Analysing Product Reviews Using Deep-Learning Model

A DISSERTATION

submitted in partial fulfillment of the requirements
for the award of the degree of

Master of Technology

in

COMPUTER SCIENCE AND ENGINEERING

by

NITESH KUMAR RAI
(17535019)

Under the supervision of

Prof. Manoj Misra
Professor, Dept. of CSE

Prof. Durga Toshniwal
Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
ROORKEE – 247667, UTTRAKHAND (INDIA)
MAY, 2019

CANDIDATE DECLARATION

I hereby certify that the work which is being presented in the M.Tech. Dissertation entitled **"Analysing Product Reviews using Deep-Learning Model"**, in partial fulfillment of the requirements for the award of the **Master of Technology in Computer Science and Engineering** is an authentic record of my own work carried out during a period from Aug 2018 to May 2019 under the supervision of **Prof. Manoj Misra** and **Prof. Durga Toshniwal , Professor** Department of Computer Science and engineering, Indian Institute of Technology Roorkee.

The matter presented in this thesis has not been submitted for the award of any other degree elsewhere.

Nitesh Kumar Rai
(17535019)

CERTIFICATE

This is to certify that thesis report entitled **"Analysing Product Reviews using Deep-Learning Model"** which is submitted by **Nitesh Kumar Rai (17535019)** towards the fulfilment of the requirements for the award of the Degree of **Master of Technology in Computer Science & Engineering**, submitted to the Department of Computer Science & Engineering, **Indian Institute of Technology Roorkee, India** is carried out by her under my esteemed supervision and the statement made by the candidate in the declaration is correct to the best of my knowledge and belief.

Date :

Prof. Manoj Misra	**Prof. Durga Toshniwal**
Professor, Dept. of CSE	Professor, Dept. of CSE
IIT Roorkee	IIT Roorkee

ACKNOWLEDGEMENT

First of all, I express my gratitude to the Almighty, who blessed me with the zeal and enthusiasm to complete this research work successfully. I am extremely thankful to my supervisors **Prof. Manoj Misra** and **Prof. Durga Toshniwal** for their motivation and tireless efforts to help me to get deep knowledge of the research area and supporting me throughout the life cycle of my M. Tech. dissertation work. Especially, the extensive comments, healthy discussions, and fruitful interactions with the supervisors had a direct impact on the final form and quality of M. Tech. dissertation work.

I am also thankful to the faculty members and supporting staff of Computer Engineering Department for their full support and heartiest co-operation.

This thesis would not have been possible without the hearty support of my friends. My deepest regards to my Parents for their blessings, affection and continuous support. Also, Last but not the least, I thank to the GOD, the almighty for giving me the inner willingness, strength and wisdom to carry out this research work successfully.

Nitesh Kumar Rai
(17535019)

ABSTRACT

Due to explosive evolution and popularity of electronic media, Online shopping and Social media sites, vast amount of user review and experience available in the form of raw data. It can be used for opining mining or sentiment mining and other pattern identification tasks. Opining mining or sentiment mining and summerization of review regarding any particular topic used to provide insights and can be used as feedback to improve or address concerns regarding that topic and helpful in future planning. Most of the work done so far in this field foced on run of the mill, well defined techniques like K-NN, SVM and others machnine learning algorithms to classify the text into two or more classes. However, traditional techniques peak out, in term of accuracy in certain limit. Additional improvement in term of accuracy reported using deep learning model LSTM-RNN with pre-trained word embedding. The aim of the present work is to improve existing techniques for opinion mining or sentiment analysis.

Contents

List of Figures

List of Tables

List of Abbreviations

LDA	Latent Dirichlet Allocation
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
k-NN	k-Nearest Neighbors
SVM	Support-Vector Machines
HWE	Hybrid Word Embedding
CNN	Convolutional Neural Network

Chapter 1

INTRODUCTION

In the era of digital evaluation, social media like Facebook, Twitter, Instagram, etc and online shopping sites like Amazon, Flipkart, etc has caused enormous amounts of public data which can be collected, processed and used to perform diferents type of analysis on that data to solve various problem. It can solve many industry problems like future needs, sell or buy prediction, and feedback of the product. Amazon and Twitter have millions of users who share their opinion regarding any product or topic on that platform , which can be used to opinion mining or sentiment analysis regarding topic. This analysis can help in decision making in various domains and provide insights of topic. Due availability of public data easily in the large amount applicable in many application like sentiment analysis or opinion mining, recommendation system, revenue prediction, etc.

1.1 Motivation

Opinion mining also is known as sentiment mining, opinion extraction, and subjective analysis. It is cost effective and fair technique to determine public opinion towards any topic. Reviews available in textual form. It has been proved that there exists a positive correlation between the sentiment analyzed on the social media data and the event organised in the actual world. In customer-relationship management for large business companies it is essential that in addition to identifying trend of the opinion of product, or to find the areas where the customers are facing diffulties so that they can be addressed by company immediately. Althought collection of data present little bit problems. Opinion mining and sentiment analysis has successfully used in various domains like determing movie rating, feedback of hotel services, food services, etc and finding of users needs. In politics to improve campaigning process by delivering speeches on topics that the public(voters) is most interested, which is obtain through opining mining and sentiment analysis on their data available through social media like Twitter and facebook. For example Obama adminstration used sentiment analysis to gauge the public opinion to policy announcement and campaign message prior to 2012 presidental election [3].

Although collection of data present little problems, it is the interpretation of data that is challenging. Traditional Techniques have not preformed well due to the presence of noise, non-standard characters, specail charters, smileys, misspelled words, regainoal slang words, short handed words might be present in the communication.

Sentiment analysis can be broadly classified into two areas lexicon-based approach and deep learning approach. In lexicon based approaches the input data after modelling is compared with the predefined lexicon with labelled classes to determine the overall sentiment of the input text. This suffers from the drawback that it cannot identify subtleties inherent to the language like sarcasm, use of metaphors. Using a lexicon based approach for sentiment analysis can however provide a reasonable between the accuracy and the complexity needed. Clever use of the obtained results can be used immediately.

Reviews consist of several redundancies and noise that is an obstacle for the lexicon-based approach. Thus pre-processing step is addressed these issues becomes a vital step in the process. Pre-processing of textual data involve these several steps.

- Removal of stop words
- Removal of duplicates
- Stemming/Lemmatization
- Noise/URLs/HTML Tags removal

Once this step is done the output of this process is given to the lexicon-based sentiment analysis . Algorithm compared with labled words and return the score of each sentence to determine sentiment of the people regarding given topic.

