# Lexicon Based Approach for Sentiment Analysis on News Articles Using Deep Learning Techniques

**Gaikwad Supriya Vilasrao**

M.E Student Shreeyash college of Engg & Technology, Department of Computer Science Engineering, Dr. Babasaheb Ambedkar Marathwada University, 431010 Aurangabad, Maharashtra, India

**Abstract:** ***Sentiment Analysis or Opinion Mining is a most popular field to analyze and find out insights from text data from various sources like News Articles, Twitter, etc. The medium of publishing news and events has become faster with the advancement of Information Technology (IT). IT has also been flooded with immense amounts of data, which is being published every minute of every day, by millions of users, in the shape of comments, blogs, news sharing through blogs, social media micro-blogging websites and many more. The medium of publishing news and events has become faster with the advancement of Information Technology (IT). IT has also been flooded with immense amounts of data, which is being published every minute of every day, by millions of users, in the shape of comments, blogs, news sharing through blogs, social media micro-blogging websites and many more. The medium of publishing news and events has become faster with the advancement of Information Technology (IT). IT has also been flooded with immense amounts of data, which is being published every minute of every day, by millions of users, in the shape of comments, blogs, news sharing through blogs, social media micro-blogging websites and many more. Manual traversal of such huge data is a challenging job; thus, sophisticated methods are acquired to perform this task automatically and efficiently. News reports events that comprise of emotions – good, bad, neutral. Sentiment analysis is utilized to investigate human emotions (i.e., sentiments) present in textual information. This paper presents a lexicon-based approach for sentiment analysis of news articles. The experiments have been performed on BBC news dataset, which expresses the applicability and validation of the adopted approach. After preprocessing we applied machine learning algorithms to classify reviews that are positive or negative. This paper concludes that, Deep Learning Techniques gives best results to classify the News Articles Reviews. LDA got accuracy 95.17% and Word 2vec got accuracy 93.54% for Emotional valence Reviews.***

**Keywords:** Sentiment analysis, Natural language processing, Deep learning, LDA algorithm, Word2vec algorithm

## 1. Introduction

With the emergence of the Internet, web and mobile technologies, people have changed their way of consuming news. Traditional physical newspapers and magazines have been replaced by virtual online versions like online news and weblogs. Online news expresses opinions regarding news entities, which may comprise of people, places or even things, while reporting on events that have recently occurred [4]. For this reason interactive emotion rating services are offered by various channels of several news websites, i.e., news can be positive, negative or neutral [5].

Sentiment Analysis or Opinion Mining is a way of finding out the polarity or strength of the opinion (positive or negative) that is expressed in written text, in the case of this paper – a news article [3] [4]. Manual labeling of sentiment words is a time consuming process. There are two popular approaches that are utilized to automate the process of sentiment analysis. The first process makes use of a lexicon of weighted words and the second process is based on approaches of machine learning. Lexicon based methods use a word stock dictionary with opinion words and match given set of words in a text for finding polarity. As opposed to machine learning methods, this approach does not need to preprocess data not does it have to train a classifier [6]. This research is based on a method for Lexicon-based sentiment analysis and Deep learning technique of news articles.

The remainder of this paper is organized as follows: Section II presents related work conducted in sentiment analysis for news articles. Section III presents the proposed methodology and experiment setup of this paper. Results have been presented in Section IV followed by limitations of the research in Section V. Finally, Section VI presents the conclusion of this research.

## 2. Related Work

Many researchers have contributed in news sentiment analysis using different approaches.

Godbole, Srinivasaiah, and Sekine built an algorithm based on sentiment lexicons which could help in finding the sentiment words and entities associated in the text corpus of two trends were analyzed in the experiment - 1) Polarity: sentiment associated with entity is positive or negative and 2) Subjectivity: how much sentiment an entity garners. Score for both polarity and subjectivity were calculated.

Reis, Olmo Benevenuto, Prates and An proposed a methodology to discover the relationship between sentiment polarity and news popularity [3]. Using different sentiment analysis methods, an experiment was conducted by utilizing the content of 69,907 headlines generated by four most reputed media corporations –The New York Times, BBC, Reuters, and Dailymail. Extracting features from text of news headlines, the research analyzed the sentiment polarity of these headlines. The research concluded that the polarity of the headline had a great impact on the popularity of the news article. The research found that negative and positive news headlines gained greater interest than news headlines that had a neutral tone.

**Algorithm 1:** Preprocessing of each word
Select news headline, then pre-process each word in it using POS tagger and perform Lemmatization, and Stemming. This is done using Natural Language Tool Kit (NLTK).

