National College of Ireland

Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning

MSc Research Project
Data Analytics

Somanath S. Chavan
x17108781

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Somanath S. Chavan
Student ID:	x17108781
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	13/08/2018
Project Title:	Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning
Word Count:	6784

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	17th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning

Somanath S. Chavan
x17108781
MSc Research Project in Data Analytics

17th September 2018

Abstract

From the era of globalization every country is cautious about its image among other countrymen. In recent years, India is looking forward for good relationship with the USA. For measuring these relationships Indian government is aggressively trying to find new ways to understand 'how the USA persuading India?'. Newspaper media plays a crucial role in developing a personal view on any topic as people trust more on newspaper media than any other means of media. News headlines are articulated in such a way that it stands for the whole news. By doing the sentiment analysis on news headlines related to India in the USA newspaper can help Indian government to understand the USA sentiments in real time. In this research project Semantic Oriented Approach which is based on SentiWordNet lexicon and machine learning techniques such as Random Forest, Support Vector Machine, Nave Bayes, Long Short-Term Memory, Concurrent Neural Network used for sentiment analysis. Results and findings of these techniques can help Indian government to do real time sentiment analysis on news headlines related to India in the USA newspapers.

1 Introduction

Indian government is spending ample amount of money to project its soft power. Soft power has been defined by Pamment (2014) as a persuasive approach to international relations, typically involving the use of economic or cultural influence. So, the Indian government are actively analyzing how the world media is presenting India to its audiences. In recent years, India is looking for good relationship with the US. Hence for the Indian government agencies it is prevalent to know the US citizens sentiments towards news related to India in the US media. News headlines are articulated in such a way that it stands for the whole news. It is a known fact that media and news plays a crucial role in developing a personal view on any topic. Also, while scanning a news on TV, newspaper or internet, we are first attracted towards headlines only. With sentiment analysis on news headlines of the US newspapers related to India can help Indian government agencies to classify the sentiments related to India. The objective of this research is to compare Semantic Oriented approach i.e. SentiWordNet lexicon-based sentiment classification method with machine learning based sentiment classification techniques for

news headlines of India in US newspapers. Using SentiWordNet, sentiments will be extracted by extracting the explicit information from the text while determining the overall feeling which the author of the article wants to convey. SentiWordNet is based on WordNet which is a sentiment lexicon and takes a semi-supervised approach for building the vocabulary database. This database includes the ability to find the polarity of the word in emotional context. While, supervised machine learning based classification methods will classify the sentiments from the news titles according to models learning ability.

1.1 Motivation and Background

Due to the increase in sheer volume of digital contents such as customer reviews, blogs and news corpora, sentiment classification has received enormous attention from large number of scholars and practitioners. Sentiment classification, also known as sentiment analysis, in printed media domain is a task of judging the opinions (positive or negative) of readers about news (document, sentence, paragraph, etc.) based on computational intelligence using machine learning. Sentiment classification provides stake holder with a tool to transform data into actionable knowledge that decision maker can use in pursuit of improved organizational performance. Sentiment analysis on news has taken a lot of attention from research community for the last two decades especially for the financial domain. In the recent years, getting sentiment analysis on news headlines especially on foreign news media is getting attention my lots of governments agencies of the countries. Dor (2003) with experiment proved relevance theory, the paper makes the claim that headlines are designed to optimize the relevance of their stories for their readers: Headlines provide the readers with the optimal ratio between contextual effect and processing effort, and directs readers to construct the optimal context for interpretation. India has assigned separate budgets to improve its image among foreign nations. A lot of countries such as France, America, UK, Australia, Japan, Germany, China etc. are investing ample amount of money to promote their countrys image among another nationalist. But apart from PEW research finding which ranks country according to its soft power there is no major research done to track this issue. Published news in foreign media for local and foreign reader are reliable source for public to know and built an idea about particular country. So, by doing the sentiment analysis on foreign news media titles for India we can do sentiment analysis in real time. This real time sentiment analysis can help the Indian government agencies to track about how world is persuading theirs country.

1.2 Research Question

RQ: "Can sentiment classification of the USA news headlines related to India using sentiment-oriented-approach (SOA) and machine learning techniques (Random Forest, Support Vector Machine, Naive Bayes, Long Short-Term Memory and Concurrent Neural Network) support/enable/help/assist the Indian government in understanding real time sentiments of India in the USA?"
To address and solve the research question, following objectives are specified, implemented, evaluated and results are presented.

1.3 Research Objectives

Objective1: An investigation of literature on SOA based Sentiment analysis and machine learning based sentiment analysis.

Objective2: Implementation, Evaluation and Results of SOA-based Sentiment classification and machine learning based Sentiment classification of news headlines for India.

Objective2(a): Implementation, Evaluation and Results of SOA-based sentiment classification.

Objective2(b): Implementation, Evaluation and Results of machine learning based Sentiment classification models- Random Forest, Support Vector Machine, Naive Bayes, Long Short-Term Memory and Concurrent Neural Network.

Objective3: Comparison of developed SOA-based sentiment classification model (Objective2(a)) and machine learning based sentiment classification models (Objective2(b)).