Opinion Mining or Sentiment analysis is the detection of attitudes. It involes identification of source, target and type of altitude (polarity).

1.2 Problem Description

"Analysing review data using deep-learning for improve performance of deep neural network classification model"

To address the problem stated, we collected the dataset from Amazon shopping sites & Twitter, and trained the deep learning model on this data using different parameters. Word embeddings for the dataset was used and also used availbale pre-trained word emmbeding like glove in Sequence Classification with LSTM Recurrent Neural Networks deep learning model to improve accuracy and perform sentiment analysis.

1.3 Organization of the Report

The rest of the report is organized as follows: Section 2 describes the previous related work in the field of sentiment analysis an analysis or opinion mining. Section 3 describes proposed work in depth about preprocessing, data collection and classification method. Section 4 shows the experiment and results. In the last part we conclude the thesis work.

Chapter 2

LITERATURE REVIEW

Due to easy availability of large dataset in field of Sentiment analysis and text mining lot of research work has been done. Many of the previously described models attempts to group the dataset into two classes the positive and negative classes and have achieved the best possible results in this area. Many research has been focused on performing fine-grained sentiment analysis which involves classifying a given data sample one among several classes each idicating a varying degree of sentiment. Lately deep learning-based approaches have taken a lead in this regard. Sentiment analysis has been tackled by using several well defined existing techniques like Naive bayes, SVM and k-Nearest Neighbors(k-NN) and other classification algorithms that establish the tradeoff between using these techniques for several datasets [4] [5] [6]. These work attempts to find which of the wel -establish techniques is best suited for sentiment analysis.

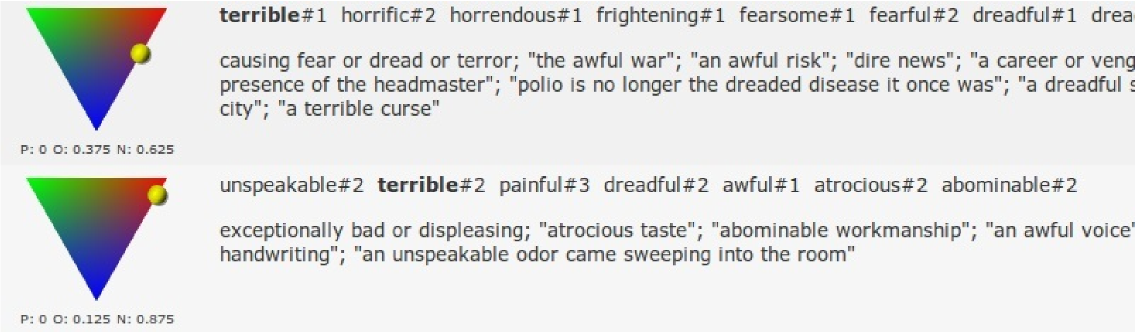
People belongs to different regions and different languages. People gave a review in own languages so that this is diffcult task to handle diffrent languages. Each language has an inherent structure that various greathly between different languages [7]. Many researchers attempts to identify the as particular variation and fine tuning of the parameters of the sentiment analysis model that are best suited for use on a particular language like chines whose sentence semantic structure varies from english [4]. Input data required for sentiment analysis alogrithm comes in various formats. So that each algorithm required own preprocessing steps [3]. Social media data crawled from public API provided by social media like Twitter provide Twitter

Streamming API [8]. Twitter has short text and unstructured making it harder for identifying the true sentiment of text [3] [9] .

Twitter has Hastag and mention with every tweets. Hashtag have metadata regarding tweet and provide additional information about tweets which is helps to identify the targets or the subject referred to in the tweet [4]. These hashtags also provide additional information related to the given tweet. Identification of aspect targets before sentiment analysis has improved the result of conventional generative probablistics model like LDA. [3] tan et al. have used a hierarchile LDA model to first identify candidate tweets. Based on these candidate tweets , the foreground and background tweets have been collected and sentiment analysis has been done specific to these tweets. Gibbs sampling has been finally used to identify the sentiment in the respective candidate tweet. Latent Dirichlet Allocation (LDA) has been found to be the most efficient topic modelling when large amounts of data are to be processed. Only some of the drawbacks of LDA have been addressed and possible approaches to perform complex sentiment analysis using variations of the basic LDA have been addressed in [3]. It has been reported that using two seprate LDA where the result of one LDA has been utlized as input to second LDA provies better analysis of Opinion mining.

2.1 Lexicon Based Approach

Lexicon Based approach is a basic approach for sentiment classification and it is a unsupervised learning model for sentiment analysis [10] [1]. The most advantage of this technique is no need to label the dataset. Labelling the data set is one of the hard-working steps. Performance of the lexicon-based approach is totally depends on the quality of lexical resource it relies on. Some of state-of-the-art resource available for lexicon based sentiment analysis like SenticNet [11], SentiWordNet [12], WordNet-affect [13] etc. In lexical resource words are mapping with a category like positive, neutral or negative or associated with sentiment score which is used by lexicon-based algorithm to acquire overall sentiment communicated by sentence or text.



terrible#1 horrific#2 horrendous#1 frightening#1 fearsome#1 fearful#2 dreadful#1 drea

causing fear or dread or terror; "the awful war"; "an awful risk"; "dire news"; "a career or veng presence of the headmaster"; "polio is no longer the dreaded disease it once was"; "a dreadful s city"; "a terrible curse"

P: 0 O: 0.375 N: 0.625

unspeakable#2 **terrible**#2 painful#3 dreadful#2 awful#1 atrocious#2 abominable#2

exceptionally bad or displeasing; "atrocious taste"; "abominable workmanship"; "an awful voice' handwriting"; "an unspeakable odor came sweeping into the room"

P: 0 O: 0.125 N: 0.875

Figure 2.1: An example of Sentiment association in SentiWordNet [1]

SentiWordNet

SentiWordNet 3.0 [12] is improved lexical resource. SentiWordNet 3.0 is enhancement of SentiWordNet 1.0 by Esuli and Sebastiani, 2006. It provides lable in three numerical class Positive, Neutral and Negative for each word present in WordNet synset [14]. SentiWordNet provide synset based sentiment score for each term. In this ach term have different sentiment score in diffrent sense. Figure 2.1 show the example of SentiWordNet, How to diffrent sentiment score for the same term in different sense i.e term terrible have two diffrent sentiment score. SentiWordNet finds the most appropriate meaning of a sentence with Word Sense Disambiguation (WSD) algorithm.

