objective which is to develop an objective method to quantify information in a collection of tweets. This approach is expected to provide a different analysis method for social media data, especially Twitter data for explanation and forecasting entities such as political momentum. Sentiment analysis is a popular approach to analyze social media data Ribeiro et al. (2016). Ahmed et al. (2015) provide an overview of sentiment analysis over social networks. Another method of social media analysis is information diffusion. Information diffusion papers mostly consist of mostly graph-based models. In this thesis, we follow a different approach that centers on the idea of tweet potential presented in TK et al. (2015). At the concept level, the potential of a tweet can be viewed as its contribution to the information measure we are interested in. The potential of a tweet depends on the words contained in the tweet and is computed as the sum of the weights of the words present in the text.

And we are also interested in comparing our model of quantifying information with other quantifying information approaches such as sentiment analysis and information propagation in twitter. The analysis is done on different type of datasets like food poisoning and Immigration.

The analysis is done on food poisoning data as the study of Foodsafety.gov has estimated that each year, millions of people in the United States get sick from contaminated food. And CDC (Table 9.1 row 5) estimates that 1 in 6 Americans gets sick from contaminated foods or beverages each year, and 3,000 die from foodborne diseases. The U.S. Department of Agriculture (USDA) estimates that foodborne illnesses cost more than $15.6 billion each year. Therefore, quantifying the present information and forecasting future events is a vital factor for society. Food poisoning data is collected from Twitter using keywords listed in Foodsafety.gov website. Further, the analysis is done on food poisoning data, and then statistics and evaluation

are done based on food poisoning data.

Next, the analysis is done on Immigration data based on immigrant family separation policy, according to Homeland Security figures, about 2,000 children have been separated from their parents. The data is collected from Twitter using keywords like immigration, illegal, child, separation, and border. The analysis is done on immigration topics and statistics, and evaluation is done based on immigration data.

CHAPTER II

REVIEW OF LITERATURE

2.1 Related Work

This research is based on different works proposed in literature. In this chapter, we review previously published works related to and contributing to this research. These works can be classified as topic modeling, information measure, information diffusion and sentiment analysis. Research related to each of the above categories are summarized in the following subsections.

2.1.1 Topic modeling

Topic modeling refers to a generative model for analyzing large quantities of unlabelled data. At the core of topic modeling is the assumption that text documents contain several topics. Documents are viewed as bags of words. The goal of topic modeling is to detect the hidden topics in documents. A topic is viewed as a probability distribution over the collection of words, and the topic model is the statistical relationship between a group of observed and unknown random variables that specifies a probabilistic procedure to generate the topics Reed (2012). One of the popular topic modeling technique is Latent Dirichlet Allocation Blei et al. (2003). Latent Dirichlet allocation (LDA) is a generative hierarchical probabilistic model that extracts the latent topics and their corresponding weights in the documents. The generative works by grouping similar keywords under a topic based on co-occurrence of words with the topic in the document. The general scheme of LDA process is given below which

generates a set of topics given a collection of documents D.

For each document w in a corpus D:

1. Choose N Poisson (ξ)

2. Choose θ Dir (α) , Dir (α) is a draw from a uniform Dirichlet distribution with scaling parameter α

3. for each of the N words w_n:

(a) Choose a topic z_n Multinomial θ.

(b) Choose a word w_n from p (w_n| z_n,β), a multinomial probability conditioned on the topic z_n.

The output from an LDA algorithm is a set of specified topics and their weights in each document. Each topic is a collection of words and associated weights.

Non-negative Matrix Factorization (NMF) is another approach to topic modeling Shi et al. (2018) This approach provides a matrix based algorithm to define topics where as LDA is Bayesian approach. It is a matrix factorization method in which a document corpus is represented as a matrix called term document matrix (TDM). If there are n documents and m words in the corpus, TDM is an m-by-n matrix. Assume that an m-by-n matrix A represents a TDM. Then entries of A are nonnegative. Several approaches are found in the literature to construct a TDM. One simple method to compute entries a_{ij} of a TDM A is count of a word i in document j. Another popular method is tf-idf defined as $a_{ij} = tf_{ij}\log\frac{N}{df_i}$ es the total documents, tf_{ij} denotes the number of words i in document j, and df_i denotes the number of documents containing the word i. The NMF method of topic modeling factors a TDM, A into two non-negative matrices W and H such that A $\approx WH^T$. Then W represents the word topic matrix and H represents document topic matrix. One method of factoring is to minimize the Frobenious norm of the matrix A-WH^T . That is minimize $\frac{1}{2}\|A-WH^T\|_F^2$.

There are many other models Rabiner (1989), Kalman (1960), Mau et al. (1999),

McLachlan and Peel (2000) and techniques related to topic models like Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Correlated Topic Model (CTM). LSA is a statistical technique which deals with extracting and representing the relations between words in a large corpus. The method of LSA helps in information retrieval from a large text Landauer et al. (1998). PLSA which evolved from LSA is a probabilistic generative model which associates unobserved variables with each occurrence of a word in a document Choi (2011). This co-occurrence of words in the document has applications in information retrieval and filtering, machine learning from text and natural language processing Landauer et al. (1998). Correlated Topic Model addresses one of the Limitation in LDA topic modeling technique. LDA is unable to model topic correlation between the generated topics from the model. Correlated topic modeling (CTM) developed by Blei and Lafferty (2007) has the capability to capture correlation between topic proportions and thus addressing a limitation of LDA. The CTM models the words of each document from a mixture model. The mixture components are shared by all documents in the collection; the mixture proportions are document specific random variables. The CTM allows each document to exhibit multiple topics with different proportions. It can thus capture the heterogeneity in grouped data that exhibit multiple latent patterns Blei and Lafferty (2007).

2.1.2 Information Measure

Claude Shannon developed information theory to study communication systems. Losee (1997) states that the origin of information theory is generally attributed to Harry Nyquist Nyquist (1924). Shannons work, The Mathematical Theory of Communication provided the currently popular measure of information known as Shannons entropy Shannon (1948).Shannons theory deals with information to be conveyed with

three communication problems: first, the accuracy of the information to be transmitted; second, how precisely the meaning is transferred and third, from all the information transferred how much is selected from the set of messages. The last aspect is the effectiveness of the information transmitted from the sender to the receiver. The information in this context deals with a message. There should be a function to choose a message from the set of possible messages. This selection process can be done with the help of logarithmic function because if the set of messages increases from 4 to 16, the logarithmic measure increases from 2 to 4 bits of information.

Shannons entropy is, therefore, the information required to describe an event or entity. Following is the entropy equation:

$H = -\Sigma p_i n_i = 1 \log p_i$.

Where p_i is the probabilities of events and n is the number of different outcomes.

2.1.3 Information Diffusion

There is a large volume of literature on information diffusion. In this section, we review several papers in this topic.

Cazabet Remy, Nargis Pervin, Fujio Toriumi, and Hideaki Takeda, Remy et al. (2013) in their paper titled Information Diffusion on Twitter: everyone has its chance, but all chances are not equal present a method to quantify propagation of information in Twitter. In their approach, the number of followers of users plays an important role as the followers have the capacity to propagate information. Authors observed that the relation between the number of followers and the retweet chain length follows the power law. From the sequence of unique tweets posted by users in the network, the relationship between retweet chain length and follower count are calculated by giving the retweet chain length as the parameter to power law p(x) $\propto X^{-\alpha}$ where x $\leq x_{min}$,

it estimates the power law parameters α and x_{min}. They concluded that tweets are propagated more widely when there are more followers. By giving the user followers count as input to the model, it randomly generates the retweet chain length which is compared to the actual retweet length. And they observed that the retweet chain length gives realistic results by the power law.

Eleni Stai, Eirini Milaiou, Vasileios Karyotis, and Symeon Papavassiliou, in the paper titled Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation Stai et al. (2018) study temporal dynamics of topic-specific information spread in Twitter. They assumed that each topic corresponds to a hashtag, where the hashtags originate from the following:

1) From tweets of users they follow or

2) Learning about the topic from sources outside twitter, and publish the topic with a hashtag in twitter.

Hashtags are divided into three categories with respect to their temporal patterns. Tweets with a particular hashtag over time are grouped as single-spike, multi-spike, and fluctuation patterns. The single spike has a single time interval with a widespread appearance of hashtags in tweets (spike). Multi-spike has multiple single-spikes among time intervals with infrequent appearances and the fluctuation is characterized by a moderate frequency of spikes over a long time interval. To validate information spread in Twitter for several hashtags chosen to cover a variety of characteristics an epidemic model is used. The susceptible-infected (SI) is an epidemic model which does not underestimate the range of spread of topic-specific information propagated in Twitter. The authors concluded that constant infection rates are mostly suitable for hashtags of fluctuation type and time-varying ones for single-spike hashtags. The equations below are used to calculate the change in susceptible, infected users at time t.

$$\frac{dS(t)}{dt} = -I(t)\frac{S(t)}{N(t)}\ K(t)^{out}_{avg}\lambda_1(t) - S(t)\lambda_2(t) - S(t)\lambda_2(t)\frac{S(t)}{N(t)}\ K(t)^{out}_{avg}$$

$$\frac{dI(t)}{dt} = -(\frac{dS(t)}{dt})$$

(dS(t)/dt) and (dI(t)/dt) stand for the continuous change (per unit of time) of the number of susceptible (have not been informed) and infected (have been informed) users, S(t) and I (t) stand for the number of susceptible, infected users at time t, respectively. N(t) is the total number of Twitter users at time t, i.e., N(t) = I (t)+S(t).$K(t)^{out}_{avg}$ is the average out-degree of users (i.e., number of followers) in Twitter at time t and $\lambda_1(t)$ and λ_2(t) denote the probability rate that an infected or susceptible users respectively and will publish a tweet with the particular hashtag of interest.

Hengmin Zhu, Yuehan Kong, Jing Wei, Jing Ma Zhu et al. (2018) proposed a model which incorporates opinion evolution into the process of topic propagation simulated to explore the impact of different opinion distributions and intervention with an opposite opinion on information diffusion. The model (epidemic SEIR) is applied on four propagation states, i.e., susceptible (an agent has never received any information about a topic), exposed (they receive the topic, but have not published their opinions in the network), infectious (received the topic and spreads) and recovered state (received it but is no longer interested in spreading it).

Opinions evolve based on Bounded Confidence model:

$$O_j^{t+1} = \begin{cases} O_j^t + (1 - conf_i) * inf_{ij} * (O_i^t - O_j^t), when \mid O_i^t - O_j^t \mid \leq \epsilon \\ \\ \mathrm{O}_j^t \text{ , when } \mid O_i^t - O_j^t \mid \leq \epsilon \end{cases}$$

where $O_j^t(O_i^t)$ is the opinion of agent j(i) at the time t, and $conf_j$ is the confidence of agent j which is set randomly at the beginning, and inf_{ij} is the influence of agent

i on j which can be calculated from network structure. spread_prob ρ_j measures the probability of an agent spreading a topic out of his specific intention.

$$\rho_j = \begin{cases} |\mathrm{O}_i\text{- } O_j \mid \text{if agent j takes the intention of debating.} \\ 1- |\mathrm{O}_i\text{- } O_j \mid \text{if agent j takes the intention of approving} \end{cases}$$

Function Prop (O_j) is defined to calculate the proportion of a single opinion. Prop $(O_j) = \dfrac{\text{number of class } (O_j)}{\text{number of the total agents}} where, class(\mathrm{O}_j)$ represents the class that O_j belongs to. Let agent j receive a topic from agent i, the probability of agent j spreading it, F_{ij}, can be regarded as the harmonic mean of Prop (O_j) and ρ_j, so it is given by the formula $F_{ij}= \dfrac{2^*\mathrm{Prop}(O_j)\times\rho_j}{\mathrm{Prop}(O_j)+\rho_j}$

They concluded that agents opinion distribution and intervention with opposite opinion can influence information diffusion to a certain extent and the topic with one-sided opinions can be reposted by more agents, hence spreads faster and more widely.

Bao-Thien Hoang and Kamel Chelghoum and Imed Kacem proposed a learning based model for predicting information diffusion in social networks Hoang et al. (2016). Information diffusion prediction analyses all factors affecting users diffusion decision such as user features, user-user interaction, crowd features and the presence of multi topics in the content item. They used a machine learning method (gradient descent) for identifying the weighting parameters of each factor. The output of this algorithm is a solution to the optimization problem.

De Wang, Aibek Musaev and Calton Pu Wang et al. (2016) present a social interaction based model FAST by taking four significant properties of social interactions into account including familiarity, activeness, similarity, and trustworthiness. The model is applied to diffusion analysis of rumor dynamics. A new metric called FD-PCI (Fractional and Directed Power Community Index) based on PCI index is

proposed to identify influential spreaders on the weighted and directed social graph. Taking k-core index, PCI, and PageRank Bickle (2010)Page et al. (1999) as baselines, FD-PCI results shows a high correlation and monotonic relationship with users information spreading capability. They inferred that k-core index and PCI are not suitable for weighing the user's information on the social graph model. PageRank has low performance in terms of correlation with users information spreading capability.

The mathematical model for FAST: $W_{ij} = F_{ij} + A_i + S_{ij} + T_i$, W_{ij} is the weight of the link from user i to user j. F_{ij} is the value of familiarity for the link from user i to user j. A_i is the value of activeness for user i. S_{ij} is the value of similarity between user i and user j. T_i is the value of trustworthiness for user i. Where F_{ij} is calculated as n_c/n_t (n_c and n_t represent the number of contacts between user i and user j through the link from i to j and number of total contacts from the user i respectively) $A_i = t_d/t_p$ where, t_d and t_p denote number of days and number of days in a period of time.

Ashwin Kumar T.K and George K.M, present a new model for microblog data analysis based on an asset price bubble model TK and George (2016). The research undertaken in this thesis is closely related to their work. A summary of their paper is described below: Since the historic data for a given topic may not be available in twitter. Therefore the conventional approaches might not be effective. So they proposed a decision methodology which is unconventional combining information diffusion and asset price bubble model associated to topic definition.
The proposed model consists of three components a topic definition, potential time-series, and B function.
Topic Definition: A variation of the AFINN approach with user input have been used. A topic Z is defined as a triple Z=(L,R,δ) where L is a set of strings, R is a set of asymmetric relations with values true or false between elements of L, and δ is a map-

ping that associates a real number with every element of L.