The rest of the technical report is structured as follows. Chapter 2 presents literature review of SentiWordNet lexicon based SOA based Sentiment analysis and machine learning classification techniques of news headlines related to India. Chapter 3 introduces and presents the scientific methodology approach used in the project. Chapter 4 presents the implementation, evaluation and results of SOA-based sentiment classification and machine learning based sentiment classification of news headlines related to India in the US newspapers. Finally, chapter 5 concludes and recommends future work.

2 Literature Review on Sentiment Classification of News Headlines (2002 -2018)

2.1 Introduction

This project focuses on a review of sentiment classification of News headlines related to India in the USA newspapers from 2002 to 2018. Sentiment analysis is broadly divided into two categories, first is subjective/objective identification and second is feature/aspect-based sentiment analysis. Hung and Chen (2016) defines Subjective text as the "linguistic expression of somebodys opinions, sentiments, emotions, evaluations, beliefs and speculations". For news headlines, subjective/objective identification-based sentiment analysis fits better. In feature and aspect-based sentiment analysis, aspect of the topic needs to know in beforehand. For example, in restaurant review aspects of the restaurant like service, dish prices etc. are easy to find out. For this project news related to India covering range of topics related to India are considered for the research. So, for analyzing the sentiments, subjective sentiment analysis fits better as author of the news article delivers his/her opinion in the article with the facts related to topic. Best way to find the sentiment about the article is to find the polarity of the headlines. Polarity is nothing, but sentiments categorized as positive sentiment or negative sentiment. Rest of the review work covers sentiment analysis of news headlines related to India in the USA newspapers. SOA based and machine learning based sentiment analysis work is reviewed with their comparison for the research topic. Finally, identified gaps in the research for sentiment

analysis of news headlines related to India in the US newspapers are discussed.

2.2 A Review of Sentiment Analysis on News Headlines

Sentiment Analysis defined by Balahur et al. (2013) is a new discipline which is combination of Information Retrieval and Computational Linguistics and it is not concerned about the topic of the document, but it concerns about opinion it expresses. Sentiment analysis on news headlines for financial domain has studied by many researchers. In financial domain news are categorized according to its type. Most of research have used news headlines only related to finance. But for sentiment analysis on news headlines related to India, news covering different domains such as economy, politics, sports, culture is required to be considered for research purpose. There are two types of methods proposed for sentiment analysis on news headlines, Semantic oriented approach and machine learning techniques (Choi and Lee; 2017). The first approach is semantic oriented approach. In this approach sentiment word lexicon is used for separating the positive sentiment from the negative sentiments. Second approach is machine learning based classification techniques. In machine learning based classification techniques, models are trained on the dataset and the performance of each model is scrutinized on the test dataset.

2.3 A Review of Sentiment Analysis using Semantic Oriented Approach

The SOA is based on identifying and selecting sentiment words in the test documents (Wang et al.; 2014). SOA is a dictionary-based approach for sentiment analysis. In dictionary-based approaches, sentiment analysis is done by using pre-developed dictionaries containing the polarity of words or phrases. Godbole et al. (2007) implemented a lexicon-based system for news and blogs analysis built on top of the Lidia text analysis system. They propose a method to expand candidate seed lists opinion words through WordNet. SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications based on synset in WordNet (Baccianella et al.; 2010). Currently, most frequently used resource is SentiWordNet, which has been employed in number of contemporary researches. Singh et al. (2013) proposed method based on SentiWordNet for review classification. It used linguistic feature consisting of adjective, adverb and verb. They have performed sentiment analysis at document level. Khan et al. (2017) proposed an approach of revised sentiment strength based on SentiWordNet and proved that the method proposed approach is superior than state-of-art techniques. Ohana and Tierney (2009) also used SentiWordNet in a study of classification and concluded that the results provided by SentiWordNet were close to the results obtained with handmade lexicons. Agarwal et al. (2016) used SentiWordNet for sentiment analysis of news headlines by considering every part of speech in the sentence.

2.4 A Review of Sentiment Analysis using Machine Learning Techniques

Pang et al. (2002) pioneered in applying machine learning techniques such as Nave Bayes, Maximum Entropy (ME), and Support Vector Machine (SVM) for binary sentiment classification for movie reviews. For this study dataset of movie reviews from IMDb.com has been used. They experimented with various feature engineering, where SVM yielded the

highest accuracy of 82.9% with unigrams features. Dang et al. (2010) classified sentiments using SVM by using different feature selection methods. Dang, Zhang, and Chen trained SVM with three collections of features set based on domain free, domain dependent, and sentiment features respectively. Information Gain (IG) was applied to reduce the number of features for different combination of features. The reduced features-set performed better on multi-domain datasets. Nassirtoussi et al. (2015) applied SVM, Nave Bayes and KNN with TF-IDF weights for sentiment analysis on news for forex currency market prediction. Hence, SVM performance is better than other two techniques.