SenticNet

SenticNet [11] is publicly available for sentiment analysis and opinion mining. SenticNet used for concept-level opining mining or sentiment analysis of text. Noval work of SentiNet is , It is multi-disciplinary paradigm for opinion mining and Sentiment Anaylsis. SenticNet label the sentiment score between -1 to 1 (-1 = most negative, 1=Most positive). SenticNet having 14000 common sense concepts. SenticNet able to assign polarity ,additional information and also provide some critical cocepts like celebrate special occasion , accomplishing goal and so on. Sentiment communicated by intensity of 16 basic emotion called Hourglass of Emotions.

WordNet-affect

WordNet-affect [13] is a linguistic resource for lexical representation of affective knowledge. In WordNet-affect , affect is lexical database with approx 1903 tuples

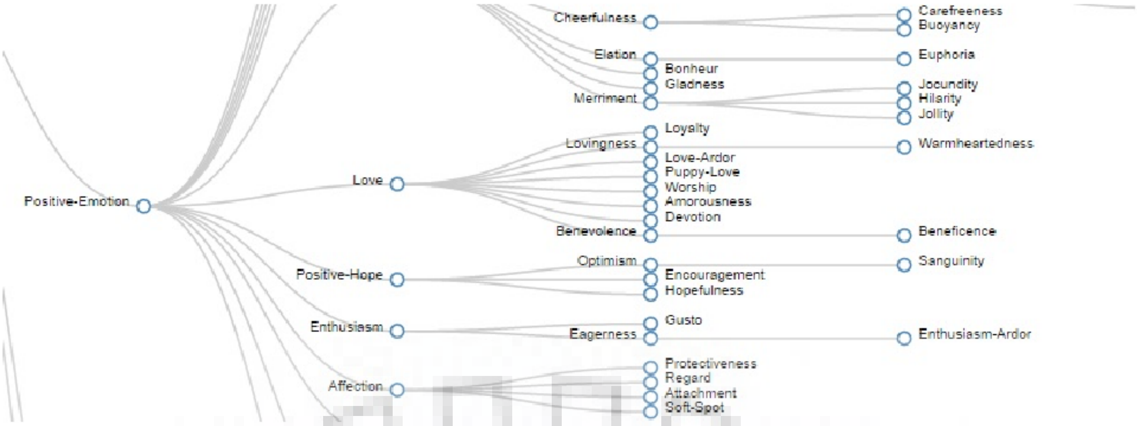


Figure 2.2: A WordNet-affect Hierarchy [1]

directly or indirectly reffering to mental (emotions) states. Figure 2.2 shows emotion hierarchy of WordNet-affect. Also information include relation between POS , English word and italian words. Currently WordNet-affects having 2,874 synset and 4,787 words.

Text data in multillingual due to dataset obtain from different region [15] [16]. To resolve the problem of multilingual, Lexicon resource created by some researchers in that language like sentiLex. Lexicon Based approach is very Easy and intuitive approach.

2.2 Naïve Bayes Classifier

The classification problem is a standard problem. In order to solve the problem, many state-of-art techniques are available. Text classification is one of the special problems. In text classification technique wide variety of task addressed such as assigning a category, topic, genres, spam detection, authorship identification, sentiment analysis or opinion mining. By defination, give a document and a set of classes, to assign the document to one of the set. Some of the most commonly used classifiers are Naïve Bayes classifier , SVM, k-NN [17].

Naïve Bayes classification is based on Bayes rule. It relies on very simple reprentation of the document known popuparly as the ”bag of words” reprentation. the bag of words model losses all information regarding the order of the words in the document. It is represented as a vector of word along with its associated count [18].

It represents a function that returns whether the class is positive or negative with respect to sentiment analysis. In this context, the entire collection of words or a subset of words can be used.

Given a corpus or document d and the distribution of classes c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{2.1}$$

C_m is the mapping of the document to a particular class

$$\begin{aligned} C_m &= argumax_{c \in C} P(d|c)P(c) \\ &= argmax_{c \in C} P(x_1, x_2 ....x_n|c)P(c) \end{aligned} \tag{2.2}$$

where $x_1, x_2 ....x_n$ is random variable representing the features(vocabulary) for the document. Independent assumption for multinomial Naïve Bayes is as stated above which is dependent on some of the simplifying assumption made such as the "bag of words" assumption and conditional independence assumption.

In conditional independence assumption, it is assumed that the feature probabilities P(x_i, c_j) are independent given a class c.

$$P(x_1, x_2 ....x_n|c) = P(x_1|c).P(x_2|c)......P(x_n|c) \tag{2.3}$$

Thus, the equation (2.3) can be used for obtaining C_m . It is a very fast algorithm, with low storage requirements. Naïve Bayes classifier based on basic bayes rule. Its robust to irrelevant features, works very well in domains having many eually important features. If the independence assumption holds then Navïe Bayes is the optimal classifier, but this assumption rarely holds. Also, Naïve Bayes is a "High bias" classifier that works well with small amounts of data.

2.3 Long Short-Term Memory (LSTM)

LSTM network is improved version of Recurrent Neural Network (RNN) that can learn long term dependencies over data with time [19] [2]. It gives better result over RNN. Recurrent neural network could not relate the previous frame to the present frame, to address the problem of recurrent neural network attached the module long-short-term memory with RNN.

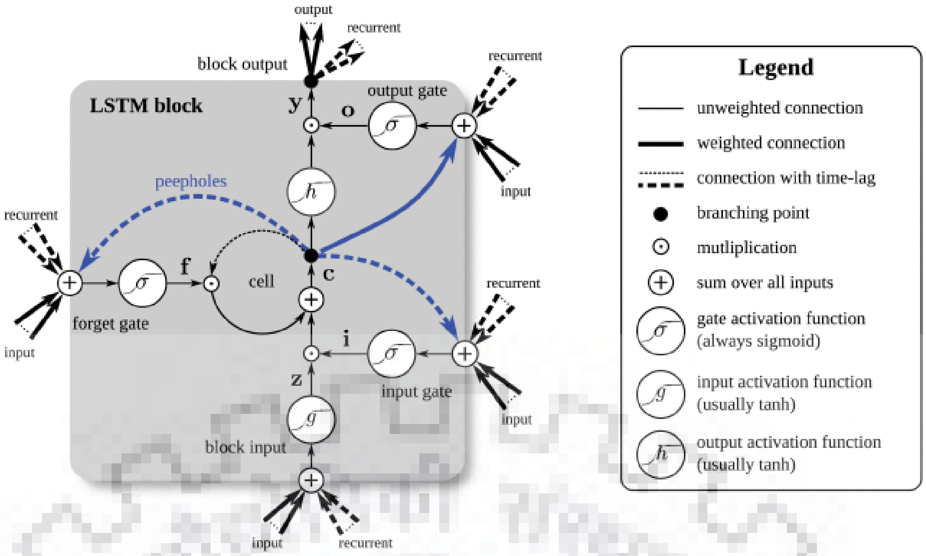


Figure 2.3: LSTM network with hidden layer of recurrent neural network(RNN) [2]