The tweet contribution is interpreted according to the topic. The effect of R here is to indicate the presence of a word affects the weight of a keyword and to set the correct context.

Tweet Potential: Potential of a topic is defined as a function of time t. The contribution of a tweet at time t to the potential is defined by an influence function $\varphi(l)$. The influence function should capture the contribution of the tweet to the topic being analyzed. The potential topic definition is essential as it produces the time series for analysis. The term tw represents a tweet and $\varphi(0) =$ is defined to be 1. Intuitively, φ(l) is the influence of a retweet of level l. The original tweet is at level 0, and so its influence is defined as 1. The formula for potential is

P_z(t)= $\sum_{\text{tw at time t}} P(tw) * (\varphi(l) \mid$ l level of tw)

B Function: It compares the time series data to a pre-selected model. The model is a pair (δ, μ) where δ is a function defined in [0, T] and μ is a measure defined for functions in [0, T] as the threshold. Decisions are made based on the values of μ for the model and the time-series under consideration.

2.1.4 Sentiment Analysis

Sentiment quantification is a part of sentiment analysis, a set of tasks concerned with the analysing of texts according to the sentiments/ opinions / emotions /judgments expressed in them. Below are few papers describing various approaches to sentiment analysis and quantification.

Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband present a machine-learning approach for sentiment analysis of Twitter Hasan et al. (2018). Opinion mining and sentiment analysis aims to explore opinions or text on different platforms of social media by calculating sentiment, subjectivity analysis or polarity.

In the lexicon-based method polarity is calculated from the dictionary that consists of a semantic score of a particular word. For sentiment analysis, a semantic orientation of words, phrases, and sentences are computed in a document. The research is focused on providing a comparison between sentiment lexicons (W-WSD, SentiWordNet, Text Blob) Navigli (2009) Esuli and Sebastiani (2007) Loria et al. (2014) so that the best can be adopted for sentiment analysis. Validating three of the sentiment analysis lexicons with two machine-learning algorithms (Nave Bayes and SVM). They concluded that the results of TextBlob were relatively better; they obtained the best result when analyzing tweets with W-WSD.

Wei Gao and Fabrizio Sebastiani Gao and Sebastiani (2015) proposed an approach to quantify information using machine learning algorithms and predicted prevalence (percentage of items in set S that belong to class c). By using CMU Twitter NLP Kiritchenko et al. (2014)(Section 5.2.1) tweets are represented in vector notation which consists of number of all-caps tokens, the number of tokens for each POS tag, the number of hashtags, the number of negated contexts, the number of sequences of exclamation and/or question marks, and the number of elongated words. Sentiment lexicons are used to calculate sentiment of the tweets (Positive, Negative, or Neutral). SVM (KLD) and SVM-perf were used to predict the prevalence. Three evaluation measures are used to estimate the quantification.

Absolute Error: is defined as the average absolute difference between the predicted class prevalence and the true class prevalence.

AE(p̂,p)= $\frac{1}{\mid C \mid} \sum_{c_j \in C} \mid \hat{p}(c_j)\text{-p}(c_j)\mid$ $Where p(c_j)$ is true class prevalence and $\hat{p}(c_j)$ is predicted class prevalence and C is set of available classes

Relative absolute error: is defined as

RAE(p̂,p)= $\frac{1}{\mid C \mid} \sum_{c_j \in C} \frac{\mid \hat{p}(c_j) - p(c_j) \mid}{p(c_j)}$ $Where p(c_j)$ is true class prevalence and $\hat{p}(c_j)$ is predicted class prevalence and C is set of available classes

And the third measure is Kullback- Leibler Divergence a measure of the inefficiency incurred when estimating a true distribution over a set of classes utilizing a predicted distribution.

KLD($\hat{\text{p}}$,p)= $\sum_{c_j \in C} p(c_j) \log \frac{p(c_j)}{\hat{p}(c_j)}$ $Wherep(\text{c}_j)$ is true class prevalence and $\hat{\text{p}}(c_j)$ is predicted class prevalence and C is set of available classes

The results indicated that SVM (KLD) excels when compared to SVM-perf.

Adebayo Adetunmbi, Oluwafemi A. Sarumi, Oluwayemisi Olutomilola, and Olutayo Boyinbode Adetunmbi et al. (2018) analyzed opinion mining of movie reviews that help users to determine which movie to purchase or watch quickly and it helps the movie producers to get the feedback from customer on their films. Cornel Movies review dataset (Table 9.1 row 4) was used in the experiment (Cornel movie dataset). After pre-processing the dataset. Term frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) are extracted and represented in vector notation. Three Machine learning techniques (K Nearest Neighbours, Support Vector Machines and Naive Bayes) were used to classify reviews based on sentiment classification of weighted classes as either positive, negative or neutral. KNN had 95.9% accuracy, NB and SVM had an efficiency of 90.6% and 92.22% respectively. The result shows that KNN gives higher accuracy than SVM and NB.

Arash Mazidi and Elham Damghanijazi proposed a sentiment analysis approach Arash and Elham (2017) using extracted Ngram feature vector and POS (Part of Speech) from the text. They find a proper combination of feature vectors so that texts can be classified into positive or negative opinions. Information gain is used to select the features and then the machine learning algorithms Boolean Multinomial Nave Bayes (BMNB) and SVM Blitzer et al. (2007) are used to find the effect of different features on sentiment analysis. Recall, precision, and F-measure are used to evaluate the classification efficiency of sentiment analysis. The accuracy of POSWord features

is higher than Ngram features indicating better information to resolve the ambiguity thereby improving classification accuracy for both SVM and BMNB. The results indicate that the accuracy of BNMB is higher than SVM.

Mondher Bouazizi and Tomoaki Ohtsuki Bouazizi and Ohtsuki (2016a) present a pattern-based approach for sentiment quantification in Twitter. Their approach detect sentiments in a tweet, and propose a way to extract different existing sentiments using a set of pattern-based features and special Unigram-based features along with other essential features, then quantifying the sentiment within tweets. The initial step is to classify the data into weighted classes positive, negative or neutral. In the next step the following features have been extracted from different approaches.

1. Sentiment-based features are ones based on the sentiment polarity (i.e., positive/negative). These features are extracted using Senti-Strength Fellbaum (2010)

2. Punctuation and syntax-based features: In addition to sentiment-based features the features such as Number of exclamation marks, Number of question marks, Number of dots, Number of all-capital words and Number of quotes were also added.

3. Unigram-based features: WordNet Bouazizi and Ohtsuki (2016b) is used to collect unigrams related to each sentiment classes (positive, negative or neutral)

4. Pattern-based features: In this approach, the words are divided into three sets (emotional, content and grammatical) replaced by another expression based on the category. The classification is done based on the POS tag of the word in the tweet.

$$\text{res(p,t)}= \begin{cases} 1, \text{if the tweet vector contains the pattern as it is,in the same order} \\ \alpha.\frac{n}{N}, \text{if n words out of N words of the pattern appear in the tweet} \\ \qquad \text{in the correct order} \\ 0, \text{if no word of the pattern appears in the tweet.} \end{cases}$$

Out of the 4 sets, pattern-based and Unigram-based features achieved better performance.

Following scores are used to quantify information.

Unigram-based score (S_u) :N_i unigrams of a sentiment class i appear in the tweet t, The Unigram-based score of the tweet for the given class i is defined as follows $S_u(i) = \sum_{k=1}^{N_i} S_k$

Pattern-based score (S_p): knn patterns of length j of a sentiment i that resembles the most to the most to the tweets patterns, and given the weights β_j given to the patterns of length j.

$S_p(i) = \sum_{j=1}^{N_L} \beta_j . \sum_{k=1}^{knn} res(p_k, t)$, S(i)= $\xi . S_u(i) + (1 - \xi) * S_p(i)$ where ξ is a weight such as $0 \leq \xi \leq 1$

For each tweet judged as sentimental, the (positive/negative) score returned is selected as the quantification of information. For each threshold 0, 1, ... , 20, then measure the precision of classification of the tweets that have a score higher than the threshold, and the number of positive/negative tweets having such score over the total number of positive/negative tweets (i.e., coverage).

2.2 Problem Statement

As outlined in the previous section, sentiment is used by researchers to quantify information for weight assignment. However, sentiment values of words will be positive, negative, or neutral and might not be the best way to compute weights for all applications. As an example, consider the case of tweets related to flu that often may contain the word tired. The sentiment value for tired may be negative, but it is appositive for flu indication. So, keyword and weight value determination is an essential area of research for different applications when quantifying information.

This thesis is to propose a different approach for quantifying and computing information in tweets based on the principles all publicity is good and information is not negative. Due to the size, speed and variety of these tweets, they fit the characterization of big data and hence, big data related environments and tools are used in data collection and analysis. Methods associated with topic modelling have been used to determine word weights. A time-series model is built based on the potential used for further analysis, and its quantification measures are compared against the previously published models.

CHAPTER III

METHODOLOGY

3.1 Tools Used

This section includes the tools used for data collection.

3.1.1 Apache Hadoop

Apache Hadoop (Table 9.1 row 1)is an open source software for reliable, scalable and distributed computing. The software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The Hadoop framework is composed of Hadoop Common, Hadoop Distributed File Systems (HDFS), Hadoop YARN and Hadoop MapReduce. Hadoop Common contains a set of libraries and utilities needed by other Hadoop modules. HDFS is a distributed file system that stores data on commodity machines provide very high aggregate bandwidth across the cluster. Hadoop YARN manages computing resources in the cluster and uses them for scheduling user's application. Hadoop MapReduce is a programming model for large-scale data processing. It is suitable for applications having large datasets and provide high throughputs access to data.

3.1.2 Apache Flume

Apache Flume (Table 9.1 row 2) is a distributed, reliable, robust and available system for efficiently collecting, large amounts of data from many different sources to a centralized data store. It has a simple and flexible architecture based on data streaming flows.

Following is the Twitter Data Streaming process: To stream data from external sources, Flume integral components such as agent, sink, source, channel, and event have been used.

- An event is a unit of data that is transferred using flume.
- The external source (i.e. Twitter) sends events to Flume in a format that is recognized by the target Flume source.
- Flume source stores events into one or more channels after receiving. The channel is a passive store that keeps the event until it's consumed by a Flume sink.
- The sink removes the event from the channel and puts it into an external repository like HDFS.
- An agent is a container for data flow.

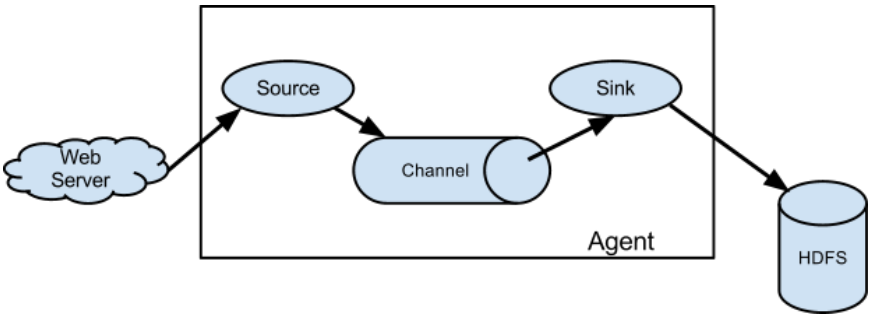

Figure 3.1: Flume Agent

3.2 Data Collection

Apache Flume (Table 9.1 row 2) is used to retrieve data from Twitter. For streaming the data, we have created a flume agent and twitter application. The twitter application contains a set of keywords related to the domain. From the application, API keys are used for streaming data from Twitter into the Hadoop cluster. For flume agent, a configuration file is created which contains tokens of the twitter application. The data obtained from twitter is in JSON format.

Two different domains of data are collected. First, we collected 71.8 GB of Food poisoning Data. The data collection period is 01/31/2018 to12/31/2018. The tweets are collected using a set of tokens Diarrhea, Abdominal Pain, Vomiting, Puke, and Fever.

Second, we have collected 219 GB of Immigration Data. The collection period is 08/01/2018 to 02/28/2019. The tweets are collected using the tokens immigration, separation, crime, illegal, and boarder.

3.3 Data Pre-processing

From the JSON format file, tweet text, user name, created date, owner name, owner time stamp and user fields are retrieved for further processing. The tweets text is used then cleaned by removing URLs, user mentions, emoticons and stop words.

CHAPTER IV

MODEL

The tweets are composed of tokens (α) which can be key words or key phrases. We define a Tweet_Set (TS) as a collection of tweets. We use the term potential (P) to refer to the information content of tweets and Tweet_Sets. Following the idea of Shannon entropy (Shannon, C.E., 1948) the potential of a tweet P(tw) is defined as the average of the information of the tokens present in the tweet. The potential of a Tweet_Set is defined as the average of the potentials of the tweets in the Tweet_Set. Formal definitions follow:

Assume that P denotes potential. Then,

P(TS)= $(\sum_{tw \in TS} P(tw) * \varphi(l))/N$, where N = $\mid TS \mid$ the size of TS,

and φ(l) a function, l is a parameter .....(1)

Intuitively speaking, φ(l) is a tweet potential modifier for retweet.We assume l to be the retweet level of the tweet tw and φ(l)= ρ^l

P(tw)= $\sum_{\alpha \in tw} p_\alpha * I_\alpha$, ,where p_α denotes the proportional weight and I_α the information content of the token α .....(2)

Assuming all tokens having the same weight. We try different methods to define I_α the information content of a token.