Various researchers have contributed to the field of sentiment analysis in different domain. Researchers have used various methodologies and approaches to obtain better results by machine learning model. For example, Nassirtoussi et al. (2015) introduced a novel approach to predict the FOREX prices by using multi-layer model. Where first layer is semantic abstraction layer which is responsible to deal with co-occurrences of words and sparsity within the data followed by sentiment integration layer which is used to find sentiments of the words by incorporating sum score method and finally implementation of dynamic model creation algorithm. This three-layered model has produced accuracy of 83% and eventually outperformed all the previous models. Another experiment conducted by Perikos and Hatzilygeroudis (2016) to predict the emotion from news article with the help of ensemble approach. Ensemble approach is a process of combining two or more models, for this research, Perikos and Hatzilygeroudis has combined two statistical methods that is Naive Bayes and Maximum Entropy learner and one knowledge-based tool. Author has conducted experiment on data collected from news media such as BBC and CNN and obtained accuracy about 86%. Also, there are few other approaches taken by researches with different methodologies, for an instance Yang et al. (2017) has implemented genetic algorithm to predict trading strategy in financial market on a dataset which is combination of tweets and news articles. This approach can help investor to make decision to buy or to sell or to hold the stock. More recently Hui et al. (2017) has conducted a study which helps to automate the process of classifying news headlines into different categories such as happy, sad, angry and amused. Author has implemented two approaches in this research first the sentiment-based category and second the polarity based, both of the approaches are conducted with the help of the KNN algorithm and provides empirical results with f score equals to 0.837 for polarity-based approach and F scores equal to 0.422 for sentiment based approach.

Narayanan et al. (2013) applied Nave Bayes on movie review for sentiment classification. It has achieved more accuracy than previous studies. Moraes et al. (2013) applied SVM, Nave Bayes and ANN for classification. For the research he used both balanced and unbalanced datasets. Also, performance checked with feature reduction methods.

2.5 Comparison of Semantic Oriented Approach and Machine Learning Based Sentiment Classification Techniques

For sentiment analysis of news headlines, many researchers suggested both SOA based and machine learning based methods. For more subjective data, SOA based methods works better than machine learning based methods. Machines learning models works with multi domain data and large datasets. Many literatures suggested that SOA based classification works better for specific domain. Denecke (2009) showed that machine

learning techniques works better for multi domain data as compare to SOA based methods. But accuracy of machine learning techniques is depending upon the size of training dataset. Also, for supervised classification techniques it is required that training data is tagged by domain expert to its best.

These studies suggest that both lexicon and machine learning techniques perform differently with the type of data. So, it is prevalent to test and evaluate both of them.

2.6 Identified Gaps in Sentiment Classification of News Headline Related to India in the USA Newspapers

There has not been much work done in sentiment analysis of news headlines related to India. Sentiment analysis of product review, movie review has been researched to great extend but those text are subjective. Sentiment analysis of news headlines has done but limited to specific domains e.g. financial domain. Most of the news headline datasets used for sentiment analysis are single domain and tagged by domain expert. There has not been any notable work done on sentiment analysis on headlines related to India in the US newspapers. The problem is unique to itself as we are doing to classify the sentiments of newspapers readers by analyzing the newspapers headlines.

2.7 Conclusion

Based on reviewed literature and identified gaps, there is need to develop SOA based and machine learning based sentiment classification on news headlines related to India in the US newspapers. The comparison of developed techniques will provide scholars and practitioners with a practical guidance for the choice of algorithms for a given problem.

3 Scientific Methodology Approach Used and Project Design

3.1 Introduction

This chapter represents scientific methodology used for this project. Developed methodology is based on CRISP-DM methodology (Wirth and Hipp; 2000). CRISDM methodology has been adopted in many researches of data mining.

3.2 Modified Methodology Approach Used

For this project modified CRISP-DM (Cross-industry Standard Process for Data Mining) methodology is adopted. While doing the research paper modified scientific methodology is referred. Fig. 1 describe modified CRISP-DM methodology for sentiment analysis of news headlines related to India in the US newspapers.

- Project Understanding: In this phase research project understanding has been developed for sentiment analysis of news headlines related to India in the US newspapers.
- Data Creation: In this phase rightful dataset is created from corpus.