In figure 2.3 shown architecture of long-short term memory block using a recurrent neural network as a hidden layer. It also has a chain block structure like a recurrent neural network with memory block called cell. It also has a running structure like a horizontal line with cell state. It is an ability to add or remove information in cell state at any time. It has three gates, to control the cell state and protect the cell state of LSTM block as given below.

- Forget Gate
- Input Gate
- Output Gate

Forget gate layer also called sigmoid layer which is responsible for what information going throw away from cell state. ”Input gate layer” has two parts first decied which value will be update next and retain and second is create vector values for new comming term in model. ”Output gate layer” finally decide what parts of memory cell going to output.

2.4 Sentiment Analysis

Sentiment analysis is problem related to classification. It is done at different level of coarseness, namely document level, sentence level and word level as described by

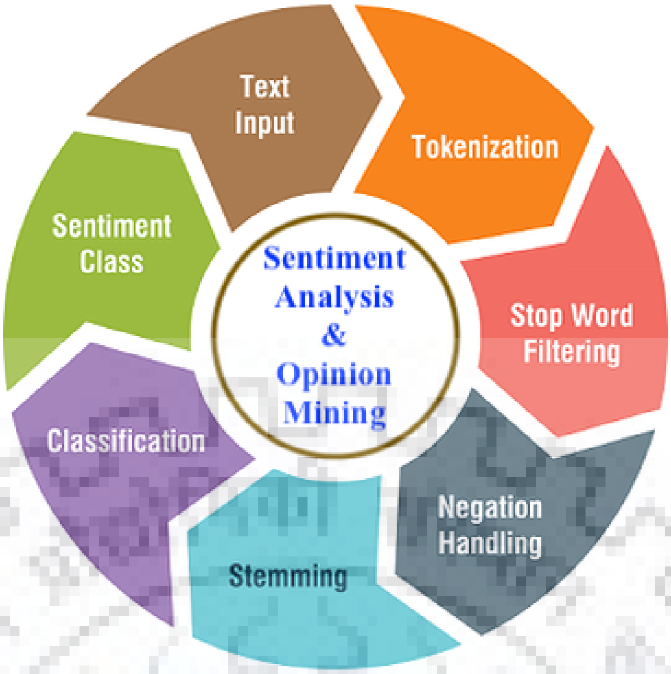


Figure 2.4: Component of Sentiment Analysis

Zhao et al. [9], With the decrease in the granularity of the sentiment analysis the accuracy of the model decrease.

In classification problem deep neural network perform better [20] than lexicon based sentiment analysis. Deep neural networks like Convolutional Neural Network (CNN), Recurrent neural network(RNN), Sequence Classification with LSTM Recurrent Neural Networks shown most promising result in sequence classification task [21]. Alexis et al. [22] has described a deep convolutional network for text classification that takes advantage of convolutional layers to train the machine. Alexis et al. show that model performance increase with depth. The author used 29 layers of convolutional layers to improve the performance of convolutional neural network. A deep convolutional neural network approach for sentiment analysis takes advantage of pre-trained word embedding. Improve the accuracy of sentiment classifier by first identifying the aspect targets in sentences and classifying the input dataset into different classes based on the number of these aspect targets.

2.5 Research Gaps

There has been lot of research works based on the diffferent types of word embedding. Each word embedding has an impact on the performance of deep learning neural network model. There are several complementry word embedding that capture different asspect of the meaning from sentence. In previous work have used only single word embedding during model training. Combination of word embedding has not yet been used to further improvement of sentiment classification model. Thus, there is a great scope to develop Hybrid word embedding. It can be used to improvement of deep learning neural network model.

Chapter 3

PROPOSED WORK

In figure 3.1 shows the flow Diagram of proposed work. Initially, Data collection is a challenging task for any product or target. In this thesis, dataset collected from Amazon shopping site and use standard movie review dataset for evaluating our deep learning model. Each of these data set preprocessed and individually used to train the independent classifier. Classification model labeled the review in positive class or negative class. In the last step use of topic modeling for knowledge extraction from a positive and negative review.

3.1 Dataset Collection

For evaluation of the model, dataset is the most important factor. We create Amzon web crawler for collection of the product reviews. We crawl the category wise product reviews one is amazon food product reviews and health product reviews and create the dataset. In this thesis work we used Amazon product review dataset and movie review dataset. Amazon review also have rating of the review. Rating is between 1 to 5 show the polarity of review or mood reviewer (1- strong neagtive and 5- strong positive). In this url (Product review page url) requester send the query to the Amazon web server and fetched the HTML web page. HTML parser fetch reviews from review section of HTML page. Data received from parser store in review database. Data stored in database send to the next section for preprocessing.

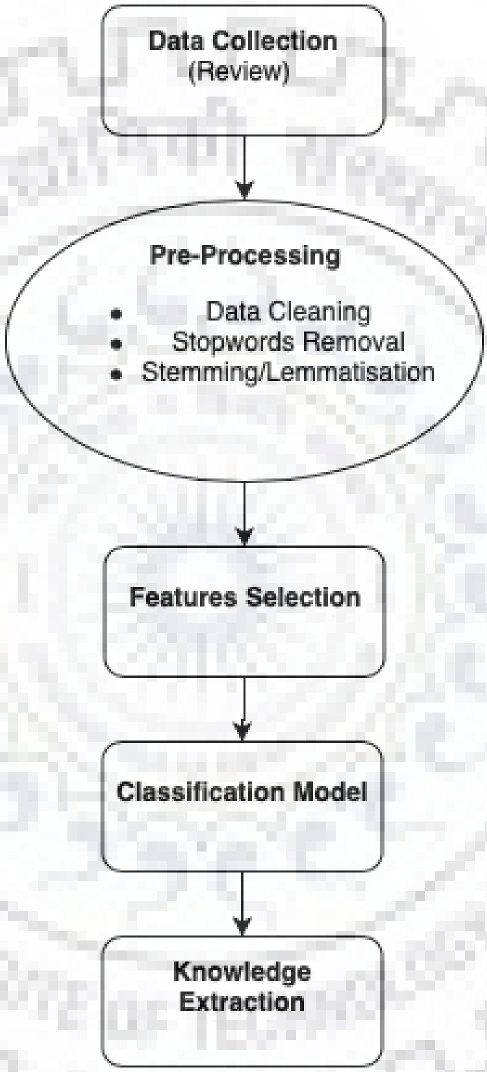


Figure 3.1: Flow Diagram of Proposed Work

3.2 Pre-processing

Pre-processing step is required for efficient use classification algorithm. Pre-processing has following steps