A. Token identification

We present two methods to identify tokens from a Tweet_Set. The first method makes use of topic modeling and selects the top words from the topics. The second method begins with a few seed words and build more words using similarity measures of words which is an iterative algorithm. For topic modeling, we adopt LDA algorithm. The

idea behind the iterative method is to start with some key words (called anchor words) and add more words from the tweets as determined by a defined measure. (The use of anchor word term is different from the use in NMF). We define similarity in abstract form as a relation between words and denote as $\delta(w_1, w_2)$. We also assume that there is a set of anchor words that we know with probability 1 are in the keyword set. The proposed iterative algorithm is described below:

ALGORITHM I: Token construction

Let S represent the set of words corresponding to tokens.

Let K be the set of all significant words taken from the tweets of the Tweet_Set

Let A be a set of anchor words.

Step 1: Set S = A; K = K-A;

Step 2. For each w_1 in S and each w_2 in K do

Step 3: If $\delta(w_1, w_2) >$ threshold add w_2 to S if it is not already in S and remove w_2 from K

Step 4: If any word is added to S, go to Step2

Step 5: Output S as the token set.

As K is finite, the procedure will terminate with a worst case performance of O($\mid K \mid^2$). The next algorithm specifies an approach to assign information measure I_α to the tokens. It is based on topic building algorithms. As one possible avenue, we make use of the NMF algorithm for our purpose. Given an m-by-n matrix M with nonnegative entries, the NMF algorithm computes a nonnegative factorization WF such that M $\approx$ WF such that W is m-by-k and F is k-by-n with nonnegative entries. The factorization is not unique.

ALGORITHM II: Information assignment to tokens

Step1: Construct a term document matrix (tdm) M with the words associated to the tokens as rows and tweet collection per time unit as document.

Step 2: Apply the NMF algorithm with k = 1 to get a vectors W and F, where M $\approx$ WF^T.

Step 3: Set W = W/$\|W\|_2, where\|.\|_2$ is the vector 2-norm.

Step 4: Output entries of W as the information of corresponding tokens.

B. Level computation

In order to apply the concepts to applications, we need to compute the retweet levels during each time interval. We have designed and implemented a map-reduce algorithm to compute the retweet levels. The mapper and reducer are described as flowcharts in Figure 4.1 and 4.2.

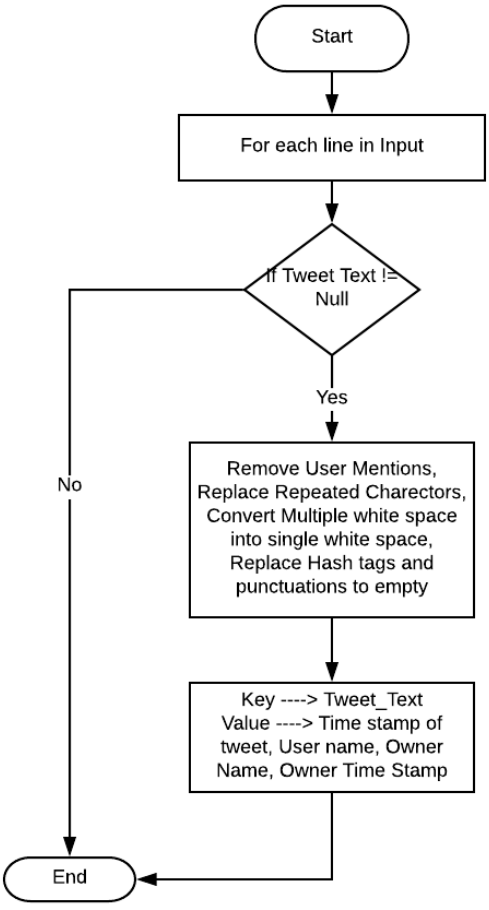

Figure 4.1: Mapper Flowchart

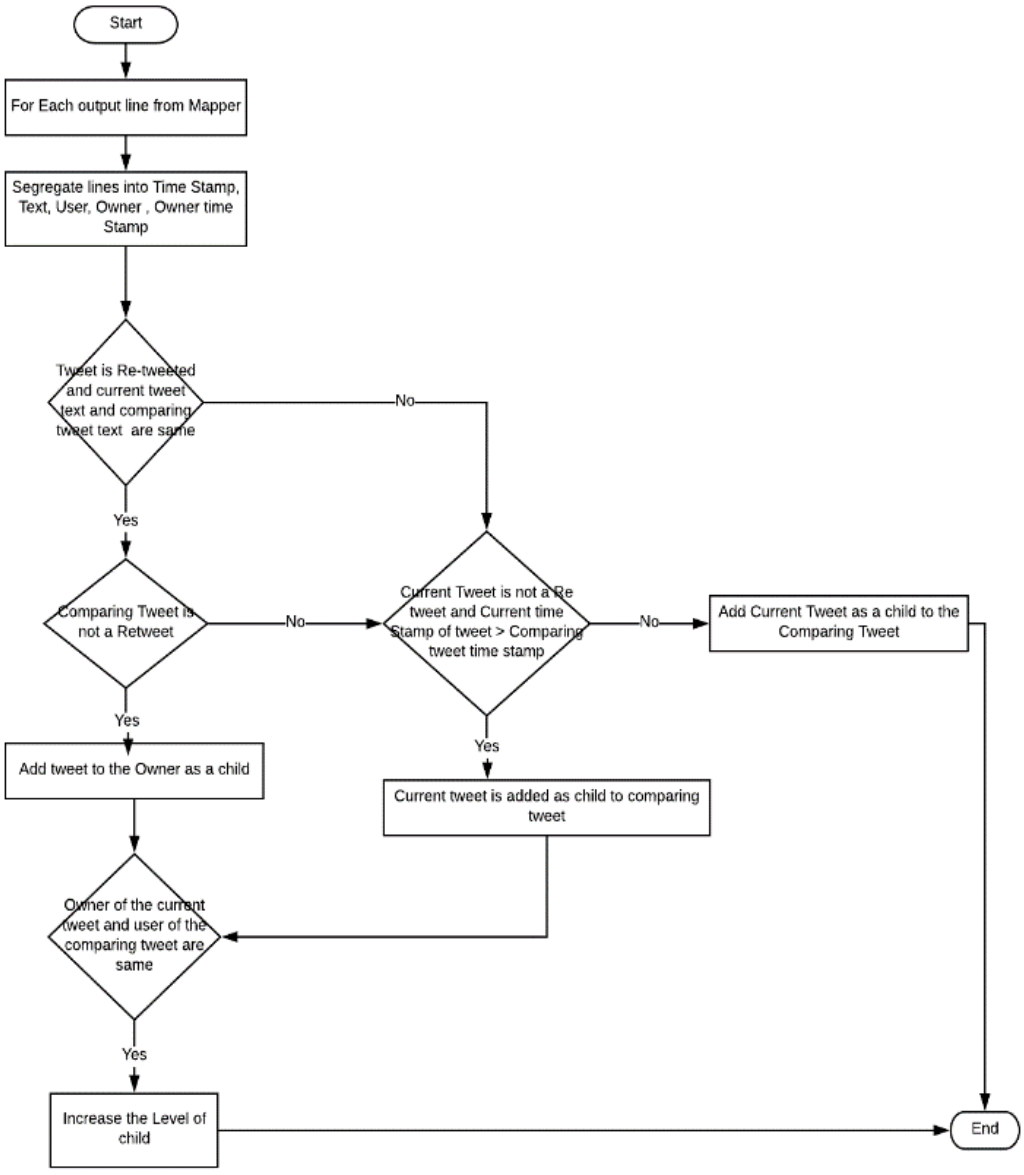

Figure 4.2: Reducer Flowchart

The information quantification method outlined in this section is applied to two sets of tweets (described in the data collection section) to demonstrate practical applicability of the model. The various results derived by the computations are given in the next section. The first step is to identify tokens and weights followed by level computation.

CHAPTER V

COMPUTED RESULTS

Topic Extraction

To determine tokens and their weights, we resort to topics. So, the first step is to identify or extract topic(s) from tweets. Topic Extraction deals with extracting information from documents, and it can be done using Topic Modelling. A topic is a set of keywords, and Topic Modelling refers to a statistical model for analyzing large quantities of unlabelled data. Latent Dirichlet allocation (LDA) is the most common technique of topic modelling. LDA is a generative probabilistic model which groups similar keywords under a topic based on co-occurrence of words with the topic in the document.

Another method for Topic Extraction is Matrix factorization, As mentioned previously, we adopt Non-negative Matrix Factorization (NMF), which is a Linear-algebraic model that factors high-dimensional vectors into a low-dimensionality representation. The underlying theme of NMF is to construct a matrix factorization which builds a term-topic matrix. By using this matrix, we can weigh the keywords in the potential model.

5.1 Application 1: Food Poisoning Data

Topic Extraction by LDA method:

We used Gensim package available in Python to execute the LDA algorithm. It returns a set of key words and frequencies. The results obtained when LDA algorithm is applied to the food poisoning data set are shown in Figure 5.1.

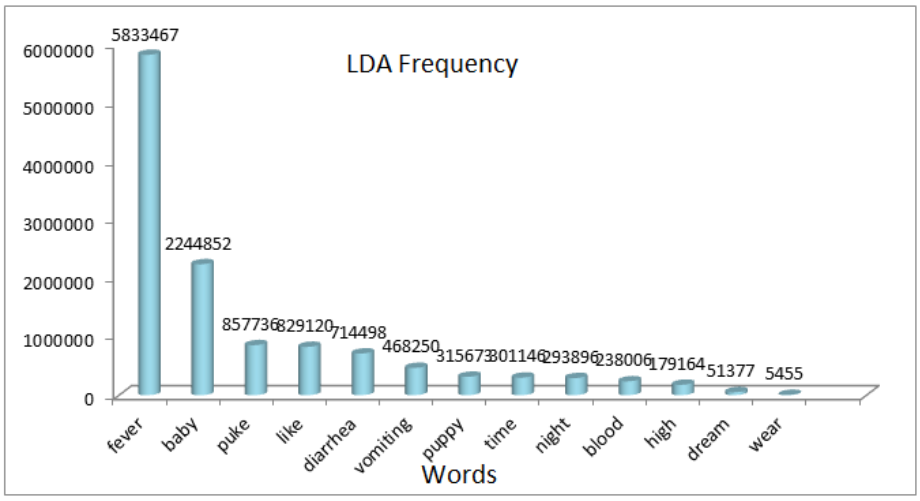

Figure 5.1: Token count by LDA method in Food Poisoning Data

Figure 5.1 shows the relationship between the keywords and their frequency in the data set. We can observe that keyword 'FEVER' is the most frequently occurring word with frequency 5833467.

Topic Extraction by NMF method:

The following are results of tokens constructed by the iterative approach of AlgorithmI.

Anchor word set used as initial seed words is {Diarrhea, Abdominal Pain, Vomiting, Puke, and Fever}. Tokens are selected based on trial and error method of the threshold value, initially when threshold is 0.90 the extracted words are given in Table 5.1.

fever	disgorgement
pyrexia	looseness
febricity	diarrhea
vomitings	Stinker
puke	Rotter
abdominal	Gits
lowlife	Skunk
stinkpots	Emesis
feverishness	Bums

Table 5.1: Tokens at Threshold value 0.90 in Food Poisoning Data

From the above threshold words, we can observe that there are words like feverishness

which has the same meaning of fever and bums that doesnt describe the topic. So the threshold is set to 0.95 and the resultant words are given in Table 5.2.

fever	disgorgement
pyrexia	looseness
febricity	diarrhea
vomitings	Stinker
puke	Rotter
abdominal	Gits
lowlife	Skunk
stinkpots	Emesis

Table 5.2: Tokens at Threshold value 0.95 in Food Poisoning Data

From the above-extracted tokens from their particular threshold, we can observe that the words are more related to the topic when the threshold value is 0.95. Considering the words at threshold value 0.95 as tokens, and by applying information assignment to tokens using Algorithm II (refer to model section), we obtain the proportional weights for tokens. The key words selected by Algorithm I and associated weights determined by Algorithm II are listed in Table 5.3.

Token	**Information Measure**
Fever	0.838697248
Pyrexia	0.0000145468
Febricity	0.000000300615
Vomitings	0.00000063655
Puke	0.092504587
Abdominal	0.005834771
Lowlife	0.0000232624
Stinkpots	0.00000011844
Disgorgement	0.000000785835
Looseness	0.00000255394
Diarrhea	0.060675229
Stinker	0.00000269339
Rotter	0.000000399
Gits	0.00000624881
Skunk	0.0000281303
Emesis	0.000018456

Table 5.3: Tokens and proportional weights for Food Poisoning Data

From Table 5.3, we can observe that token fever has the highest information measure value 0.83869742. Figure 5.2 shows the frequency of words listed in Table 5.3.

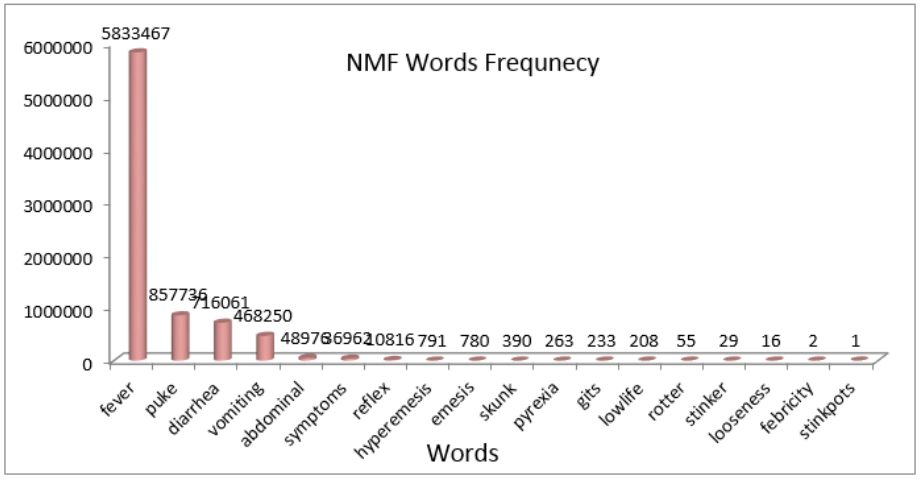

Figure 5.2: Token count by NMF method in Food Poisoning data

Figure 5.2 shows the relationship between tokens and the frequency of tokens in Food poisoning dataset, we can observe that the token fever had the highest frequency of 5833467 in Food poisoning dataset.