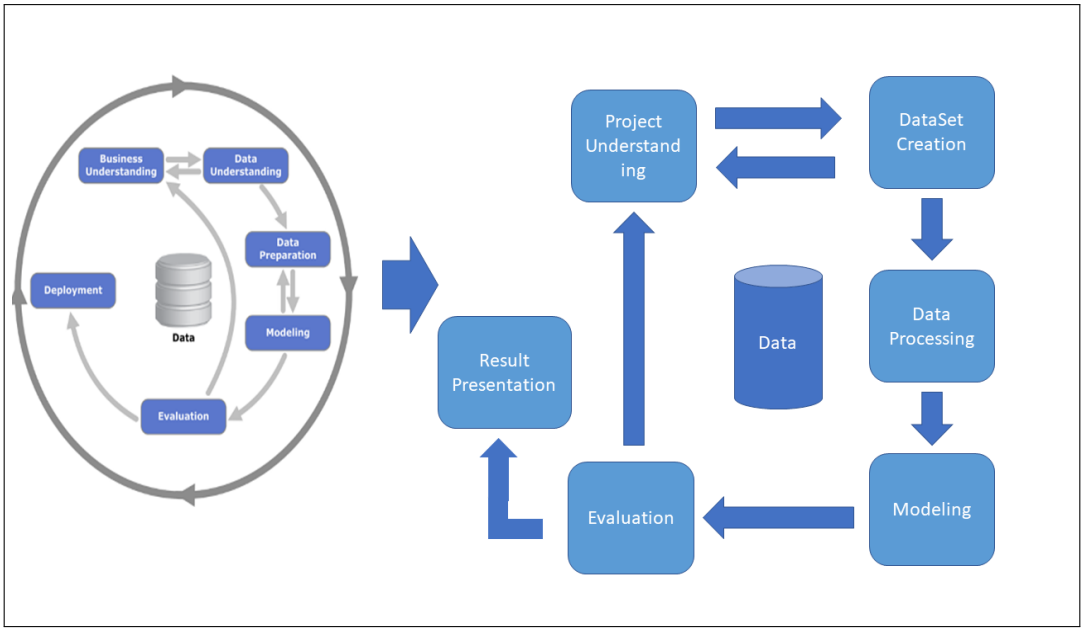

Figure 1: Modified Scientific Methodology Used

- Data Processing: In this phase data is cleaned with NLP techniques. Data prepared for next step.
- Modelling: In this phase supervised classification techniques viz Random Forest, SVM, Nave Bayes, LSTM, RNN are used. Also, SentiWordNet lexicon-based method is developed and used for sentiment analysis on created dataset.
- Evaluation: In this phase performance of developed method are evaluated with different metrices i.e. accuracy, precision, recall, F1 Score, AUC score.
- Result Presentation: Finally, results are presented in grid view for more understanding.

4 Implementation, Evaluation and Results of Semantic Oriented Approach Based Sentiment Classification and Machine Learning Based Sentiment Classification on News Headlines Related to India in the USA Newspapers

4.1 Introduction

In this chapter, implementation and evaluation of SOA based and machine learning based sentiment analysis of news headlines on India in the US newspaper has been presented. Selection of corpora and building of dataset is discussed in the details. Performance analysis by considering different evaluation metrics are discussed in the details. Finally, this chapter compares performance of developed SOA based method with supervised machine learning based classification techniques.

4.2 Process Flow Diagram

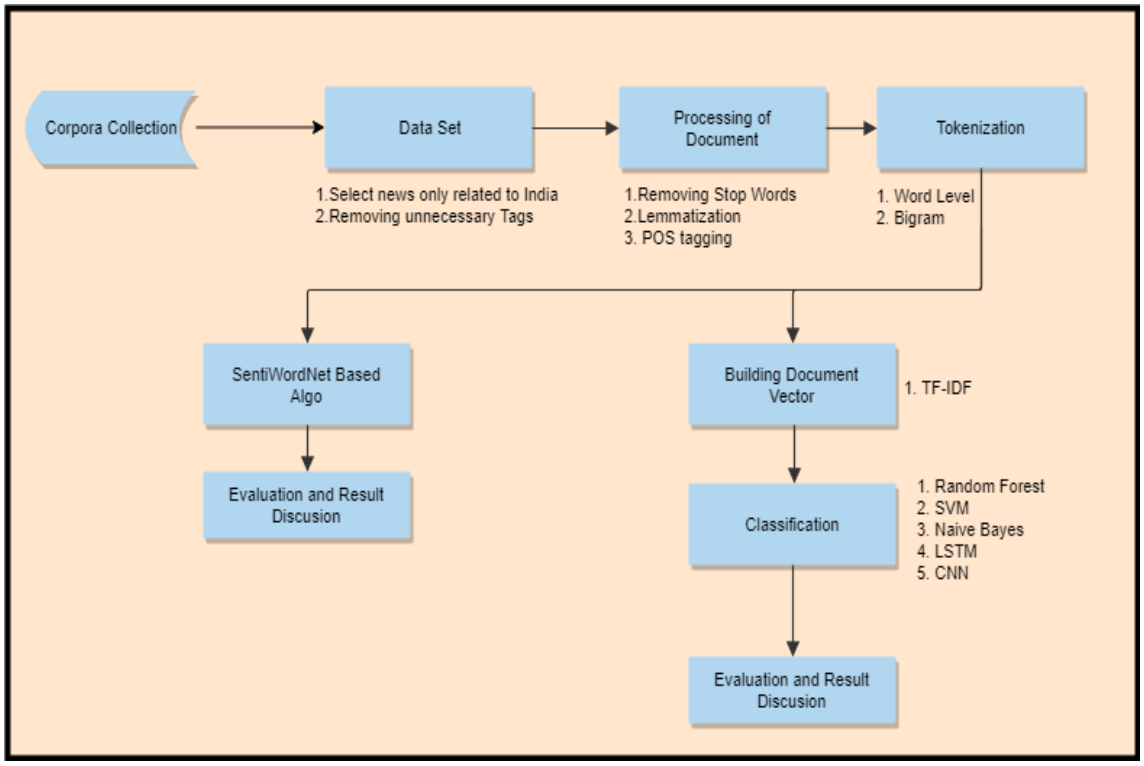

Figure 2: Process flow diagram

The process flow diagram of sentiment analysis of news headlines related to India in US newspaper is shown in (Fig. 2). Process starts with selection of corpora. This corpus is a news headlines of US newspaper. Process is mainly divided into four parts. First, dataset creation from the corpora. Second, Data processing using NLP and tokenization. Third, building document vector and applying SOA and machine learning classification techniques. And in last, performance evaluation and result discussion.