- Noise Removal
- Text Normalization
- Word standardization

Noise Removal

In noise removal step, remove unwanted characters, numbers, and symbols from text. In classification problem stop words also has as a noise. Urls and html tags are remvoed from text data because urls and htmls tags have not much important information for classification problems. After noise removal step text is much more clean.

Text Normalization

In text normalization process, stemming and lematization used for normalization of text. One or more similar word represented by one word. Similars words like 'likes', 'like', 'liked' ,'likely' all are stemmed in root word 'like'. Stemming technique also suffer from two problems overstemming and under stemming.

Word Standardization

In word standardization step text represent in a standard form like all text in lower case or upper case. Convert slang words in standard form and send to the next section.

3.3 Features Selection

In this section cleaned text prepare for classification model. Requirment of vector of text data for Classifiction model. Many algorithm available for features selection like Bow, Tf-Idf , Word2Vec etc. In this report, we have used word2vec model and pre-trained word embedding for LSTM network which is used for classification.

3.3.1 Word Embedding

Word embedding, each word represented in a multi-dimension vector. Word embedding for individual words is used to direct the neural network that performs sentiment analysis. This notation provides for representing words with a limited size of the vector of real numbers. There are several methods of arriving at the position for each word. Some of the method to arrive at embedding are explained in further part of the report.

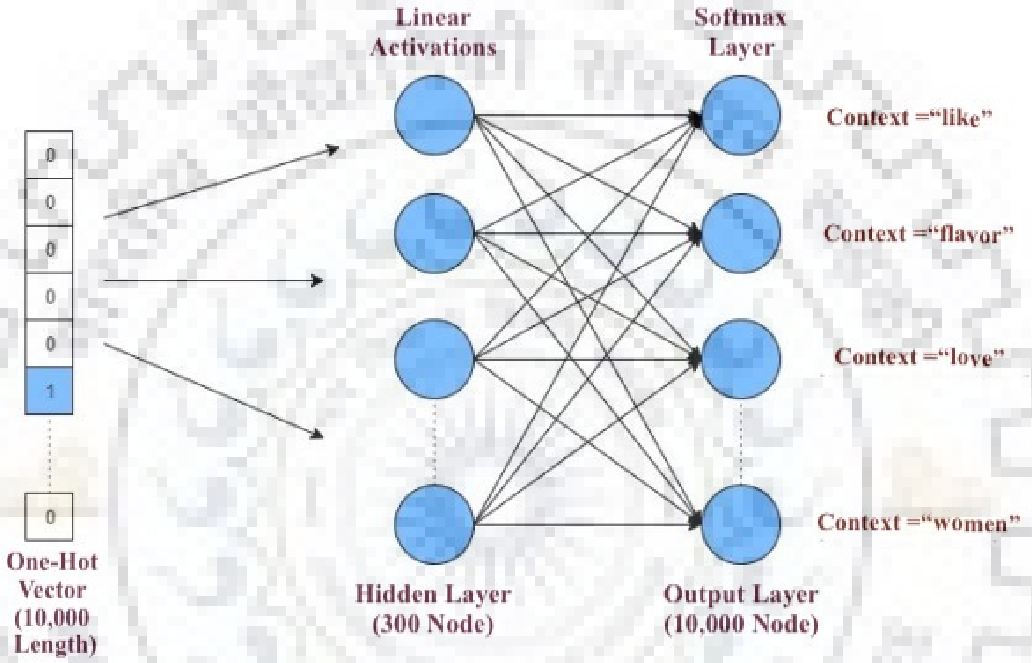


Figure 3.2: Word2Vec Architecture [24]

Word2vec Model

Word2Vec model is produced word embedding of corpus [23] [24]. Word2vec model is patend by Google Inc. It was researched by Google researcher T.Mikolov et al. [24]. Word2vec model have two layer linear activation layer and softmax layer shown in figure 3.2. It takes input as large document and generate vector for each word in document, nearly in hundereds dimensions. Word2vec vectors generated by skip-gram model. Word2vec vector represented in vector space s.t. similar context word pointed in close propinquity to each other in vector space.

Skip-gram model is to generate word notation such that are useful to learn conext of word. A skip-gram model for training the word vector is shown in figure 3.4.

Objective of this model is to minimize the following cost function:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}(Log\ P(w_{t+j}|w_t)) \tag{3.1}$$

In equation 3.1 T is the total size of the vocabulary, c is the window size, t is the current word.

Basic skip-gram function $P(w_{t+j}|w_t)$ define as (softmax function) :

$$P(w_{t+j}|w_t) = \frac{exp(v'_{w0}T_{v_{w1}})}{\sum_{w=1}^{W} exp(v'_{w0}T_{v_{w1}})} \tag{3.2}$$

In equation 3.2 v_w is "input" vector representation and v'_w is "output" vector representation of w. W is the vocabulary size. This equation is impractical because cost computing is propotional to W and W is too much large. The model that uses this embedding shall henceforth be referred to as word2vec embedded(W2Vec) classifier.

3.3.2 Proposed Classification Models

Glove have pre-trained word embedding to use in training of neural network model [6]. Glove embedding vectors involves the generation of a co-occurence matrix(X) for each pair of words. In figure 3.3 vector space reprentation of comparative and superlative words through glove words representation. The model that uses this embedding shall henceforth be referred to as glove embedded (GvE) classifier. For Classification, Proposed model uses Google Based embedding shall henceforth be referred to as Google Embedding (GooE) Model. One more Model proposed for classification, which used multiple embeddings. The classifier used multiple embeddings shall henceforth be referred to as Hybrid Word Embedding (HWE) Model.

