INTERNATIONAL CONFERENCE ON ADVANCES IN SCIENCE, ENGINEERING AND ROBOTICS TECHNOLOGY (ICASERT-2019)

Statistical Approach for Classifying Sentiment Reviews by Reducing Dimension using Truncated Singular Value Decomposition

Asmaul Husna[1], **Habibur Rahman**[2] and Emrana Kabir Hashi[2]

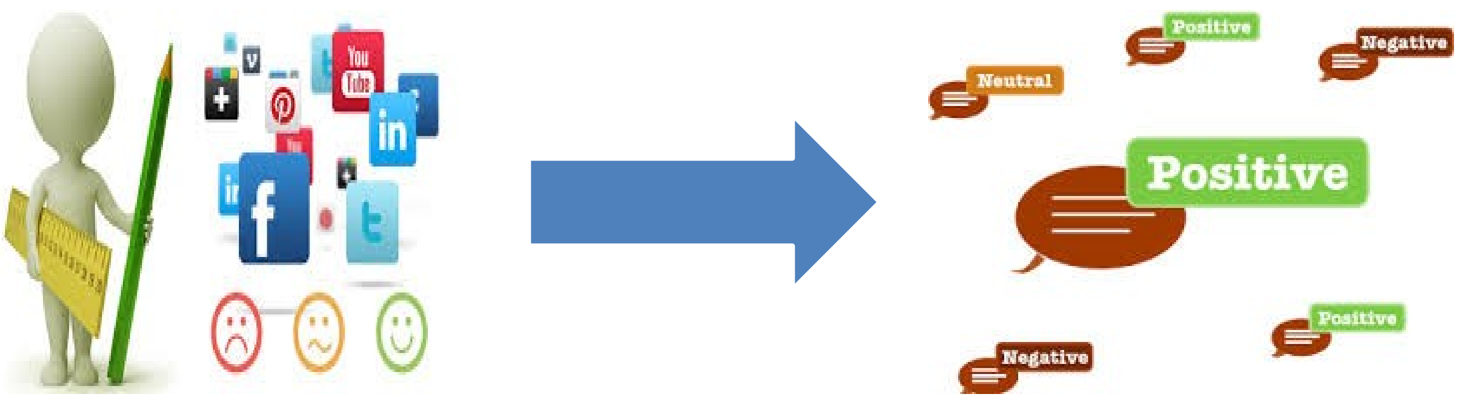
[1] Computer Science and Engineering, University of Information Technology and Sciences (UITS)
[2] Computer Science and Engineering, Rajshahi University of Engineering and Technology (RUET)

Outline

- Introduction
- Motivation
- Objectives
- Background Study
- Related Works
- Existing System
- Proposed System
- Methodology
- Implementation
- Result Analysis
- Comparison
- Conclusion, Limitation and Future Work
- References

Introduction

- Sentiment analysis is the computational study of people opinions or reviews expressing in online media.



Motivation

The opportunity to **capture the opinions** of the general public about **social events, political movements, company strategies, product preferences** which has **raised increasing interest** both in the

- **scientific community** for the exciting open challenges

and in the

- **business world** for the remarkable fallouts in marketing and financial prediction

Objectives

To design a system using suitable feature generation and extraction method with less computational cost

To create a time and cost-effective framework with fast learning speed machine learning algorithm

Background Study

Existing approaches of sentiment analysis can be grouped into four main categories [1] :

1. **Keyword spotting**
2. **Lexical affinity**
3. **Statistical methods**
4. **Concept-based techniques**

Keyword Spotting:

- Text is classified into **positive and negative** based on the presence of affect words like **'happy', 'sad', 'afraid', and 'bored'.**

Lexical Affinity:

- Assigns arbitrary words as **probabilistic 'affinity'** for a particular emotion.
- Example: **'accident'** might be assigned a **75%** probability of indicating a **negative affect**

Background Study(Cont'd)