4.3 Creation of Dataset

For creation of data set, news corpus of news headline of the US newspapers with its sentiment score are selected (Moniz and Torgo; 2018). This corpus is available for research on UCI machine learning repository. This corpus contains Date, news headline, source, topic and sentiment score and news article from different news channels. For the research purpose dataset is built by selecting news only related to India. Dataset is created from corpus by searching India word in the headlines of the corpus. Selected rows with the sentiment score are considered as a dataset for this project. Negative sentiment score is labelled as 0 and positive sentiment score is labelled as 1. Data set contains total 1260 headlines, 750 headlines labels with sentiment score as 1 and 510 are labels with 0. With this fact, dataset is considered balanced while processing it for machine learning techniques.

4.4 Data Processing

Given dataset is processed using NLP techniques. For NLP, python NLTK library is used. Stolcke et al. (2000) first printed NLP process in detail. Before the getting the sentiments from the text it is necessary to process the text in such a way that machine understands it. In NLP process these tasks are done with some science. In naturally occurring texts very word is not point of concerns. Dataset is processed as follow:

1. **Removing Unnecessary Tag:** Data is cleaned by removing all unnecessary tags. Dataset is cleaned by removing commas, semicolons, periods, exclamation marks, question marks, intra-word dashes and apostrophes (e.g., "I'd like"). Numbers are removed from the dataset. All extra white spaces are removed. Then remained sentences converted to lower cases.

2. **Lemmatization:**Removing inflectional endings of the words are called lemmatization. Lemmatization process helps to aggregate the word having same meaning but inflectional endings. Dataset is lemmatized.

3. **Parts of Speech Tagging:** Parts of the speech (POS) is a base of English language. English sentences are made up of combination of parts of speech. Tagging each word in sentence with its parts of speech helps in building the features which are necessary. Nouns, adjectives, adverbs mainly influence the tone of speech. Tagging words from text snippet with its POS tag helps in selecting the more accurate features. Lemmatized text is POS tagged and only nouns, adjectives, and adverbs are selected for further processing.

4. **Tokenization:**Tokenization is a crucial step of NLP. In tokenization process every word is separate from all other words and presented as a feature of a document. Manning et al. (2014) in their revised version of The Standford CoreNLP natural language processing toolkit explained that tokenizing the text by only on white space character can miss the meaning of the word. For getting more unambiguous meaning of the word in text snippet, we can divide text into in unigram, bigram or n-gram. Dividing text in grams help to detect the meaning of word in context more accurately. Most of research for long and formally written document suggested to use either unigram or bigram tokenization methods. POS tagged data is tokenized at word level and bigram.

5. **Building Document Vector using TF-IDF for Machine Learning Classifiers:**TF-IDF stands for term frequency-inverse document frequency, and the TF-IDF weight is a weight often used in information retrieval and text mining. TF-IDF weight of a word is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. Word relevance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. TF-IDF term weights are the result of simplified probabilistic retrieval model that simulates human relevance decision-making (Wu et al.; 2008). Python sklearn library is used for building TF-IDF document vector.

4.5 Evaluation Metrics

For the evaluation of performance of developed models Accuracy, Recall, Precision metrics have been used. Developed binary sentiment classifier on test dataset will yield

results in four category, a true positive (TP), a false positive (FP), a true negative (TN) and a false negative (FN). Test data supplied to classifier are already labeled so we know the true label of the instance and can categorize our predictions to one of the mentioned results. A TP is an instance with a positive predicted label and positive actual label, likewise a TN is an instance with a negative predicted label and negative actual label. A FP is an instance with a positive predicted label but is actually labeled negative with the reverse being true for the FN which has a predicted label of negative but has an actual label of positive.

A classifier which performs acceptably by correctly predicting the test instance's label will have a large number of TP + TN counts relative to the overall count, also known as accuracy. Formally defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is the ratio of the retrieved true positives instances that are correctly labeled defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as True Positive Rate (TPR), is the ratio of actual true positives correctly labeled defined as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Performance of every developed algorithms is verified by accuracy, precision and recall metrics

4.6 Implementation, Evaluation and Results of Semantic Oriented Approach-based Sentiment Classification

For implementing SOA-based sentiment analysis SentiWordNet lexicon is used. SentiWordNet is based on English dictionary i.e. WordNet. WordNet is a lexical database for English language. Synsets is a heart of WordNet database. It groups English words in such a way that word is grouped with synonyms of that word with shortest definition and its usage examples. It also records a number of relation among these synonym sets or their members. Thus, WordNet can be used as a combination of dictionary and thesaurus. SentiWordNet assigns each synset (or synonym) in WordNet with three different sentiment polarities- positive, negative, and neutral. Each label has specific value in range of 0 to 1 and sum of three terms is equal to zero.

Implementation: Using SentiWordNet, an algorithm is developed to sentiment analysis. Algorithm is developed using Python 2.0, NLTK package. SentiWordNet 3.0 lexicon is used. Pseudo-code of algorithm is shown in figure 3.

```
For every document in the TestDataSet:
    For each sentence in the document:
        TaggedSentences = POS (sentence)
        For SentiCandiate (adverb, adjective, and verb) in TaggedSentence:
            PolarityScore +=LookupSentiWordNet(SentiCnadidate)
            TotalCandidateCount++
    AveragePolarity = PolarityScore/TotalCandidateCount
    IF(AveragePolarity>0): RETURN 1
    ELSE: RETURN 0
```

Figure 3: SOA based algorithm for sentiment analysis

Evaluation and Result: Overall accuracy achieved by SOA is 66 percent. Precision and recall achieved using SOA are 67.2 and 79.8 respectively.