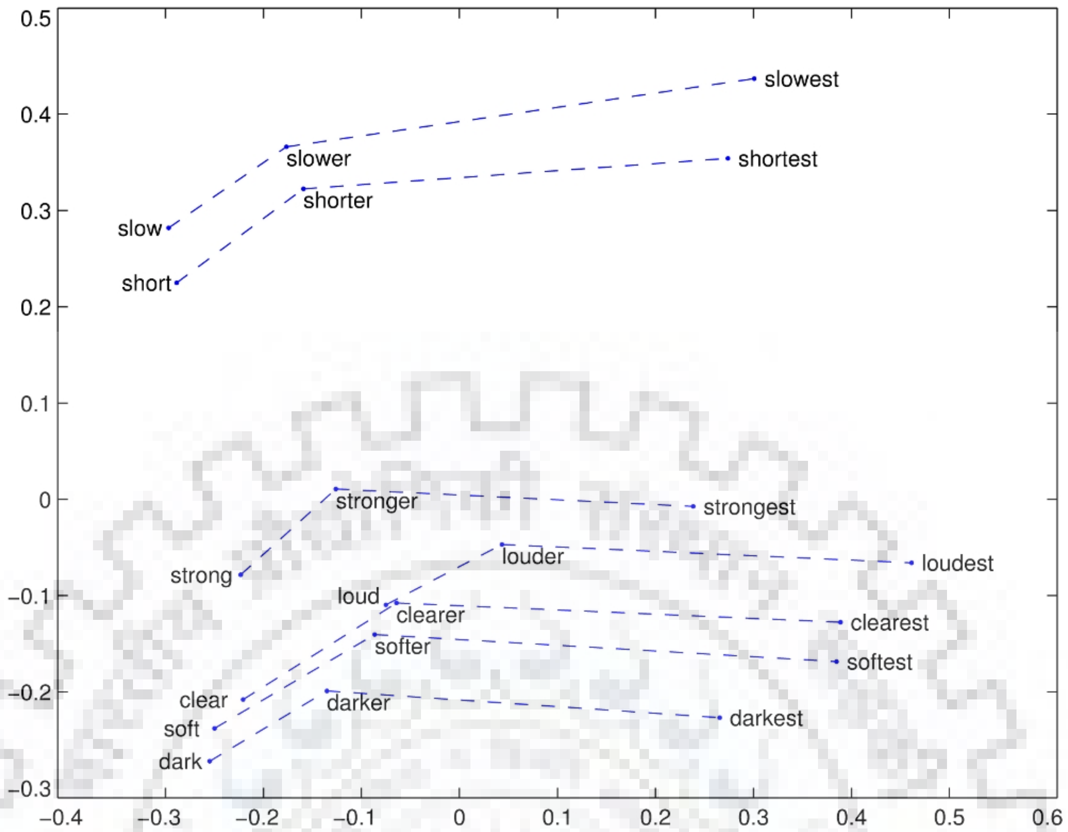


Figure 3.3: Comparative/Superlative words represented in Glove Embedding

3.4 Sequence Classification with LSTM

In the previous section described the LSTM network, which is used for the classification problem. LSTM network is improved version of RNN, In which memory unit included called "cell". It is learn privious context and connect with present situtaion.

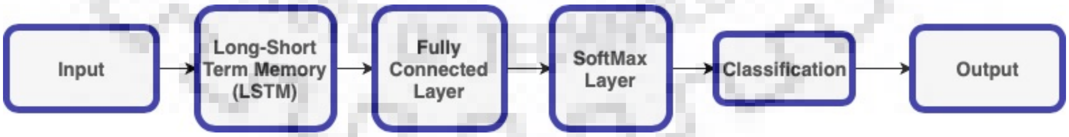


Figure 3.4: Simple LSTM network for Classification

In figure 3.4 Shows how LSTM network used for solve to classification problem. LSTM use three gate Input Gate, Output Gate and forgate gate. Forget gate layer also called sigmoid layer which is responsible for what information going throw away from cell state. "Input gate layer" has two parts first decied which value will be update next and retain and second is create vector values for new comming term in model. "Output gate layer" finally decide what parts of memory cell going to output.

3.5 Latent Dirichlet allocation (LDA)

Latent dirichlet allocation is a generative probabilistic model for collection of discreate data [18]. Documents are represented as a random mixture over latent topics. It is a hierarchical three level Bayesian model, in which each document is modelled as a finite mixture over an underlying set of topics. Each topic in turn is modelled as an infinte mixture over an underlying set of topic probabilities. It gives the most probable topic of the document.

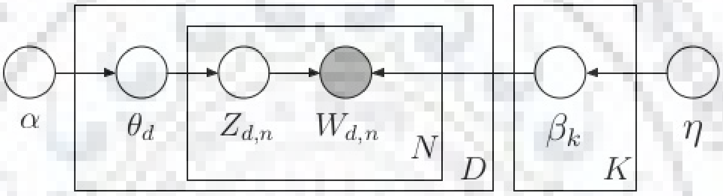


Figure 3.5: LDA as a graphical probabilistic model

In figure 3.5 shown graphical probabilistic model of LDA. In model , "D" denote total number of document , "N" denote total number of words in document ,"K" denote total number of topics, θ_d denote per-document topic proportions, $Z_{d,n}$ denote per-word topic assignment, $W_{d,n}$ denote observed word in document, β_k denote topic distribution over vocabulary and α , η are dirichlet parameters.

Chapter 4

RESULT AND DISCUSSION

4.1 Dataset Description

In this work we used food review dataset, health product review, Election tweet dataset and movie review dataset for experiment on proposed model.

4.1.1 Food Review Dataset

Food review dataset consists of 586,454 tuples, each having 10 attributes. We takes only 3 attributes "Review header" , "Rating" and "Review Text" for classification and knowledge extraction process. Table 4.1 Shows overview of dataset.

like
coffee
taste
good
great
just
can
one
love
product
flavor
tea
really
much
will
amazon
best
make
also
tried
dont
find
recommend
chocolate
food
price
cup
buy
use
time
well
little
sugar
water
drink
get
now
made
better

Figure 4.1: Word Cloud : Food review dataset

Table 4.1: Food Product Review Dataset Overview

	Food Product Review
Size of Dataset(Number of Reviews)	586,454
Attributes	10
Rating	Between 1 to 5

Figure 4.1 shows word cloud of food review dataset. Score is between 1 to 5 (1-strong negative and 5-strong positive) , we aggregate review in two class score 1,2 as negative and 4,5 is as positive review. After removing redundant review we have 307,061 positive review and 57,110 negative tweet. Figure 4.2 shown bar chart of food review dataset.