Level Computation Performance for tweets:

A map-reduce algorithm (refer to model section) computes retweet levels of the tweets. Since every tweet need to be compared with all other tweets in the dataset, the computational time for a large amount of data is relevantly longer. By using map-reduce parallel computation the execution time for 71.8 GB of data is lowered to 25 minutes. Figure 5.3 shows the frequency of tweets at different levels.

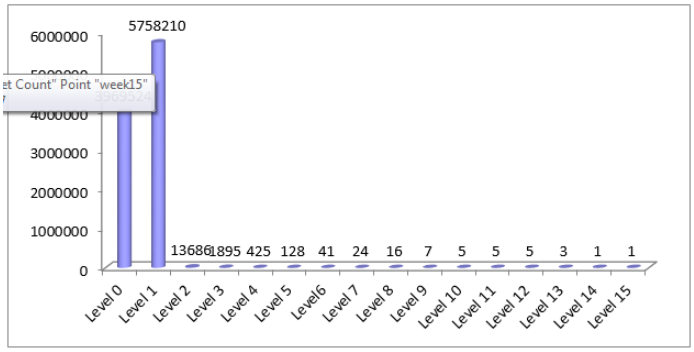

Figure 5.3: Retweet Levels in Food Poisoning Data

We can observe that the frequency of tweets is highest at level 1 and is getting diminished from level 2 and this retweet chain of tweets stops at level 15.

Computing Potential for the tweet set:

Tweet set potential is computed using equation 1 (refer to section model) with two different arbitrary constants 0.5 and 1.5. The potential of tweets for the two term-weighing approaches LDA and NMF are shown in Figure 5.4 and Figure 5.5.

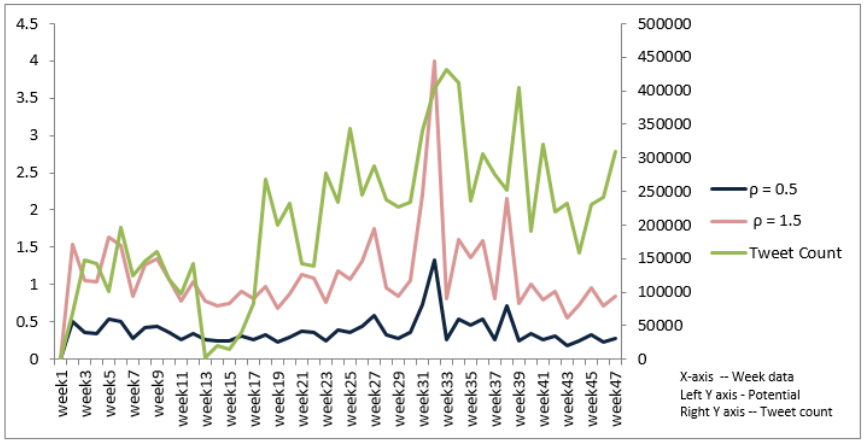

Figure 5.4: Tweet Count vs Potentials on LDA approach for Food Poisoning data

Figure 5.4 depicts the potential of Food poisoning dataset as time-series, where tokens are extracted by the LDA approach with two different rho (ρ) values. They are similar to tweet count for the period of analysis, but we can observe from the graph that from week13 to week15 even though the tweet count is less the amount of information gain (potential) is high, because the occurrence of the keywords(FEVER, BABY, PUKE and DIARRHEA)and weights extracted for the keywords by LDA method are higher.

Figure 5.5, illustrates the potentials using tokens obtained by NMF approach with two different rho (ρ) values which are similar to weekly tweet count from week 17 to week 47. The potential with rho value 1.5 from week 13 to week 16 is higher even though the tweet count is less thereby giving more information regardless of lower tweet count. Since, occurrence of the keywords(FEVER, VOMITING, PUKE and DIARRHEA)and weights extracted for the keywords by NMF method are high.

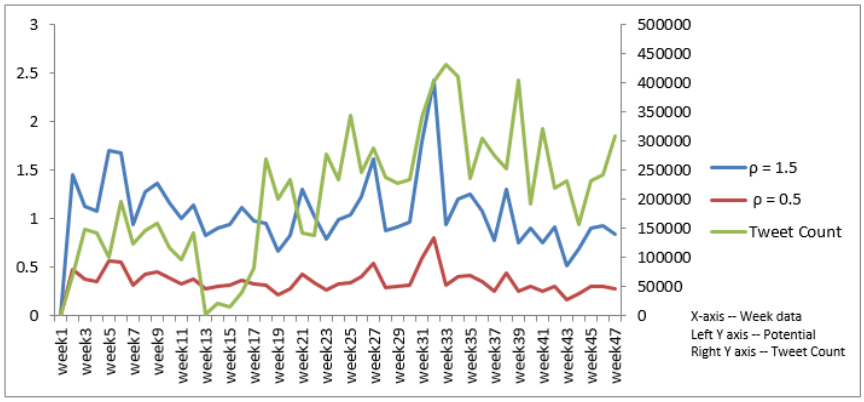

Figure 5.5: Tweet Count vs Potentials on NMF approach for Food Poisoning data

5.2 Application 2: Immigration Data

Topic extraction by LDA approach:

We used Gensim package available in Python to execute the LDA algorithm. It returns a set of key words and frequencies. The results obtained when LDA algorithm is applied to the immigration data set are shown in Figure 5.6.

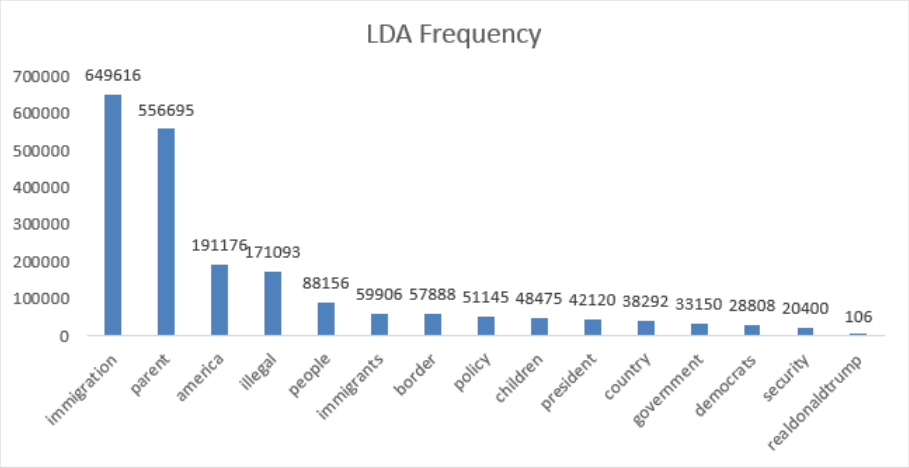

Figure 5.6: Token count by LDA method for Immigration Data

Figure 5.6 shows the relationship between the keywords and its frequency in the data set. We can observe that immigration is the most frequently occurring word with

frequency 649616.

Topic extraction by NMF approach:

The following are results of tokens constructed by the iterative approach of Algorithm I.

Anchor word set used as initial seed words is {"immigration", "separation", "crime", "illegal", "boarder", "parent" }. Tokens are selected based on trial and error method of the threshold value, when threshold is 0.90 and 0.95 the extracted words are given in Table 5.4.

Threshold 0.90	**Threshold 0.95**
Lessee	parent
Immigration	fraud
Lodgers	highjack
Tenant	immigration
Crime	burglary
Illegal	burglaries
Boarder	tenant
Crimesssss	isolation
Lodger	violation
Children	felony
Leaseholder	crime
Separation	disassociation
Renter	illegal
Mom	separate
Dad	migration
Papa	adopter
Hijack	disconnection
Violations	boarder
Progenitors	
Perpetration	
Felonies	

Table 5.4: Tokens obtained at different threshold values for Immigration Data

From the above-extracted tokens from their particular threshold, we can observe that the words are more related to the topic when the threshold value is 0.95. Considering the words at threshold value 0.95 as tokens, and by applying information assignment

to tokens using Algorithm II (refer to model section), we obtain the proportional weights for tokens. The key words selected by Algorithm I and associated weights determined by Algorithm II are listed in Table 5.5.

Tokens	Information Measure
parent	0.291820449
fraud	0.01870399
highjack	0.00000187172
immigration	0.511645885
burglary	0.00000651022
burglaries	0.00000172387
tenant	0.0000237514
isolation	0.0000488703
violation	0.000763516
felony	0.000438828
crime	0.006779052
disassociation	0.00000028187
illegal	0.159967581
seperate	0.002236284
migration	0.006299751
adopter	0.00000728928
disconnection	0.00000102312
boarder	0.000499426

Table 5.5: Tokens and proportional weights for Immigration data

From Table 5.5, we can observe that token immigration has the highest information measure value 0.511645885. Figure 5.7 shows the frequency of words listed in Table 5.5.

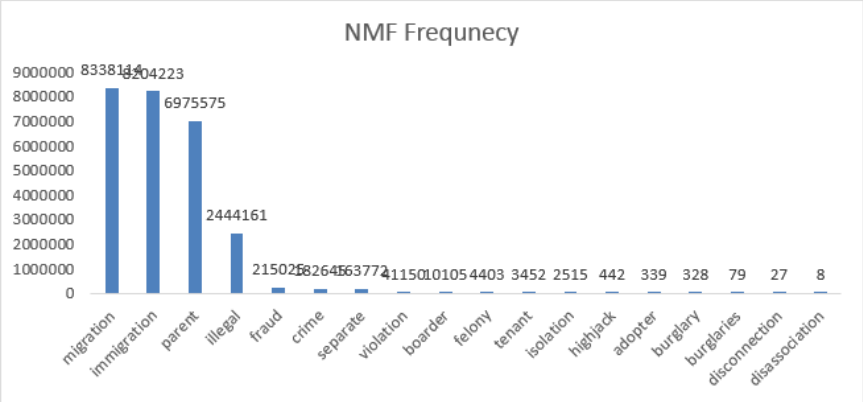

Figure 5.7: Token count by NMF method for Immigration Data

Figure 5.7 shows the relationship between tokens and the frequency of tokens in immigration dataset, we can observe that the token 'migration' had the highest frequency of 8338114 in immigration dataset.

Level Computation Performance for tweets:

Figure 5.8 shows the frequency of tweets at different levels.

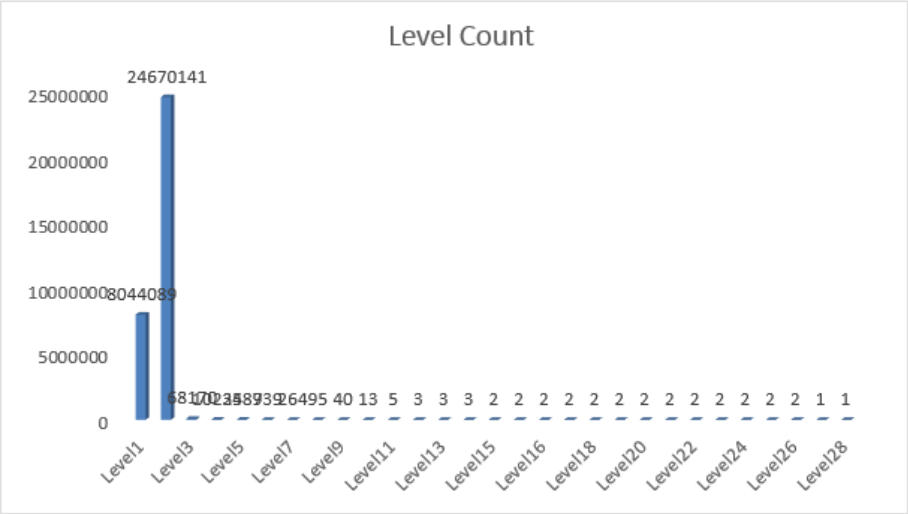

Figure 5.8: Retweet Levels for Immigration Data

We can observe that the frequency of tweets is highest at level 1 and is getting diminished from level 2 and this retweet chain of tweets stops at level 28.

Computing Potential of the tweet set:

Tweet set potential is computed using equation 1 (refer to section model) with two different arbitrary constants 0.5 and 1.5. The potential of tweets for the two term-weighing approaches LDA and NMF are shown in Figure 5.9 and Figure 5.10.