4.7 Implementation, Evaluation and Results of Machine Learning Sentiment Classification Models- Random Forest, Support Vector Machine, Nave Bayes, and Concurrent Neural Network

This chapter represent implementation, evaluation and result of machine learning sentiment classification models- Random Forest, Support Vector Machine (SVM), Nave Bayes and Concurrent neural network. Dataset is divided into training and testing subsets. 80 percent data is taken for training and 20 percent data is taken for testing. Python 2.0 is used for the implementation.

4.7.1 Implementation, Evaluation and Results of Support Vector Machine Classification Model

Support vector machine (SVM) is a linear learning technique that finds an optimal hyperplane to separate two classes. SVM is a supervised learning technique. SVM seeks to maximize the distance to the closest training point from either class in order to achieve better generalization/classification performance on test data. Classification is based only on those training data points which are at the margin of the decision boundary. These points are called support vectors and are illustrated in Fig. 2(a). Instead of minimizing a global error function in a gradient descent process, which suffers from the existence of multiple local minima solutions, the parameters of the optimal separating hyperplane can be obtained by solving a convex optimization problem, for which there are standard software packages available (Tsytsarau, M. et al 2014).
As shown in fig. 2(b) when classes are not linearly separable, the input data space is transformed into a higher-dimensional feature space in order to make data linearly separable and suitable for the linear SVM formulation. Generally, this transformation is achieved by using a kernal function. It makes possible to determine a nonlinear decision boundary, which is linear in the higher-dimensional feature space, without computing the parameters of the optimal hyperplane in a feature space of possibly high dimensionality. Hence, the solution can be written as a weighted sum of the values of certain kernel function evaluated at the support vectors.

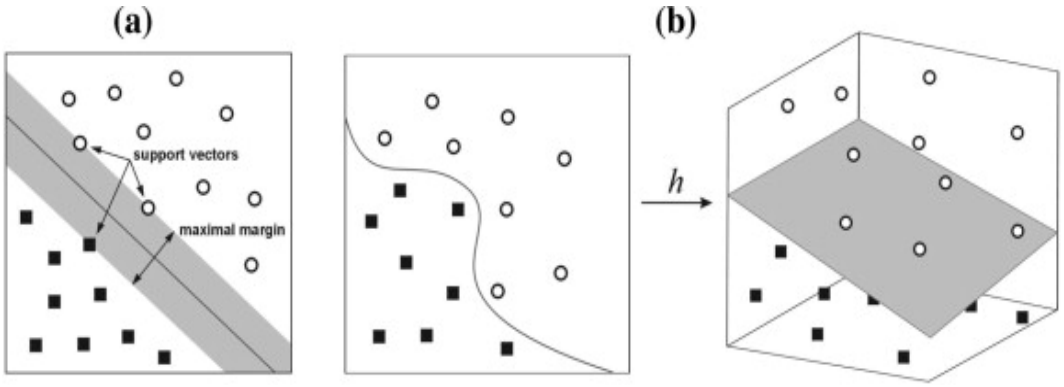

Figure 4: SVM hyper plane transformation

Implementation:SVM is implemented using SVC library of sklean in python. For the implementation linear kernel is used with gamma value as 1e-4. SVM is implemented on both word level tokenization and bigram. Also decomposition which helps in feature reduced is used on word level tokenized data.
Evaluation and Results: Overall accuracy achieved by SVM is 75 percent with bigram. Precision and recall achieved by SVM are 71.5 and 69.4 respectively. F1 score which shows trade off between precision and recall is 70.1 percent for the SVM.

4.7.2 Implementation, Evaluation and Results of Random Forest Classification Model

Random Forests builds on bagging technique. Breiman (2001) define Random Forests is a bagged classifier h_c combining a collection of T classification or regression trees (i.e. forest of trees), here T classification trees. Each tree t is grown on a different bootstrap sample S_t containing N_1 randomly drawn instances with replacement from the original training sample. Besides bagging Random Forests also employs random feature selection. At each node of the decision tree t, m variables are selected at random out of the M input vectors and the best split selected out of these m. Each decision tree t is grown using CART methodology to the largest extent possible. To classify a new instance, put the input vector down the T trees in the forest. Each tree votes for the predicted class. Finally, the bagged predictor is obtained by majority vote, i.e. the instance is classified into the class having the most votes over all T trees in the forest. The two sources of randomness, random inputs and random features, make Random Forests accurate classifiers in different domains (Huang et al.; 2005).

Implementation : For implementation RandomForestClassifier library of sklearn is used. Random Forest is tested for both word level tokenized dataset and bigram tokenized dataset. Random forest is tested for different values of number of tree. For number of trees equal to 2000, random forest has given maximum accuracy.

Evaluation and Results : Over all accuracy achieved by Random Forest is 72.6 percent on word level. Precision and recall achieved by random forest for word level tokenization are 70.9 and 93.8 percent respectively. Overall accuracy achieved by Random forest is 72.6 percent by bigram tokenization. With precision and recall 70.7 and 94.1 percent respectively.