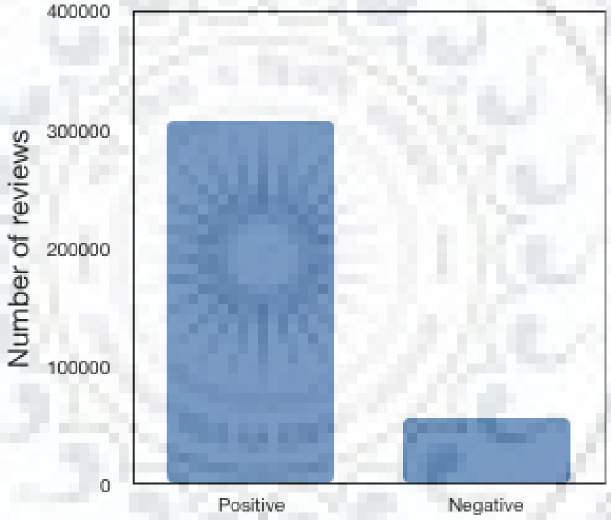


Figure 4.2: Sentiment ditribution in food review dataset

4.1.2 Movie Review Dataset

Movie review dataset have 10,136 reviews. In this 5,094 are positive and rest 5042 are negative reviews. Initally all these reviews are pre-processed to remove punctuations. The number of unique word in the vocabulary is 18,757 after pre-processing the reviews. The maximum length of any sentence is 56. Movie review data have english review about films. Table 4.2 shows Movie review dataset statics.

Table 4.2: Movie Review Dataset Overview

	Movie Review Dataset
Size of Dataset(Number of Reviews)	10,136
Attributes	4
Sentiment	Positive or Negative

4.1.3 Gujarat Election Tweets

During the Gujarat assembly election, many people talk about political party and current problems going on. We crawl the tweet related to assembly election.

Table 4.3: Gujarat Assembly Election Tweets Dataset overview

	Election Tweet DataSet
Size of Dataset (Number of Tweets)	821,251
Attributes	5
Geo-Location	India

This dataset consists of total 821,257 tweet regarding political parties and political leaders. This dataset collected in mainly december 2017 during assembly election 2017. Table 4.3 Showing tweet distribution in Gujarat assembly election dataset. Each these tweet have assign positive , neagtive and neutral.

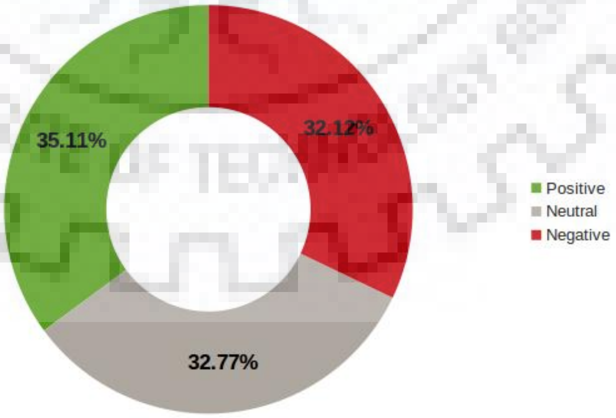


Figure 4.3: Sentiment Distribution of Election Dataset

Figure 4.3 Shows the sentiment distribution of tweet. Those tweet are labelled positive that indicate the opinion regarding the election or political party spoken

about is positive. Tweets showing negative opinion regarding topic labelled as negative. Some tweets are showing neutral opinion about the topic labelled as neutral.

4.1.4 Health Product Related Review

We have also test our model on health product related reviews. Health product related review have 15167 review ragrading health devices like thermometer, Nebulizer, etc and 10678 review regarding sports nutritions product. In health related review score is also one attribute attached to each review. Score is between one to five for each review, one is stand for strong negative opinion and 5 is stand for strong positive opinion for given product. Table 4.4 shows Health product related review dataset statics.

Table 4.4: Health Product Related Review Dataset Overview

	Health Product Related Review Dataset
Size of Dataset(Number of Reviews)	25,845
Attributes	6
Rating	1 to 5

4.2 Results

4.2.1 Sentiment Classification result

The performance of various model on the various dataset are shown below in table 4.5 :

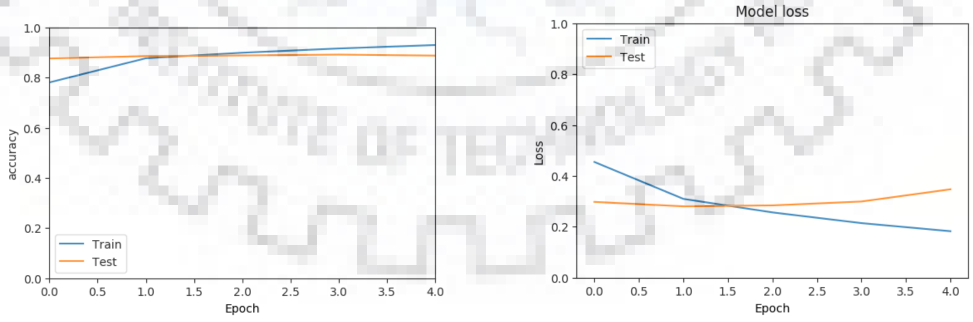
Each model is trained to run up to 10 epochs with early stopping and the model having the best test accuracy is used to evaluate the model. It is observed that for most dataset the best accuracy on either Hybrid Word Embedding (HWE) model or Glove Embedding (GvE) model. As shown in table majority of the proposed models outperform the state-of-the-art models by an acceptable margin.

Table 4.5: Experiment Tabel: Classification Accuracy of the Model (Accuracy in Percentage(%)) PR= Product Review

Model	Embedding Size	Food PR Dataset	Moive Review Dataset	Health PR Dataset	Election Dataset
Existing Approaches					
Naive Bayes	-	86.88	85.45	86.78	78.56
k-NN	-	79.33	78.23	79.65	75.52
LSTM	300	86.65	86.21	85.89	82.67
Proposed Approaches					
GvE model	300	87.9	**88.96**	88.23	83.56
GooE model	300	87.1	86.45	87.82	82.67
HWE model	300	**89.23**	88.57	**88.67**	**84.56**

4.2.2 Effect of Embedding Size on Accuracy

Model accuracy and time and space also depend on the embedding size. Large dimensonality word embedding take more time and space and give more accurate output. Accuracy of the model increase as increase the size of word embedding. After optimal embedding size 300, any further increase in dimensions results in deteriorating performance of the models.