For Figure 5.9 and Figure 5.10, X-axis represents the weekly data, left Y-axis represents potential, and right Y-axis is the tweet count. The figures portray the potential of Immigration data at two different rho (ρ); we observe similar results of high potential in week4, week18, week22, week24, week27 and week 31 at lower tweet-count when the rho value is 1.5. Since, the occurrence of the keywords(Migration,

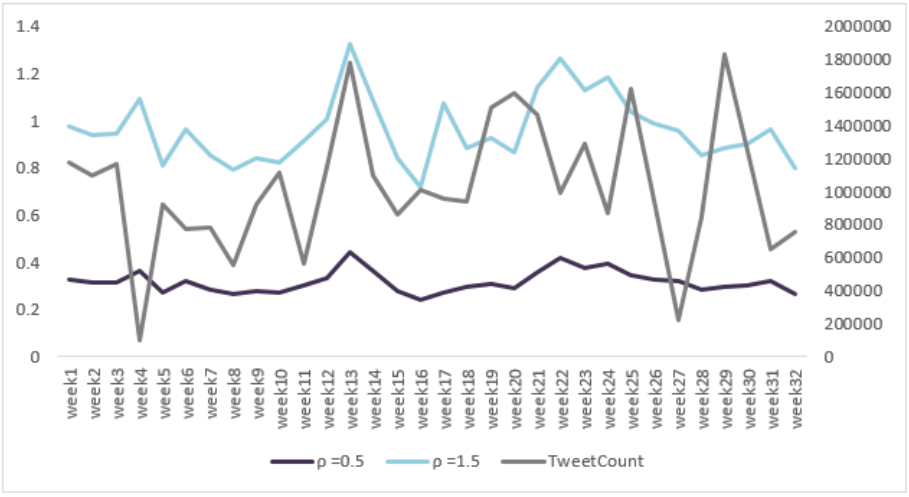

Figure 5.9: Tweet Count vs Potentials on LDA approach for Immigration data

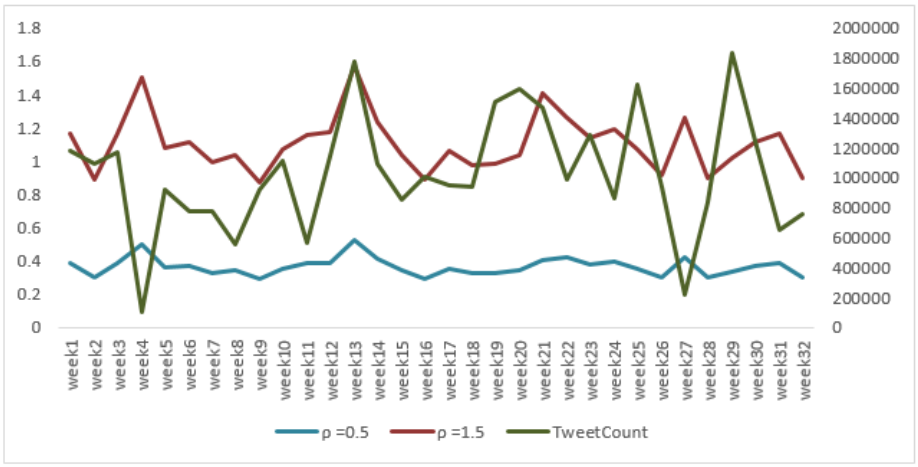

Figure 5.10: Tweet Count vs Potentials on NMF approach for ImmigrationData

Immigration, Fraud, Illegal, Parent and Border)and weights extracted for the keywords by topic extraction approaches are high.

The results of potential for Food poisoning and Immigration data are then compared with two other models from the literature. The models used for comparison are Sentiment quantification Gao and Sebastiani (2015) and Information Diffusion Remy et al. (2013).

CHAPTER VI

COMPARING MODELS

6.1 Model 1: Quantifying information by sentiment analysis

As explained in the related work Gao and Sebastiani (2015) to compute the sentiment of tweets AFINN (Table 9.1 row 3) database is used and a CMU tool tagger is used to calculate the POS and is represented in vector notation, and evaluation measures AE(Absolute Error) RAE (Relative Absolute Error) and KLD (Kullback - Leibler Divergence) are used to quantify the information by prevalence obtained from SVM (a supervised machine learning algorithm that analyze data used for classification and regression analysis).

Application 1: Food Poising Data

Computed prevalence for the analysis period is shown as time-series in Figure 6.1. The X-axis represents weeks and Y-axis represents prevalence. We can observe that the Negative prevalence is more in the Food poisoning data set, i.e., there are many negative sentiment tweets in the data, but for Food poisoning data it can be considered as a positive context.

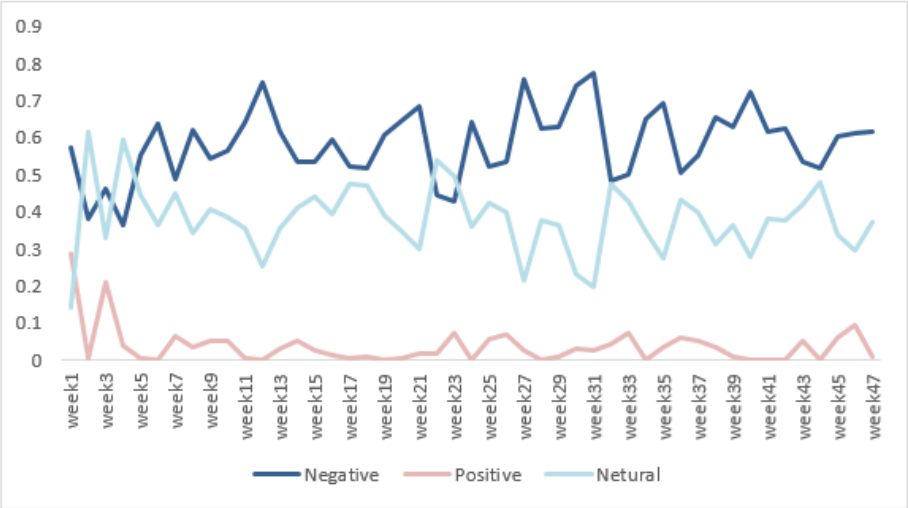

Figure 6.1: Prevalence in Food poisoning data

Application 2: Immigration Data

In Figure 6.2 X-axis represents the week data and Y-axis represents prevalence. Neutral prevalence is more in Immigration data.

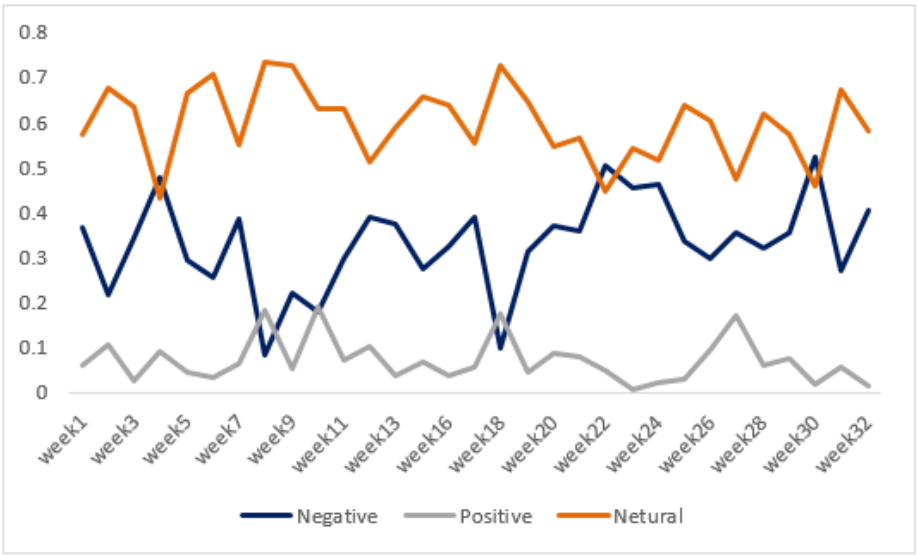

Figure 6.2: Prevalence in Immigration data

6.2 Model 2: Information Diffusion

Remy et al. (2013) quantify propagation of information in Twitter by the number of followers of users. Followers play an important role and have the capacity to propagate information. Based on retweet count the power law model will generate an alpha value that predicts the retweet count of the user when the follower count is given as input.

Application 1: Food Poising Data

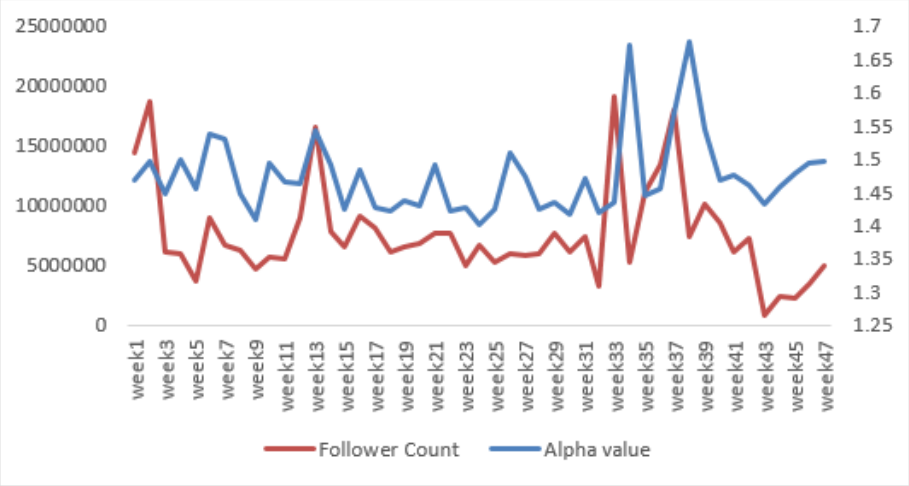

Figure 6.3: Follower count and Alpha value in Food Poisoning data

In Figure 6.3, X-axis represents weeks, left Y-axis represents Follower count, and right Y-axis represents the Alpha value, we can observe that the alpha value is proportional to the follower count which means if there are more number of followers more retweets can be expected.

Application 2: Immigration Data

In Figure 6.4, X-axis represents week data, left Y-axis represents Follower count, and Right Y-axis represents the Alpha value, we can observe that the alpha value is mostly proportional to the follower count.

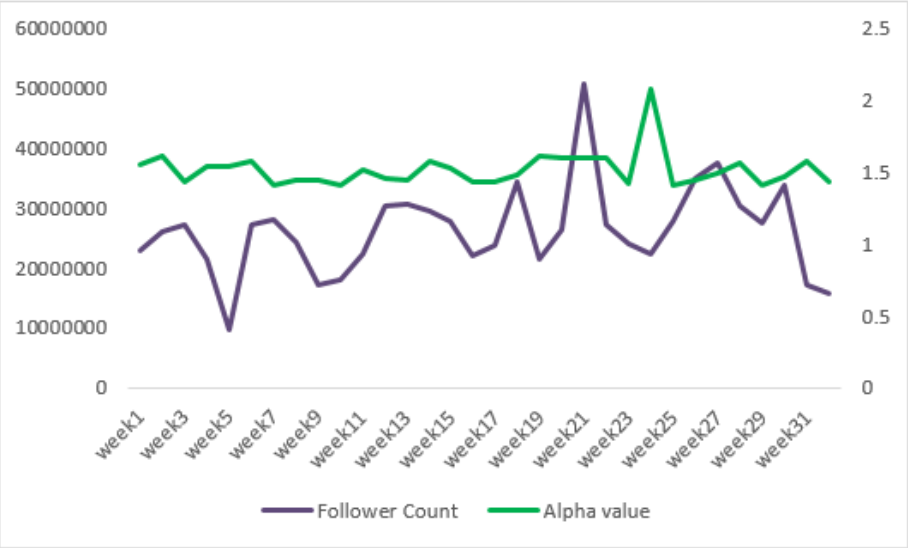

Figure 6.4: Follower count and Alpha value in Immigration data

6.3 Comparison Measures

Correlation and Normalization are used as evaluation measures to compare the models.

Correlation analysis is used to quantify the degree to which two variables are related. We can evaluate the correlation coefficient that tells us how much one variable changes when the other one does.

Normalization analysis is used commonly when the relationship between two dataset is non-linear. We transform data to reach a linear relationship.

6.3.1 Correlation Measure

Correlation between our proposed model and sentiment analysis model are given in tables 6.1 and 6.2.

Application 1: Food Poisoning Data

The table 6.1 lists correlation between the potential computed with different rho values and prevalence (positive, negative or neutral) for food poisoning data. Correlation

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Positive	-0.28981	-0.29125	-0.19943	-0.20018
Negative	0.077268	0.076085	0.058667	0.058138
Neutral	0.083024	0.085006	0.051633	0.052577

Table 6.1: Correlation between Potential and Prevalence for Food Poisoning data

between positive prevalence and potential of NMF and LDA approaches at two different rho values 0.5 and 1.5 are negatively correlated. As we can observe from the Figures 6.5 and 6.6 they are in opposite directions.

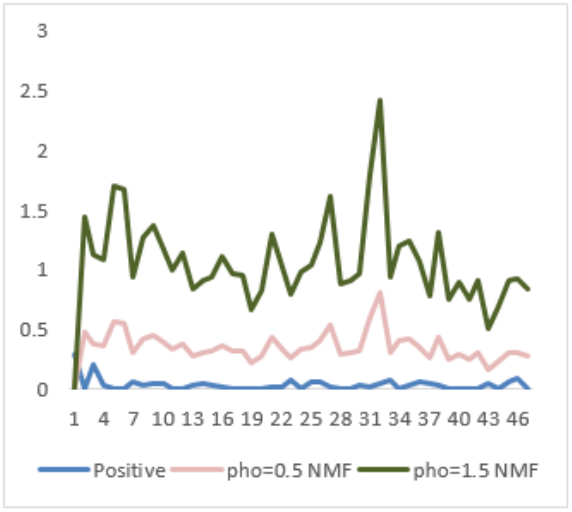

Figure 6.5: Correlation between Positive prevalence and NMF

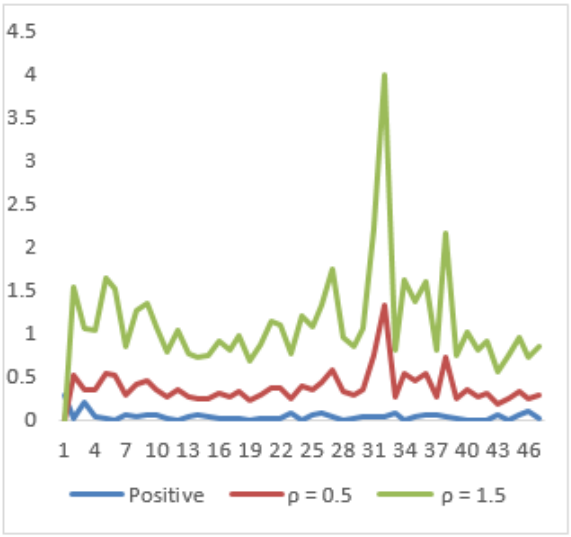

Figure 6.6: Correlation between Positive prevalence and LDA

Application 2: Immigration Data

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Positive	-0.29033	-0.28834	-0.29295	-0.27014
Negative	0.026724	0.057556	0.130459	0.051356
Neutral	-0.32989	-0.36637	-0.38129	-0.34003

Table 6.2: Correlation between Potential and Prevalence for Immigration data

Table 6.2 shows the correlation between prevalence and potentials of Immigration data computed by different algorithms and parameters. Positive and neutral sentiment prevalences are negatively correlated with potential. But negative sentiment

prevalence is positively correlated. Figures 6.7 and 6.8 display the time-series.