**Algorithm 2:** Analyzing news headlines
After pre-processing pass each word in to SentiWordNet 3.0 dictionary to find positive, negative and objective scores. If positive score > negative score then mark news headline as positive. And if positive score < negative score then mark news headline as negative.

## 3. Research Methodology

The methodology used for sentiment analysis of news articles in this paper is based on the Lexicon-based approach and Deep Learning Technology. Sentiment analysis can generally be carried out using supervised or unsupervised approaches. Unsupervised or Lexicon-based approaches to sentiment analysis do not require any training data. In this approach, the sentiments conveyed by a word are inferred on grounds of the polarity of the word.

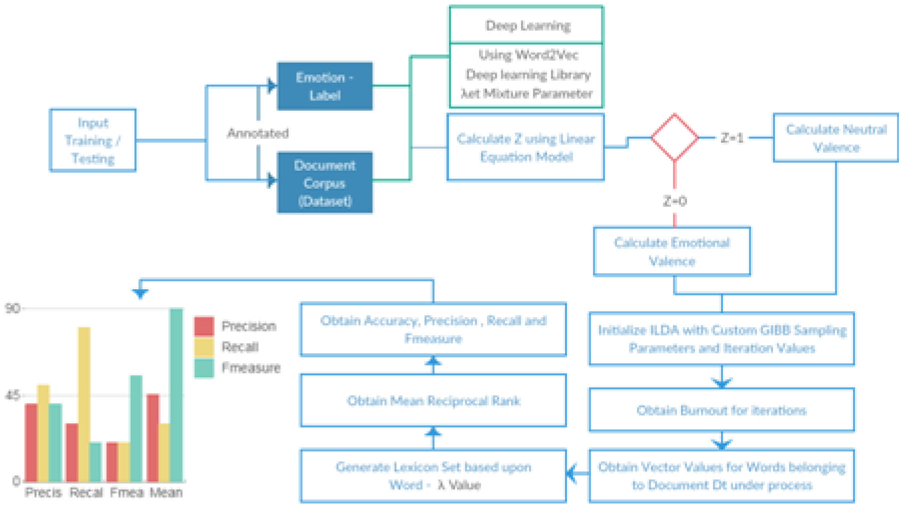
Sentiment analysis can be done on document level, sentence level, word level or phrase level. This paper explores sentiment analysis on the document level. Similar to [13] [14], this research identifies whether the documents new articles expressed opinions are positive, negative or neutral. The dictionary based approach has been used for sentiment analysis of news articles utilizing the wordNet lexical dictionary.

There are 6 emotion categories that are widely used to describe humans’ basic emotions, based on expressions *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness* and *Surprise*. In recognition task, the 3 most common approaches are rule-based, statistic-based and hybrid, and their use depends on factors such as availability of data, domain expertise, and domain specificity.

In the case of sentiment analysis, this task can be tackled using lexicon-based methods, machine learning, or a concept-level approach

**Objectives**
1) To extract features which disentangle the hidden factors of variations
2) To obtain factors like positive reviews to check the disentanglement of the dataset.
3) To consider unlabeled data and labels from a single domain and follow the two-step procedure for sentiment analysis.
4) To use word2vec algorithm that will train the linear classifier on transformed labeled data thereby assisting in emotion detection.
5) To develop a system with emotion dataset and training dataset and to obtain valency in the form of emotional and neutral.
6) To train the system by eliminating stop words and neutral words to obtain polarity of emotions and then to distinguish them in their respective classes.

Proposed System Architecture:


Input
Training /
Testing
Annotated
Emotion - Label
Document Corpus (Dataset)
Deep Learning
Using Word2Vec
Deep learning Library
λet Mixture Parameter
Calculate Z using Linear Equation Model
Z=1
Calculate Neutral Valence
Z=0
Calculate Emotional Valence
Initialize ILDA with Custom GIBB Sampling Parameters and Iteration Values
Obtain Burnout for Iterations
Obtain Vector Values for Words belonging to Document Dt under process
Generate Lexicon Set based upon Word - λ Value
Obtain Mean Reciprocal Rank
Obtain Accuracy, Precision , Recall and Fmeasure
90
45
0
Precis
Recal
Fmea
Mean
Precision
Recall
Fmeasure

The chief characteristics of the proposed model are as follows:
1) Every subject is entitled by patterns
2) Information is filtered using Stanford NLP framework.
3) Provide a more precise document modeling method for classification.