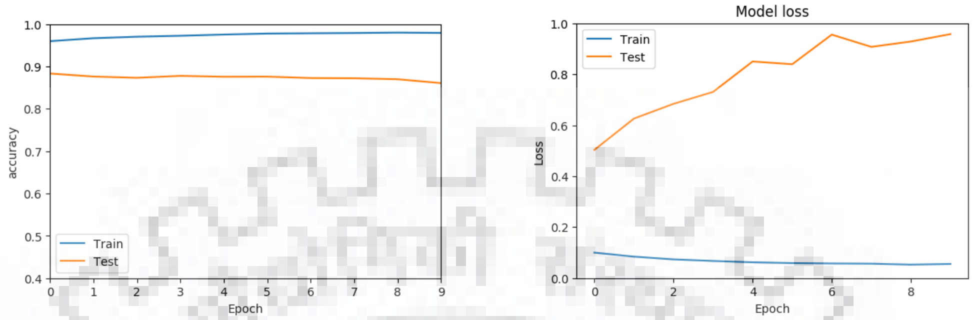
(a) Accuracy Curve

(b) Loss Curve

Figure 4.4: Accuracy and loss curve of HWE model

Figure 4.4(a) the accuracy of the proposed mode Hybrid word embedding (HWE) on the train data and test data at periodic intervals is recorded on food data set.

Figure 4.4(b) the loss of the proposed mode Hybrid word embedding (HWE) on the train data and test data at periodic intervals is recorded on food data set. It is traind on the 300 dimensions word embedding vectors. It gives better result from other dimensions vector and other model.



(a) Accuracy Curve

(b) Loss Curve

Figure 4.5: Accuracy and Loss curve of GvE model

Figure 4.5(a) the accuracy of the proposed mode Hybrid word embedding (HWE) on the train data and test data at periodic intervals is recorded on food data set.

Figure 4.5(b) the loss of the proposed mode Hybrid word embedding (HWE) on the train data and test data at periodic intervals is recorded on food data set.

Chapter 5

CONCLUSION & FUTURE SCOPE

In this work, we have proposed a model to optimze the peoples sentiment classification. In text classification, preprocessing of input dataset is a vital part inorder to improve the accuracy of classification model. We have removed noise from the datasets, and the text normalization is performed to group similar types of words, and further slang words have also standardized to achieve clean text. Our work shows that the accuracy of classification model greatly depends on word embeddings applied. Our proposed model applied Hybrid Embedding which is the compilation of Glove and Google news Embedding. Further the accuracy of the proposed model has been compared with the state-of-the-art classification techniques like Naive bayes, k-NN and LSTM. Proposed model is trained on the four dataset Food Review dataset, Health Product Review dataset, Election tweet dataset and Movie Review dataset. The results of work show that our proposed model improves the sentiment classification accuracy when compared with the techniques like k-NN, Naive Bayes.

Large word embedding dimensions increase the time and space complexity of the model therefore, dimensionality reduction on embedding vector can further improve the model in terms of time and space with some margin. Thus, the proposed model given a new base system which can be further improved and refine.

Bibliography

[1] C. Musto, G. Semeraro, and M. Polignano, “A comparison of lexicon-based approaches for sentiment analysis of microblog posts,” *Information Filtering and Retrieval*, vol. 59, 2014.

[2] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[3] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, “Interpreting the public sentiment variations on twitter,” *IEEE transactions on knowledge and data engineering*, vol. 26, no. 5, pp. 1158–1170, 2013.

[4] J. Li and M. Sun, “Experimental study on sentiment classification of chinese review using machine learning techniques,” in *2007 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2007, pp. 393–400.

[5] Z. Hao, R. Cai, Y. Yang, W. Wen, and L. Liang, “A dynamic conditional random field based framework for sentence-level sentiment analysis of chinese microblog,” in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 1. IEEE, 2017, pp. 135–142.

[6] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[7] F. Morin and Y. Bengio, “Hierarchical probabilistic neural network language model.” in *Aistats*, vol. 5. Citeseer, 2005, pp. 246–252.

[8] C. Korson, “Political agency and citizen journalism: Twitter as a tool of evaluation,” *The Professional Geographer*, vol. 67, no. 3, pp. 364–373, 2015.

[9] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, “Comparing twitter and traditional media using topic models,” in *European conference on information retrieval.* Springer, 2011, pp. 338–349.

[10] T. Schmidt and M. Burghardt, “An evaluation of lexicon-based sentiment analysis techniques for the plays of gotthold ephraim lessing,” in *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2018, pp. 139–149.

[11] E. Cambria, D. Olsher, and D. Rajagopal, “Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis,” in *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[12] S. Haccianella, A. Esuli, and F. S. Sebastiani, “3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

[13] C. Strapparava, A. Valitutti *et al.*, “Wordnet affect: an affective extension of wordnet.” in *Lrec*, vol. 4, no. 1083-1086. Citeseer, 2004, p. 40.

[14] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[15] L. V. Avanço and M. d. G. V. Nunes, “Lexicon-based sentiment analysis for reviews of products in brazilian portuguese,” in *2014 Brazilian Conference on Intelligent Systems.* IEEE, 2014, pp. 277–281.

[16] M. J. Silva, P. Carvalho, and L. Sarmento, “Building a sentiment lexicon for social judgement mining,” in *International Conference on Computational Processing of the Portuguese Language.* Springer, 2012, pp. 218–228.

[17] M. Steyvers and T. Griffiths, “Probabilistic topic models,” *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” 1999.

[20] S. Chen, C. Peng, L. Cai, and L. Guo, “A deep neural network model for target-based sentiment analysis,” in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[21] T. Chen, R. Xu, Y. He, and X. Wang, “Improving sentiment analysis via sentence type classification using bilstm-crf and cnn,” *Expert Systems with Applications*, vol. 72, pp. 221–230, 2017.

[22] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” *arXiv preprint arXiv:1606.01781*, 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

Publications

[1] Nitesh Rai, Durga Toshniwal and Manoj Misra , ”*Analysis of Gujarat Assembly Elections 2017 through Micro-Blog: Twitter*” in 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) : 7th International Workshop on Natural Language Processing for Social Media (SocialNLP@IJCAI 2019), Macao, China, August 2019 . (ERA: A , Qualis: A1) [Accepted]