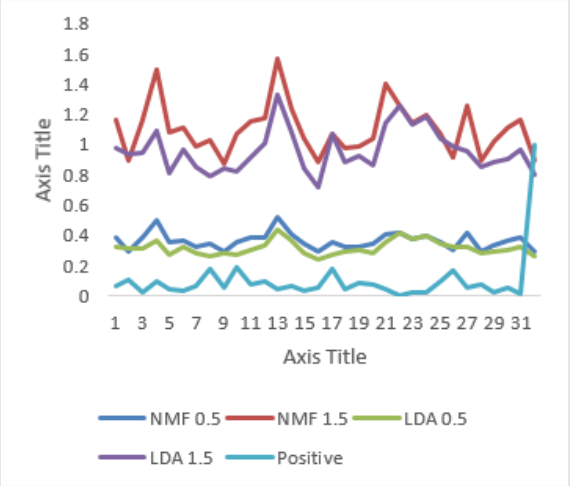

Figure 6.7: Correlation between Positive Prevalence and Potentials

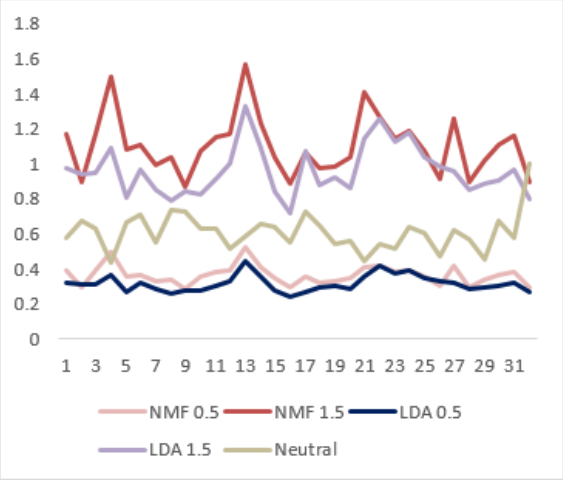

Figure 6.8: Correlation between Neutral Prevalence and Potentials

Correlation between proposed model and Information diffusion model

Application 1: Food Poisoning Data

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Alpha value	0.059451	0.05885	0.118331	0.118071

Table 6.3: Correlation between Alpha value and Potential on Food Poisoning data

Application 2: Immigration Data

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Alpha value	0.164799	0.179793	0.362735	0.330362

Table 6.4: Correlation between Alpha value and Potential on Immigration data

From Table 6.3 and Table 6.4 the correlation between alpha value and potentials of tweets are positively correlated and LDA approach is more positively correlated when compared to NMF approach.

6.3.2 Normalization

Z-score Normalization Gopal and Kishore (2015) is a technique which normalizes values or range of data from the original unstructured data by using mean and standard deviation.

Z-score normalization is calculated as $v_i\prime = (v_i\text{-}\overline{E})/std(E)$

Where,

$v_i\prime$is Z-score normalized one values.

v_i is value of the row E of ith*column*

$std(E) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(v_i - \overline{E})^2}$

$\overline{E} = \frac{1}{n}\sum_{i=1}^{n} v_i$ Or mean value

Here the scale of potentials varies from 0 to 2, prevalence range is between 0 and 1 and alpha values scales between 1 and 3. Since the model's scales are different and unable to compare we have used z-score, (more commonly referred to as a standard score) a measure of how many standard deviations below or above the population mean a raw score is. The population here is potential for the proposed model, the prevalence for sentiment analysis and alpha value for Information Diffusion.

	NMFρ0.5	**NMFρ1.5**	**LDAρ0.5**	**LDAρ1.5**	**Prevalence**	**Alpha value**
Food Poisoning	0.72513	0.74694	0.6452	0.6447	0.7135	0.710008
Immigration	0.761198	0.7597	0.7483	0.77892	0.74162	0.670175

Table 6.5: Normalized values

Normalization is used as a standardized method where the values range between 0 and 1. And this values are used to compare the models. For the food poisoning data, the normalized value is high at ρ =1.5 for the NMF topic extraction model. In Immigration data, the normalized value is high at ρ=1.5 for the LDA approach, which indicates that these approaches are better to quantify information.

CHAPTER VII

FORECASTING

Time series forecast is the process of predicting future events based on historical data. Time series Forecast can be split into two terms Time series and Forecast, where Time series is a sequence of observations taken sequentially in time and Forecast means making predictions about a future event.

In this section, we will see the analysis of forecasted time series data using a deep learning technique long short term memory (LSTM) (Table 9.1 row 6) algorithm. The core components of an LSTM network are a sequence input layer and an LSTM layer. A sequence input layer inputs sequence or time series data into the network. An LSTM layer is a type of recurrent neural network (RNN) that learns long-term dependencies between time steps of sequence data.

Long Short-Term Memory models can predict an arbitrary number of steps into the future. An LSTM module (or cell) contains 5 essential components which allows it to model both long-term and short-term data.

Cell state (c_t) It represents the internal memory of the cell which stores both short term and long-term memories.

Hidden state (h_t) It is the output state information calculated with respect to current input, previous hidden state and current cell input which is eventually used to predict the future values.

Input gate (i_t) Decides how much information from the current input flows to the cell state.

Forget gate (f_t) - Decides how much information from the current input and the pre-

vious cell state flows into the current cell state

Output gate (o_t) - Decides how much information from the current cell state flows into the hidden state.

Absolute error is used as an evaluation measure in the forecasting model.

Absolute Error: The mean of absolute difference between the predicted and actual values.

Absolute Error = $\sum (Predictedvalue - Actualvalue)/N$ Where N is number of observations.

Since the model is trained with the historical data points, we can predict the future data in the long run.

7.1 Forecasting Proposed Model

Application 1: Food Poisoning data

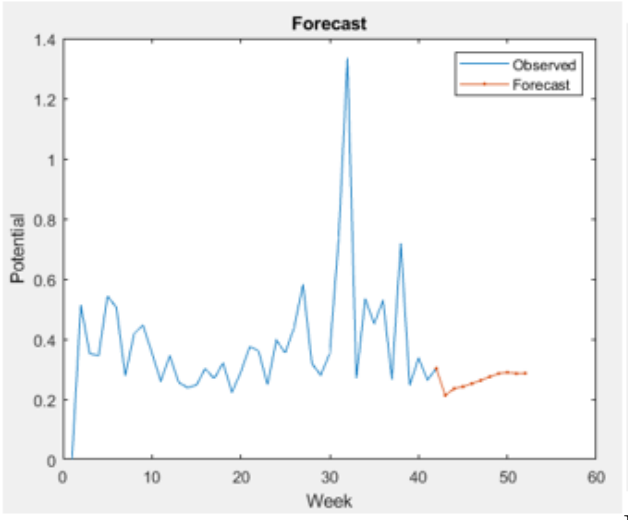

Figure 7.1: Forecasted potential for LDA at $\rho = 0.5$

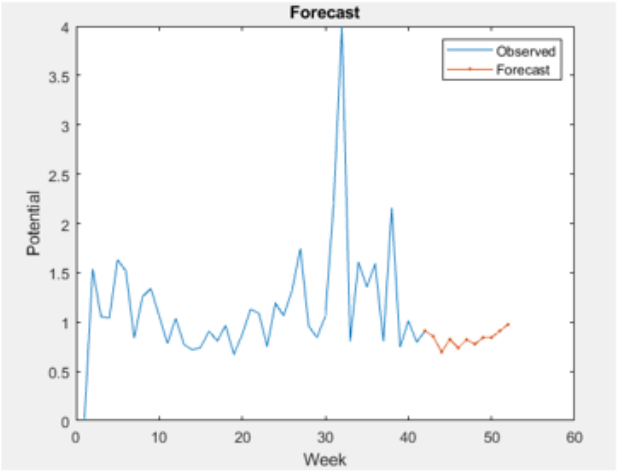

Figure 7.2: Forecasted potential for LDA at $\rho = 1.5$

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Absolute Error	0.1920	0.06259	0.094	0.2638

Table 7.1: Absolute error for the forecasted potentials on Food Poisoning Data

Figure 7.1, 7.2, 7.3 and 7.4 are forecasting the potentials of Food posing data for

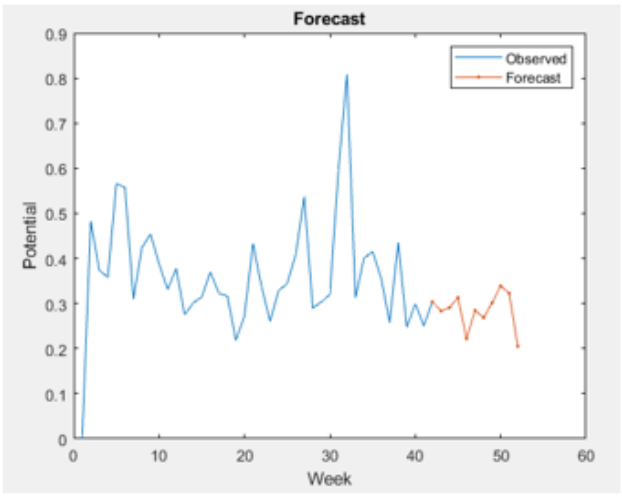

Figure 7.3: Forecasted potential for NMF at $\rho = 0.5$

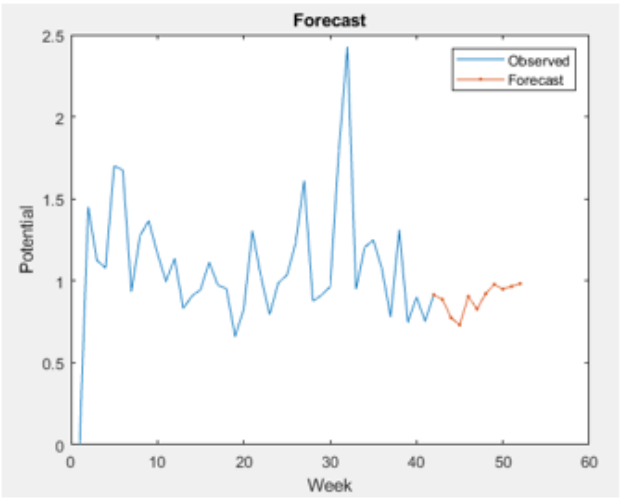

Figure 7.4: Forecasted potential for NMF at $\rho = 1.5$

different term extraction approaches LDA and NMF and at two different arbitrary constants 0.5 and 1.5. Below Figures 7.5, 7.6, 7.7 and 7.8 are the forecasting results for the potentials on Immigration data.

Application 2: Immigration Data

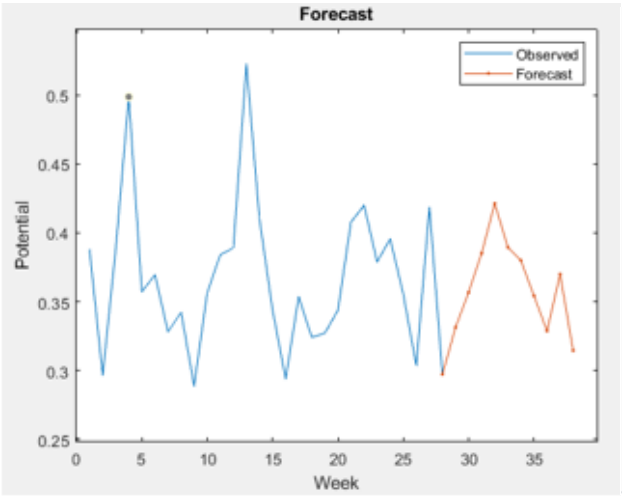

Figure 7.5: Forecasted potential for NMF at $\rho = 0.5$

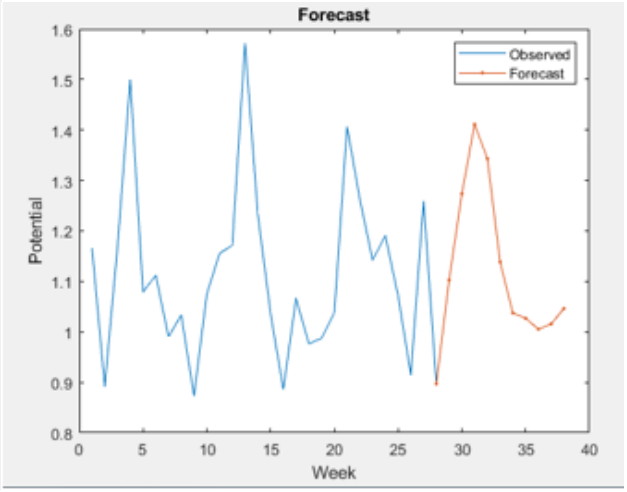

Figure 7.6: Forecasted potential for NMF at $\rho = 1.5$

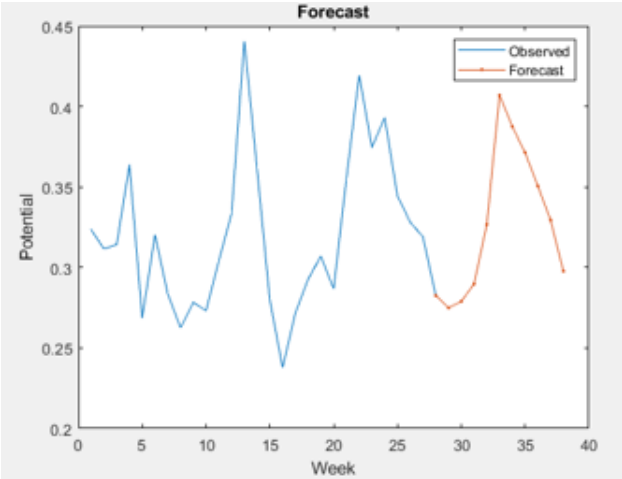

Figure 7.7: Forecasted potential for LDA at $\rho = 0.5$

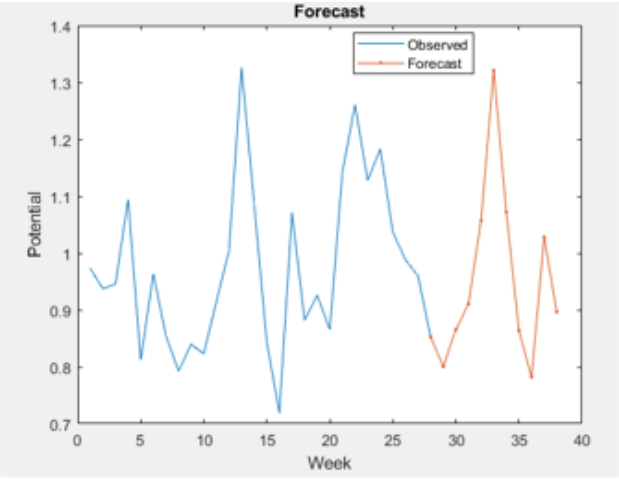

Figure 7.8: Forecasted potential for LDA at $\rho = 1.5$

	NMF ρ=0.5	**NMF ρ=1.5**	**LDA ρ=0.5**	**LDA ρ=1.5**
Absolute Error	0.05334	0.200338	0.0522	0.1407