Statistical Methods:

- Not only learn the affective valence of **affect keywords** (as in the **keyword spotting** approach), but also to take into account the valence of other arbitrary keywords (like **lexical affinity**), **punctuation**, and **word co-occurrence frequencies**

Concept-Based Techniques:

- Focus on a **semantic analysis** of text through the use of web ontologies or semantic networks
- Handle the **conceptual and affective information** rather than affective words but use **complex approach** than **statistical methods**

Related works

1. B. Pang, L. Lee, and S. Vaithyanathan [2]

- **Contribution:** They have used a mixture of lexical features such as unigrams, bigrams, POS with Naive Bayes, Maximum Entropy and Support Vector Machine.

- **Limitation:** Lower accuracy with higher dimension .

2. D. V. N. Devi, C. K. Kumar, and S. Prasad [3]

- **Contribution:** They have used SVM in a novel way to find out the overall positive and negative scores for a particular feature.

- **Limitation:** They have showed better **accuracy only** in **higher dimensional** feature space.

Existing System

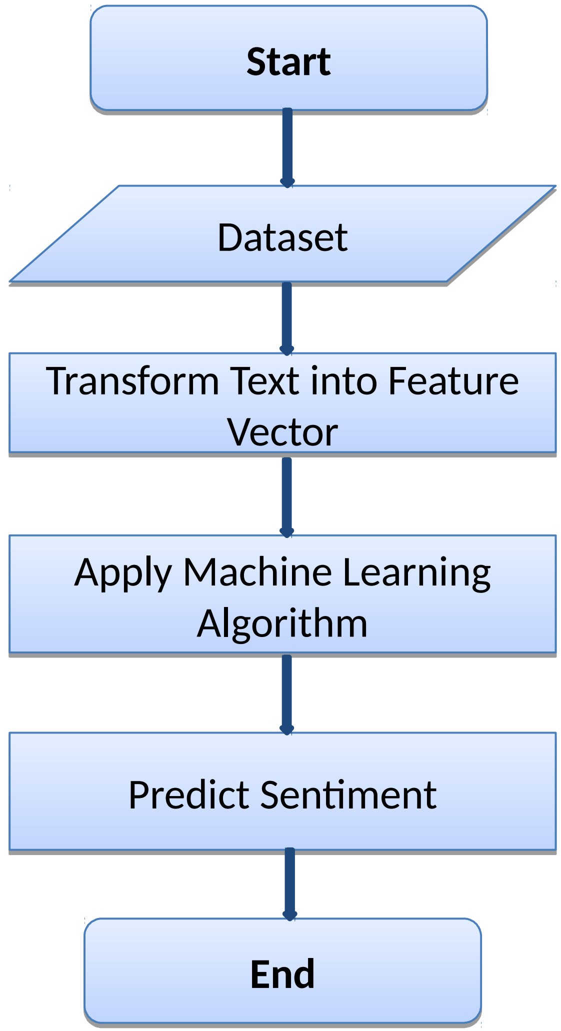