4.7.3 Implementation, Evaluation and Results of Long Short-Term Memory Model

The Long Short Term Memory architecture (Gers et al.; 2002) was motivated by an analysis of error flow in existing RNNs (Hochreiter et al.; 2001), which found that long time lags were inaccessible to existing architectures, because back propagated error either blows up or decays exponentially. An LSTM layer consists of a set of recurrently connected blocks, known as memory blocks. These blocks can be thought of as a differentiable version of the memory chips in a digital computer. Each one contains one or more recurrently connected memory cells and three multiplicative units the input, output and forget gates that provide continuous analogues of write, read and reset operations for the cells. More precisely, the input to the cells is multiplied by the activation of the input gate, the output to the net is multiplied by that of the output gate, and the previous cell values are multiplied by the forget gate. The net can only interact with the cells via the gates.

Implementation: LSTM is implemented using sklearn with recurrent neural network. Model is trained for epochs of 10. With epochs 10 we have achieved 90 percent training accuracy.

Evaluation and results: LSTM achieved overall 71 percent accuracy with 82 percent of loss.

4.7.4 Implementation, Evaluation and Results of Nave Bayes Model

Nave bayes is most practiced machine learning algorithm for sentiment classification. Nave Bayes is easy to implement and skip the any complicated iterative parameter estimation schemes (Wu et al.; 2008). Based on the bag-of-words model, Nave Bayesian based sentiment classification defines the likelihood of a document (d) to be positive or negative as a sum of total probability over all mixture components, i.e., $P(d) = \sum jP(d|positive)P(positive)$ for positive; where P(positive) is the probability of the positive and P(d—positive) is the probability of the document belonging to positive. For balanced training dataset, P(positive) and P(negative) are equal to 0.5 as the equal number of documents are used for positive and negative. To compute the likelihood of being positive or negative for given document, Nave Bayesian approach applies the so-called "nave assumption" that all words are independently used in all document, (Melville et al.2009), which implies that P(w.i).=.P(w.i.—.wj.) where.w.j.can be any other words. Based on this assumption, the probability of a document.d being generated in positive is P(d—.positive).=..i.P(w.i.—.positive), where i is the number of words in a document.

The Nave Bayes classification rule uses Bayes theorem to compute the probabilities of a document belonging to class c.j.as follow

$$P(positive|d) = \frac{P(positive)\Pi ip(wi|positive)}{P(d)} \tag{4}$$

and the label with the highest likelihood is predicted, i.e.,

$$argmaxlabelP(label)\Pi iP(wi|label)argmaxlabelP(label)\Pi iP(wi|label) \tag{5}$$

Implementation: Naive Bayes is implemented for both word level and bigram tokenization. Bigram tokenization is decomposed for feature reduction

Evaluation and results: Overall accurecy achieved by Naive bayes by bigram decomposition is 64.2 percent.

4.7.5 Implementation, Evaluation and Results of Convolution Neural Network Model

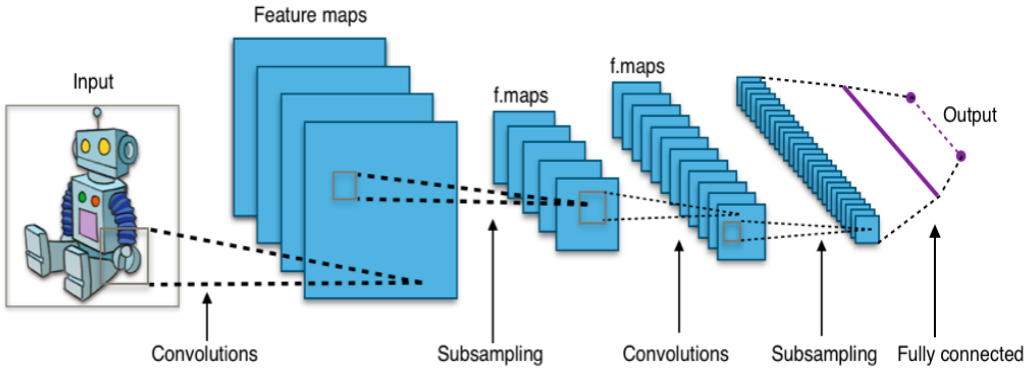

Figure 5: Convolution Neural Network

Kim (2014) applied CNN for sentiment classification. In CNN as shown in Fig 3, consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutions layers, pooling layers, fully connected layers and normalization layers. A process by which one feature is extracted from one filter. The model uses multiple filters (with varying window sizes) to obtain multiple features. These features form the penultimate layer and are passed to a fully connected softmax layer whose output is the probability distribution over labels.

Implementation: For implementation keras library for Python is used. For pooling in CNN global pooling is used. For convolution layer relu activation function is used. For first out payer relu and for second output layer sigmoid activation function is used.

Evaluation and results: Overall accuracy achieved by CNN is 68.7.

4.8 Comparison of developed Semantic Oriented Approach-based sentiment classification model (Objective2 (a)) and machine learning based sentiment classification models (Objective2(b)).

As shown in the table 1, we can notice that SVM accuracy is 76.5 which is highest among all developed algorithm. Naive Bayes has achieved 64.2 percent accuracy which is lowest accuracy among all developed algorithms among machine learning techniques. For sentiment classification accuracy performance matrix is useful for the stake holder. For sentiment analysis both positive and negative sentiment are equally important. We have successfully achieved objective3 of the project.