Table 7.2: Absolute error for the forecasted potentials on Immigration Data

7.2 Comparing Model 1: Sentiment Analysis Forecasting

Application 1: Food Poisoning data

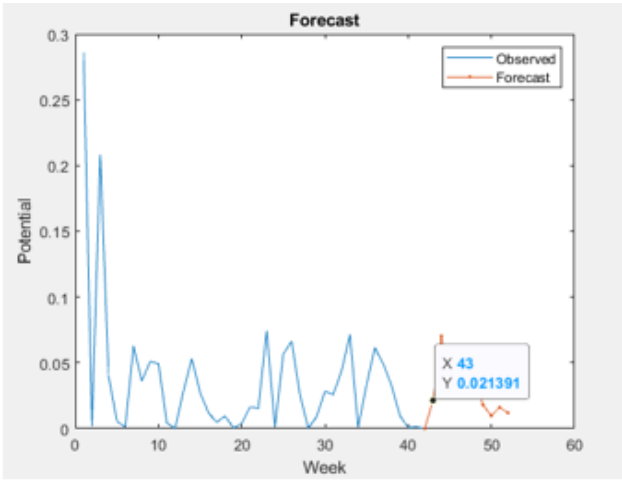

Figure 7.9: Forecasted Positive prevalence value

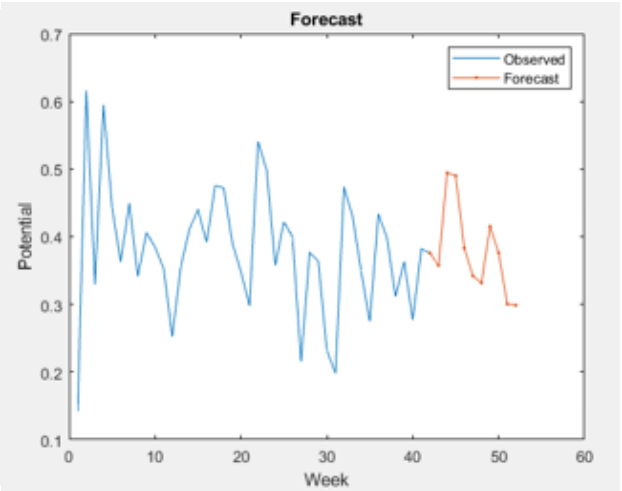

Figure 7.10: Forecasted Neutral prevalence value

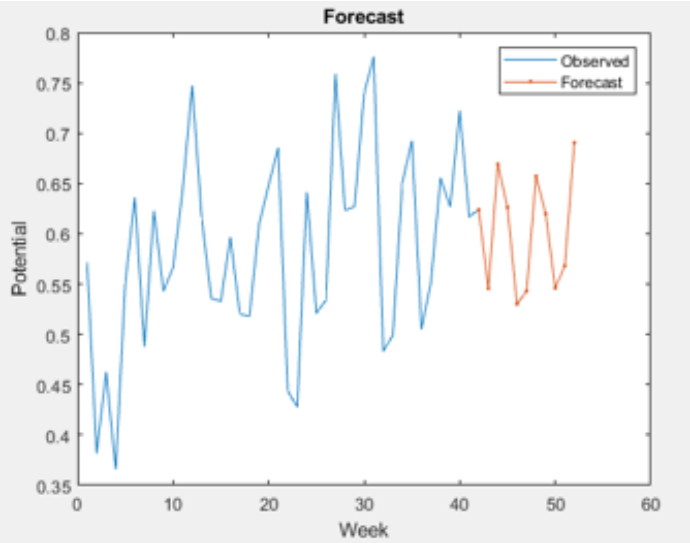

Figure 7.11: Forecasted Negative prevalence value

	Positive	**Negative**	**Neutral**
Absolute Error	0.0466	0.05737	0.068894

Table 7.3: Absolute error for the forecasted prevalence on Food Poisoning Data

Figure 7.9, 7.10 and 7.11 are forecasting the prevalence (Positive, Negative and Neutral) of Food posing data. Below Figures 7.12, 7.13, and 7.14 are the forecasting the prevalence on Immigration data.

Application 2: Immigration Data

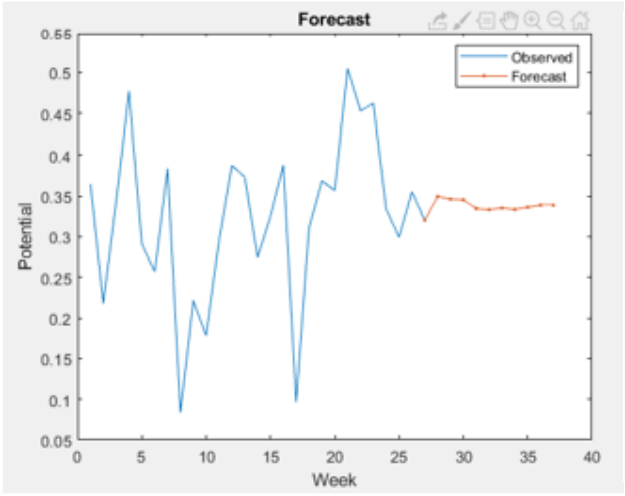

Figure 7.12: Forecasted Negative prevalence value

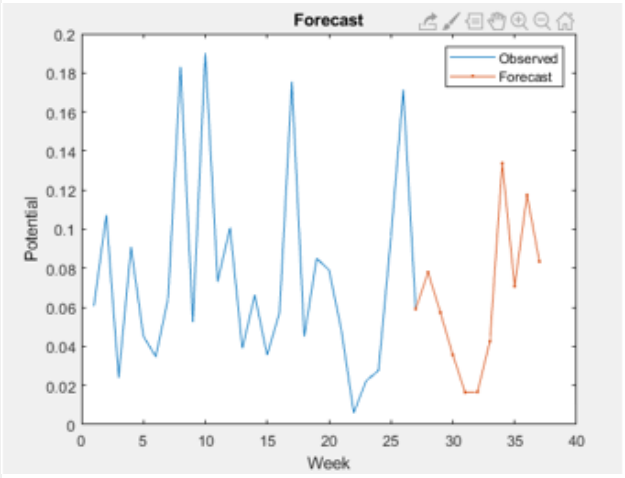

Figure 7.13: Forecasted Positive prevalence value

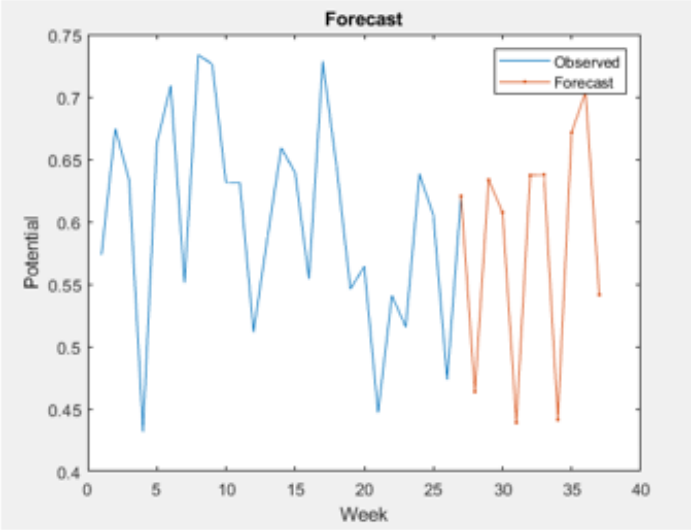

Figure 7.14: Forecasted Neutral prevalence value

	Positive	**Negative**	**Neutral**
Absolute Error	0.03125	0.0633	0.128803

Table 7.4: Absolute error for the forecasted prevalence on Immigration Data

7.3 Comparing Model 2: Information Diffusion Forecasting

Application 1: Food Poisoning data

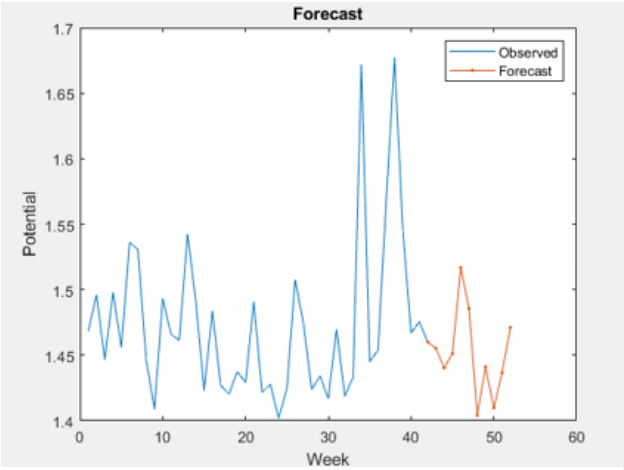

Figure 7.15: Forecasted alpha value

Application 2: Immigration Data

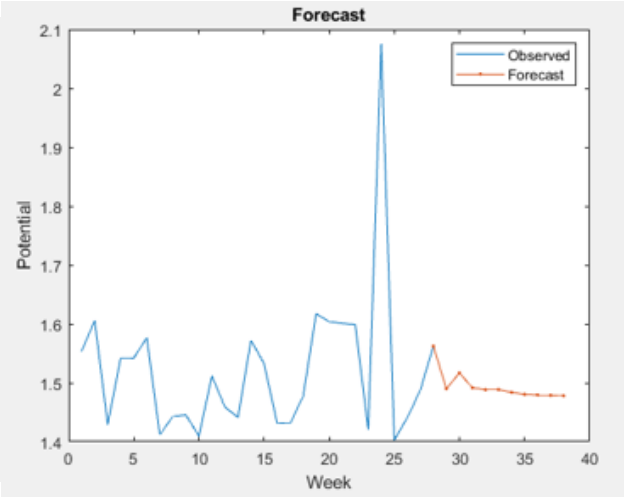

Figure 7.16: Forecasted alpha value

Figure 7.15 is forecasting the Alpha value of Food poising data and Figure 7.16 is forecasting the alpha value of Immigration data.

	Food Poisoning	**Immigration**
Absolute Error	0.0350	0.06856

Table 7.5: Absolute error for the forecasted alpha value

From Figures 7.1, 7.2, 7.3, 7.4, 7.12 and 7.16, the forecasted pattern varies from the observed pattern. Therefore, the LSTM model might not be the best way to predict future events, and so, we couldn't come up with an evaluation measure for comparing the models.

CHAPTER VIII

ANALYSIS

Application1: Food Poisoning

From Table 6.1, negative and neutral prevalences are positively correlated with potentials from the topic extraction approaches. Compared to other pairs in the table, neutral prevalence and topic extracted by the NMF approach are shown to have better positively correlated values.

From Table 8.1, the correlation between alpha value and the potentials are positive. Alpha values are more positively correlated for NMF topic extraction method compared to LDA topic extraction method.

From Figure 6.1, the negative prevalence of sentiment quantification has more negative values compared to other prevalences which indicates that there are more negative tweets. In domains such as food poisoning, negative words are considered to be positive indicators. Therefore, considering the negative sentiment in sentiment analysis might not be the best approach to quantify information. These results show that information quantification methods are depended on the type of topic being considered.

Normalized data with respect to weeks is represented in Table 9.2. Tokens extracted by the NMF approach from the proposed model excelled when compared to sentiment quantification and information diffusion models. For the domain like food poisoning, topics obtained by the NMF approach would be better to quantify the information.

Application2: Immigration data

In Table 7.2, negative prevalence is positively correlated with the potential measures. Whereas, positive and neutral prevalences are negatively correlated with potentials. This result and the result from table 6.1 mentioned in application1 indicate that sentiment data measured separately may not measure the same concept.

From Table 8.2, LDA topic extraction methods are more positively correlated with alpha value compared to NMF topic extraction method.

Normalized data with respect to weeks is represented in Table 9.2. Tokens extracted by the LDA approach from the proposed model outperformed when compared to sentiment quantification and information diffusion models. From Figure 9.1, we observed an inconclusive graph for the power law model where the followers count for week5, week21 and week23 are not proportional to the alpha value, and the observed retweet count and the real retweet count varies. This might not be the best approach to quantify information.

Therefore, for the domain like Immigration, topics extracted by the LDA method would be the optimal approach to quantify the information.

CHAPTER IX

CONCLUSION

As social media become a source of information overtaking the traditional media, where people post their real-time experiences and their opinions on various day-day issues, methods to quantify information from tweets would be beneficial. In this thesis, we proposed a method to quantify the information in tweets. The proposed model is based on weight assignment to tokens in tweets. Two approaches are proposed for building tokens associated with a set of tweets. One approach is topic modeling. And the other is an iterative approach; new algorithms are developed for the selection and assignment of weights to the tokens. The proposed model is compared against two previously published models. The comparison shows that the domain of tweets influences quantification. The usefulness of quantification for forecasting is also demonstrated.

Including external factors such as user influence in the potential computation are proposed as future work.