Fig. 1: Flowchart of existing system

Proposed System

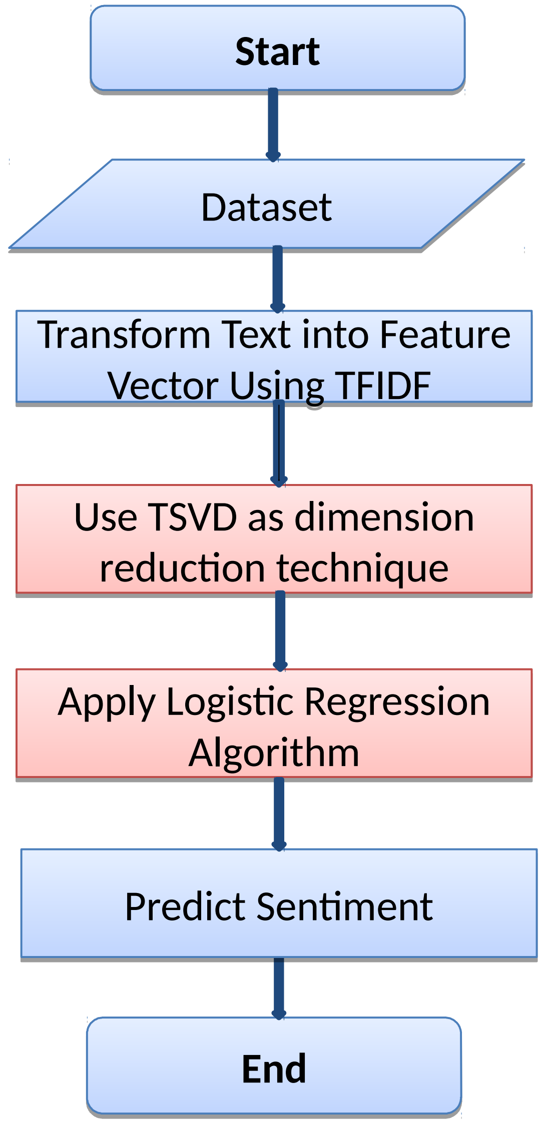


Fig. 2: Flowchart of proposed system

Methodology

- **Transform Text into Feature Vector Using TFIDF(Term Frequency-Inverse Document Frequency):**

Dataset(Text) → TFIDF → Term-document matrix of dataset

<u>Term- document matrix</u>

$$\begin{array}{c|ccc} & d_1 & d_2 & d_3 \\ \hline t_1 & w_{1,1} & w_{1,2} & w_{1,3} \\ t_2 & w_{2,1} & w_{2,2} & w_{3,3} \end{array}$$

Here,

d_1, d_2, d_3 = document/ sentence

t_1, t_2 = term/ word of each sentence

And $w_{1,1}$, $w_{1,2}$ …… $w_{3,3}$ = tf (t, d). idf(t, D) =weighted vector which represents the occurrence rate of each term in a particular document

Here, $tf(t, d)=\log(f_{t,d})$

$idf(t, D)=\log(N/N_{tEd})$

Methodology(Cont'd)

- **Example of tf-idf**

  Term for document 1= {this, is, a, sample}

  Term count for document 1={1,1,2,1}

  Term for document 2= {this, is, another, example}

  Term count for document 2={1,1,2,3}

  tf(example,d2)=log(3)=0.48

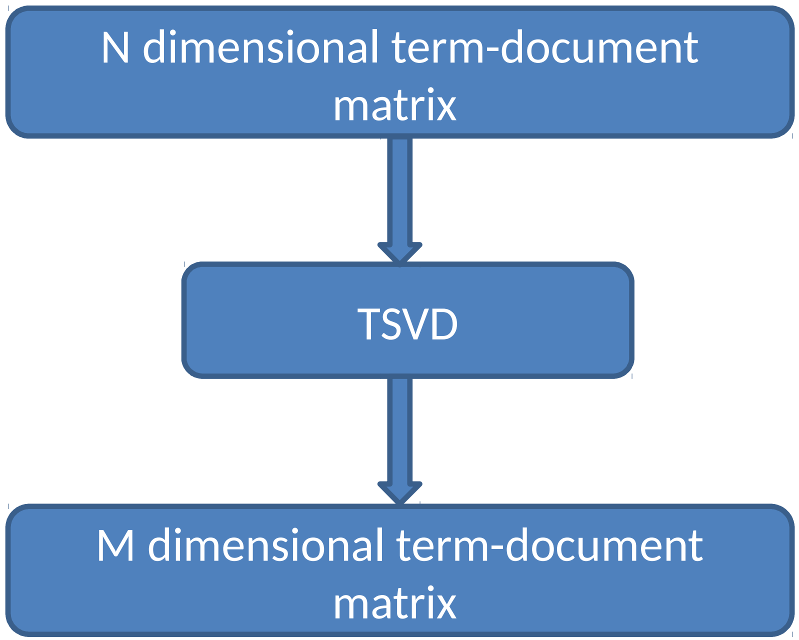
  idf(example,D)=log(2/1)=0.301

  Finally,

  tf-idf(example,d2)=0.48*0.301=0.144

Methodology(Cont'd)