Algorithm	Accuracy
SOA	66
Random Forest	74.5
SVM	76.5
Naive Bayes	64.2
LSTN	71
CNN	68.7

Table 1: Performance of the algorithms

5 Conclusion and Future Work

In this research project sentiment classification on news headlines related to India in the USA have been successfully implemented using semantic oriented approach and machine learning based techniques. Support vector machine has achieved maximum accuracy of 76.5 percent. SOA based algorithm achieved 66 percent accuracy which is lowest among all the algorithms except Naive Bayes, but which is much equal to contemporary researches. This finding can be very helpful for the Indian government agencies as a reference work for sentiment analysis on news headline related to India in the USA newspapers using semantic oriented approach and machine learning approach. Thus, we have successfully answered the research question and have achieved all the objectives defined in introduction chapter.

Future Work: In the future, one can use bigger corpus and apply machine learning algorithms on it. As machine learning algorithms are proven to get better with more size of data. Also, for this research project news from different domain are considered at one time for SOA based sentiment classification. SOA based classifications works better on single domain data. So, in future one can separate news using topic modelling and then can test SOA based sentiment classification. For the research project news only in the USA newspapers are considered for the experiment. In future one can use news headlines from countries apart from the USA for the experiment.

Acknowledgement: I would specially like to thank my Supervisor Dr. Catherine Mulwa for her continuous guidance and supporting me throughout the research project. Catherine always encouraged me for adding more innovation in the project. I would also like to acknowledge my friend Kunal Khule for his review and comments for the project implementation.

References

Agarwal, A., Sharma, V., Sikka, G. and Dhir, R. (2016). Opinion mining of news headlines using sentiwordnet, *Colossal Data Analysis and Networking (CDAN), Symposium on*, IEEE, pp. 1–5.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., *Lrec*, Vol. 10, pp. 2200–2204.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M.,

Pouliquen, B. and Belyaeva, J. (2013). Sentiment analysis in the news, *arXiv preprint arXiv:1309.6202* .

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Choi, Y. and Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications, *Information Systems Frontiers* **19**(5): 993–1012.

Dang, Y., Zhang, Y. and Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews, *IEEE Intelligent Systems* **25**(4): 46–53.

Denecke, K. (2009). Are sentiwordnet scores suited for multi-domain sentiment classification?, *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, IEEE, pp. 1–6.

Dor, D. (2003). On newspaper headlines as relevance optimizers, *Journal of Pragmatics* **35**(5): 695–721.

Gers, F. A., Schraudolph, N. N. and Schmidhuber, J. (2002). Learning precise timing with lstm recurrent networks, *Journal of machine learning research* **3**(Aug): 115–143.

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs., *Icwsm* **7**(21): 219–222.

Hochreiter, S., Younger, A. S. and Conwell, P. R. (2001). Learning to learn using gradient descent, *International Conference on Artificial Neural Networks*, Springer, pp. 87–94.

Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S. J., Miller, L. W. and Hall, J. (2005). A comparative study of discriminating human heart failure etiology using gene expression profiles, *BMC bioinformatics* **6**(1): 205.

Hui, J. L. O., Hoon, G. K. and Zainon, W. M. N. W. (2017). Effects of word class and text position in sentiment-based news classification, *Procedia Computer Science* **124**: 77–85.

Hung, C. and Chen, S.-J. (2016). Word sense disambiguation based sentiment lexicons for sentiment classification, *Knowledge-Based Systems* **110**: 224–232.

Khan, F. H., Qamar, U. and Bashir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on sentiwordnet, *Knowledge and information Systems* **51**(3): 851–872.

Kim, Y. (2014). Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* .

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014). The stanford corenlp natural language processing toolkit, *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60.

Moniz, N. and Torgo, L. (2018). Multi-source social feedback of online news feeds, *arXiv preprint arXiv:1801.07055* .

Moraes, R., Valiati, J. F. and Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between svm and ann, *Expert Systems with Applications* **40**(2): 621–633.

Narayanan, V., Arora, I. and Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced naive bayes model, *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, pp. 194–201.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. and Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment, *Expert Systems with Applications* **42**(1): 306–324.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet.

Pamment, J. (2014). Articulating influence: Toward a research agenda for interpreting the evaluation of soft power, public diplomacy and nation brands, *Public Relations Review* **40**(1): 50–59.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, pp. 79–86.

Perikos, I. and Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers, *Engineering Applications of Artificial Intelligence* **51**: 191–201.

Singh, V. K., Piryani, R., Uddin, A. and Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification, *Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on*, IEEE, pp. 712–717.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V. and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational linguistics* **26**(3): 339–373.

Wang, H., Yin, P., Zheng, L. and Liu, J. N. (2014). Sentiment classification of online reviews: using sentence-based language model, *Journal of Experimental & Theoretical Artificial Intelligence* **26**(1): 13–31.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, pp. 29–39.

Wu, H. C., Luk, R. W. P., Wong, K. F. and Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions, *ACM Transactions on Information Systems (TOIS)* **26**(3): 13.

Yang, S. Y., Mo, S. Y. K., Liu, A. and Kirilenko, A. A. (2017). Genetic programming optimization for a sentiment feedback strength based trading strategy, *Neurocomputing* **264**: 29–41.