External Links

Sr No	Links
(1)	Apache Hadoop, https://hadoop.apache.org/
(2)	Apache Flume, https://flume.apache.org/
(3)	Afinn, http://corpustext.com/reference/sentiment_afinn.html
(4)	Cornel Movie Dataset, http://www.cs.cornell.edu/people/pabo/movie-review-data/
(5)	CDC, https://www.cdc.gov/foodsafety/cdc-and-food-safety.html
(6)	LSTM, https://medium.com/microsoftazure/neural-networks-for-forecasting-financial-and-economic-time-series-6aca370ff412

Table 9.1: External Links

REFERENCES

Adetunmbi, A. O., Sarumi, O. A., and Boyinbode, O. (2018). Machine learning approach to sentiment analysis of users movie reviews. pages 327–332.

Ahmed, K., El Tazi, N., and Hossny, A. H. (2015). Sentiment analysis over social networks: An overview. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2174–2179. IEEE.

Arash, M. and Elham, D. (2017). A sentiment analysis approach using effective feature reduction method. pages 10–14.

Bickle, A. (2010). The k-cores of a graph.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation journal of machine learning research (3).

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Bouazizi, M. and Ohtsuki, T. (2016a). Sentiment analysis in twitter: From classification to quantification of sentiments within tweets. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE.

Bouazizi, M. and Ohtsuki, T. O. (2016b). A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

Choi, S. (2011). Probablistic latent semantic analysis.

Esuli, A. and Sebastiani, F. (2007). A high-coverage lexical resource for opinion mining. Technical report, Technical Report 2007-Tr-02. Istitutiodi Scienza e Technologie dell .

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Gao, W. and Sebastiani, F. (2015). Tweet sentiment: From classification to quantification. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 97–104. ACM.

Gopal, S. and Kishore, K. (2015). Normalization: A preprocessing stage. In *CSE & IT department, VSSUT*.

Hasan, A., Moin, S., Karim, A., and Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, 23(1):11.

Hoang, B.-T., Chelghoum, K., and Kacem, I. (2016). A learning-based model for predicting information diffusion in social networks: Case of twitter. In *2016 International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 752–757. IEEE.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., et al. (2014). Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.

Losee, R. M. (1997). A discipline independent definition of information. *Journal of the American Society for information Science*, 48(3):254–269.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics*, 55(1):1–12.

McLachlan, G. and Peel, D. (2000). Finite mixture models. wiley-interscience.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, 43:412–422.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Reed, C. (2012). Latent dirichlet allocation: Towards a deeper understanding. *unpublished*.

Remy, C., Pervin, N., Toriumi, F., and Takeda, H. (2013). Information diffusion on twitter: everyone has its chance, but all chances are not equal. In *2013 International*

Conference on Signal-Image Technology & Internet-Based Systems, pages 483–490. IEEE.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1105–1114. International World Wide Web Conferences Steering Committee.

Stai, E., Milaiou, E., Karyotis, V., and Papavassiliou, S. (2018). Temporal dynamics of information diffusion in twitter: Modeling and experimentation. *IEEE Transactions on Computational Social Systems*, 5(1):256–264.

TK, A. K. and George, K. (2016). Application of an asset bubble model to microblog data analytics. In *2016 IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW)*, pages 19–27. IEEE.

TK, A. K., George, K., and Thomas, J. P. (2015). An empirical approach to detection of topic bubbles in tweets. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, pages 31–40. IEEE.

Wang, D., Musaev, A., and Pu, C. (2016). Information diffusion analysis of rumor dynamics over a social-interaction based model. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 312–320. IEEE.

Yang, J. and Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE.

Zhu, H., Kong, Y., Wei, J., and Ma, J. (2018). Effect of users opinion evolution on information diffusion in online social networks. *Physica A: Statistical Mechanics and its Applications*, 492:2034–2045.

APPENDICES

Twitter data streaming configuration file:

TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = CSiRReE3UPZrfdcPysN6mlV9D
TwitterAgent.sources.Twitter.consumerSecret =
DTOqrR7zcOpLTpcU7AlhHmQHP18GPo04NhSqwq2PRiYxyEtkUX
TwitterAgent.sources.Twitter.accessToken = 735984908959547392-
nGGuCX9QQwIbOY964ycKUyTtOYR2sUA
TwitterAgent.sources.Twitter.accessTokenSecret =
psYQPQFaYIubys8aIWrHxHo4D4FktVIc1trvutkSYOchC
TwitterAgent.sources.Twitter.keywords = Diarrhea, abdominal pain, vomiting,
puke, fever
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path=
hdfs://hadoop1:9000/rramine/Food_data/%Y/%m/%d/%H
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwtterAgent.sinks.HDFS.hdfs.rollCount = 0
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 10000

Sample JSON format file

{"extended_tweet":{"entities":{"urls":[],"hashtags":[{"indices":[129,140],"text":"IndianArmy"}],"user_mentions":[{"indices":[10,23],"screen_name":"richardrekhy","id_str":"134055679","name":"Richard Rekhy","id":134055679},{"indices":[79,85],"screen_name":"adgpi","id_str":"1227253801","name":"ADG PI - INDIAN ARMY","id":1227253801}],"symbols":[]},"full_text":"Thank you @richardrekhy sir for your kind praise. I am humbled. \n\nThis is what @adgpi taught me, to keep commitment. \n\nJaihind \n\n#IndianArmy","display_text_range":[0,140]},"quoted_status":{"extended_tweet":{"entities":{"urls":[{"display_url":"twitter.com/richardrekhy/s\u2026","indices":[281,304],"expanded_url":"https://twitter.com/ri

chardrekhy/status/1058562665570885632","url":"https://t.co/XY26ihMZma"}],"hashtags":[],"user_mentions":[{"indices":[84,95],"screen_name":"MajDPSingh","id_str":"423362558","name":"Major D P Singh","id":423362558}],"symbols":[]},"full_text":"I am learning so much about these brave hearts . We had the opportunity of inviting @MajDPSingh to our three events at 3 diff cities while at KPMG. He totally inspired and mesmerised the audience. What is more on one occasion he was running high fever but he kept his commitment.
https://t.co/XY26ihMZma","display_text_range":[0,280]},"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat Nov 03 04:33:18 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone<\/a>","quoted_status_id":1058562665570885632,"retweet_count":0,"retweeted":false,"geo":null,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":true,"id_str":"1058577838926950405","in_reply_to_user_id":null,"favorite_count":4,"id":1058577838926950405,"text":"I am learning so much about these brave hearts . We had the opportunity of inviting @MajDPSingh to our three events\u2026
https://t.co/puoHH3ieM6","place":{"country_code":"IN","country":"India","full_name":"Bengaluru, India","bounding_box":{"coordinates":[[[77.373474,12.919037],[77.373474,13.231381],[77.739371,13.231381],[77.739371,12.919037]]],"type":"Polygon"},"place_type":"city","name":"Bengaluru","attributes":{},"id":"1b8680cd52a711cb","url":"https://api.twitter.com/1.1/geo/id/1b8680cd52a711cb.json"},"lang":"en","quote_count":0,"favorited":false,"possibly_sensitive":false,"coordinates":null,"truncated":true,"reply_count":0,"entities":{"urls":[{"display_url":"twitter.com/i/web/status/1\u2026","indices":[117,140],"expanded_url":"https://twitter.com/i/web/status/1058577838926950405","url":"https://t.co/puoHH3ieM6"}],"hashtags":[],"user_mentions":[{"indices":[84,95],"screen_name":"MajDPSingh","id_str":"423362558","name":"Major D P Singh","id":423362558}],"symbols":[]},"display_text_range":[0,140],"quoted_status_id_str":"1058562665570885632","contributors":null,"user":{"utc_offset":null,"friends_count":925,"profile_image_url_https":"https://pbs.twimg.com/profile_images/666563812141592576/HvBhLCFQ_normal.jpg","listed_count":126,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","default_profile_image":false,"favourites_count":7897,"description":"I don't know how my story will end but nowhere in my text will it ever read 'I GAVE UP';Passionate.Views expressed are personal.RT's do not imply endorsements.","created_at":"Sat Apr 17 09:52:20 +0000 2010","is_translator":false,"profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png","protected":false,"screen_name":"richardrekhy","id_str":"134055679","profile_link_color":"1DA1F2","translator_type":"none","id":134055679,"geo_enabled":true,"profile_background_color":"C0DEED","lang":"en","profile_sidebar_border_color":"C0DEED","profile_text_color":"333333","verified":true,"profile_image_url":"http://pbs.twimg.com/profile_images/666563812141592576/HvBhLCFQ_normal.jpg","time_zone":null,"url":null,"contributors_enabled":false,"profile_background_tile":false,"profile_banner_url":"https://pbs.twimg.com/profile_banners/134055679/1447756187","statuses_count":11052,"follow_request_sent":null,"followers_count":7010,"profile_use_background_image":true,"default_profile":true,"following":null,"name":"Richard Rekhy","location":"ÃœT: 28.368607,77.184336","profile_sidebar_fill_color":"DDEEF6","notifications":null}},"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat Nov 03 07:00:46 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android<\/a>","quoted_status_id":1058577838926950405,"retweet_count":0,"retweeted":false,"geo":null,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":true,"id_str":"1058614949927403520","in_reply_to_user_id":null,"favorite_count":0,"id":1058614949927403520,"text":"Thank you @richardrekhy sir for your kind praise. I am humbled. \n\nThis is what @adgpi taught me, to keep

commitment\u2026
https://t.co/jbWGJgFGnC","place":null,"quoted_status_permalink":{"expanded":"https://twitter.com/richardrekhy/status/1058577838926950405","display":"twitter.com/richardrekhy/s\u2026","url":"https://t.co/tAAmvjarcs"},"lang":"en","quote_count":0,"favorited":false,"coordinates":null,"truncated":true,"timestamp_ms":"1541228446834","reply_count":0,"entities":{"urls":[{"display_url":"twitter.com/i/web/status/1\u2026","indices":[117,140],"expanded_url":"https://twitter.com/i/web/status/1058614949927403520","url":"https://t.co/jbWGJgFGnC"}],"hashtags":[],"user_mentions":[{"indices":[10,23],"screen_name":"richardrekhy","id_str":"134055679","name":"Richard
Rekhy","id":134055679},{"indices":[79,85],"screen_name":"adgpi","id_str":"1227253801","name":"ADG PI - INDIAN
ARMY","id":1227253801}],"symbols":[]},"quoted_status_id_str":"1058577838926950405","contributors":null,"user":{"utc_offset":null,"friends_count":530,"profile_image_url_https":"https://pbs.twimg.com/profile_images/909587248890355712/gj9T0_xv_normal.jpg","listed_count":60,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","default_profile_image":false,"favourites_count":8220,"description":"Fought KargilWar &enjoying its woundsðŸ˜‰.1st amputee marathoner of India.4 Limca records. Inspirational speaker.Founder 'The Challenging Ones'
@majdpsingh_TCO","created_at":"Mon Nov 28 12:10:23 +0000
2011","is_translator":false,"profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png","protected":false,"screen_name":"MajDPSingh","id_str":"423362558","profile_link_color":"89C9FA","translator_type":"none","id":423362558,"geo_enabled":true,"profile_background_color":"C0DEED","lang":"en","profile_sidebar_border_color":"C0DEED","profile_text_color":"333333","verified":true,"profile_image_url":"http://pbs.twimg.com/profile_images/909587248890355712/gj9T0_xv_normal.jpg","time_zone":null,"url":"http://www.majordpsingh.com","contributors_enabled":false,"profile_background_tile":false,"profile_banner_url":"https://pbs.twimg.com/profile_banners/423362558/1499767353","statuses_count":12332,"follow_request_sent":null,"followers_count":14721,"profile_use_background_image":true,"default_profile":false,"following":null,"name":"Major D P Singh","location":"new
delhi","profile_sidebar_fill_color":"DDEEF6","notifications":null}}
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat Nov 03 07:00:47 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone<\/a>","retweet_count":0,"retweeted":false,"geo":null,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":false,"id_str":"1058614954507542528","in_reply_to_user_id":null,"favorite_count":0,"id":1058614954507542528,"text":"slight tw for vomiting // I\u2019m definitely feeling sick as hell now and I almost threw up in the bathroom a few minutes ago it\u2019s
disgusting","place":null,"lang":"en","quote_count":0,"favorited":false,"coordinates":null,"truncated":false,"timestamp_ms":"1541228447926","reply_count":0,"entities":{"urls":[],"hashtags":[],"user_mentions":[],"symbols":[]},"contributors":null,"user":{"utc_offset":null,"friends_count":553,"profile_image_url_https":"https://pbs.twimg.com/profile_images/1057506030702206976/Pm92hiRH_normal.jpg","listed_count":9,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","default_profile_image":false,"favourites_count":19501,"description":"\u201cyou don\u2019t get to destroy who i am.\u201d | real-life jessica jones","created_at":"Fri Jul 20 01:55:39 +0000
2018","is_translator":false,"profile_background_image_url_https":"https://abs.twimg.com/imag

es/themes/theme1/bg.png","protected":false,"screen_name":"WomynOfMarvel","id_str":"1020125048378667009","profile_link_color":"7FDBB6","translator_type":"none","id":1020125048378667009,"geo_enabled":false,"profile_background_color":"000000","lang":"en","profile_sidebar_border_color":"000000","profile_text_color":"000000","verified":false,"profile_image_url":"http://pbs.twimg.com/profile_images/1057506030702206976/Pm92hiRH_normal.jpg","time_zone":null,"url":"https://curiouscat.me/WomynofMarvel","contributors_enabled":false,"profile_background_tile":false,"profile_banner_url":"https://pbs.twimg.com/profile_banners/1020125048378667009/1540964061","statuses_count":6298,"follow_request_sent":null,"followers_count":870,"profile_use_background_image":false,"default_profile":false,"following":null,"name":"karen page defense squad ï¸½âœµï¸½","location":"The New York Bulletin","profile_sidebar_fill_color":"000000","notifications":null}}

VITA

RUCHISHYA RAMINENI

COMPUTER SCIENCE

Master of Science

Thesis: AN APPROACH TO QUANTIFY INFORMATION IN TWEETS

Major Field: Computer Science

Biographical:

Education:
Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in May 2019.