In (form) pattern based topic model, which has been employed in Information Filtering, can be acknowledged as a ”Post-LDA” model based on the patterns that are produced from the topic representations of the LDA model. Patterns can represent more specific meanings than single words. By comparing the word-based topic model with pattern-based topic models, the pattern based model can be used to represent the semantic content of the user’s documents more accurately than word based document.

Feature representation is the next stage in the process that includes representations and the use of skip-gram and n-gram, characters instead of words in a sentence, inclusion of a part-of-speech tag, or phrase structure tree.

Next process is to obtain aspect of data, using heuristic rules that we can define from our NLP framework and Penn Tree bank and obtain different aspects such as Nouns, Pronouns, Adjectives etc.

We need to calculate that the number of neurons and layers in a neural network has on an emotion classification task.

**Proposed System Steps:**

**Step 1:** Learn initial model from training data.
**Step 2:** Set mixture parameter λ using Word2Vec representation.

Set estimation of hidden variable *Zw.*
Perform Maximization step (M-step) and obtain parameter Theta(*e )*
Generate model for each document. (LDA, Theta Value for Dt)
Set Burnout Parameter.
Perform Gibbs Sampling
Calculate Emotional Valence
Calculate Neutral Valence
Obtain vector value for words and generate lexicon

***Data Collection***

The News articles dataset was utilized for this experiment. There are total 253 news headlines used in this proposed system.

***Preprocessing***

In preprocessing tokenization, stop word removal, stemming, punctuation marks removal, etc., has done. It has converted in bag of words. Preprocessing is important in sentiment analysis and opinion mining.

***Score Generation***

In this step, every sentence has analyzed and calculated sentiment score. To calculate sentiment score dataset has compared with opinion lexicons i.e. 2006 positive words and 4783 negative words and calculated sentiment score for every sentence.

***Sentiment Classification***

Using score and different features different machine learning algorithms has applied and different accuracy measurements calculated. Proposed method uses the

LDA Enhancement Algorithm:

```
Input: user interest model UE = { E(Z1), . . ., E(ZV)}, a list of incoming document Din
Output: rankE(d), d ∊ Din
rank(d) = 0
for each d ∊ Din do
    for each topic Zj ∊ [Z1, Zv] do
        for each equivalence class ECjk ∊ E(Zj) do
            5: scan ECk,j and find maximum matched pattern which exists in d
            update rankE (d) using equation(1)
            7: rank(d) := rank(d) + $\left| \right|^{0.5}$ * fjk * υD,j * uniform distribution * equivalent class frequency
        end for
    end for
end for
```

Word2Vec Algorithm :

**Step 1:**
word list (key = "getVecFromWord).
300-dimensional vector representation of a given word

**Step 2:**
(Required): List of 300-dimensional vectors (key = "getWordFromVec")

**Step 3:**
The top 10 words that are most consistent with the vector defined in the vector space

**Step 4:**
(Required): Two words list (key = "similarity between the words")

**Step 5:**
Similarity scores between the two words received

**Step6:**
(required): word list (key = "doesntMatch")

**Step 6:**
Return words that do not match the other words in the list.

**Step 7:**
(required): vector arithmetic using the algorithm proposed in the original word2vec paper (key = "vectorArithmetic")

***Sentiment Results***

News articles were classified in to positive, negative and neutral classes by looking at their total sentiment score. News articles sentiment was then calculated as the average value of total word sentiments.

## 4. Conclusion and Future Work

The main objective of using deep learning is that they aim to extract those features which disentangle the hidden factors of variations. This will help to perform the transfer across different domains. In this case, they were expecting the concept which characterized the review. They considered some of the factors like positive reviews to check the disentanglement of the dataset

**Table 2:** Evaluation parameters for classifiers of datasets

| Dataset | Accuracy | Precision Recall | F measure | Accuracy | Precision Recall | F measure |
|---|---|---|---|---|---|---|
| Camera | 98.17 | 98.3 | 99.03 | 93.54 | 93.58 | 96.66 |
| Laptops | 90.22 | 90.01 | 94.74 | 88.16 | 88.52 | 93.71 |
| Mobile phones | 92.85 | 91.64 | 95.64 | 92.85 | 91.64 | 95.64 |
| Tablets | 97.17 | 98.73 | 98.31 | 84.12 | 84.31 | 91.37 |
| TVs | 90.16 | 90.17 | 94.72 | 88.49 | 85.56 | 93.89 |
| Video surveillance | 91.13 | 89.95 | 94.71 | 79.43 | 84.25 | 88.53 |