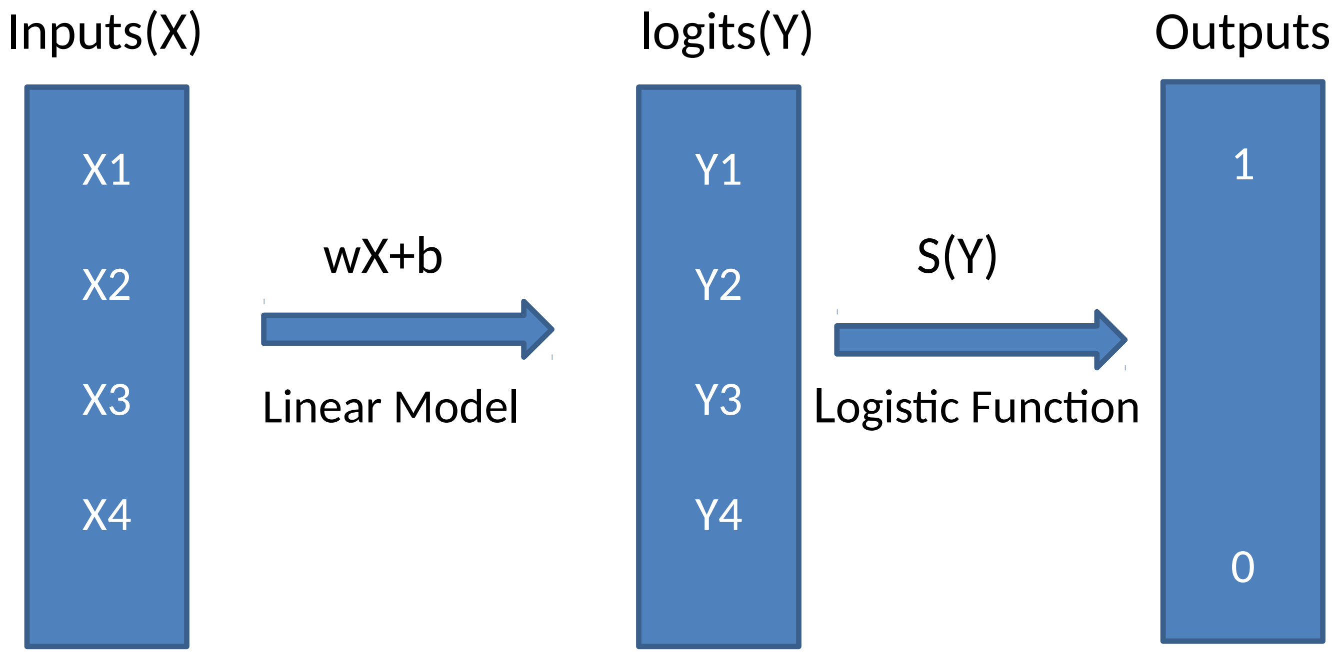
- **TSVD(Truncated Singular Value Decomposition) as Dimension Reduction Technique:**



Here, M<N

Methodology(Cont'd)

- **Logistic Regression Algorithm for Classification:**



Here, w = weight of the corresponding input

b = bias input

$S(Y) = e^Y/(1+e^Y)$

- **SVM(Support Vector Machine) for Classification:**
  Linear SVM has been used as a binary classifier to classify positive and negative sentiment.

The hyperplane can be defined as
$f(x)= b+ W^TX$
Here,
W=weight vector
X= input
b= bias input

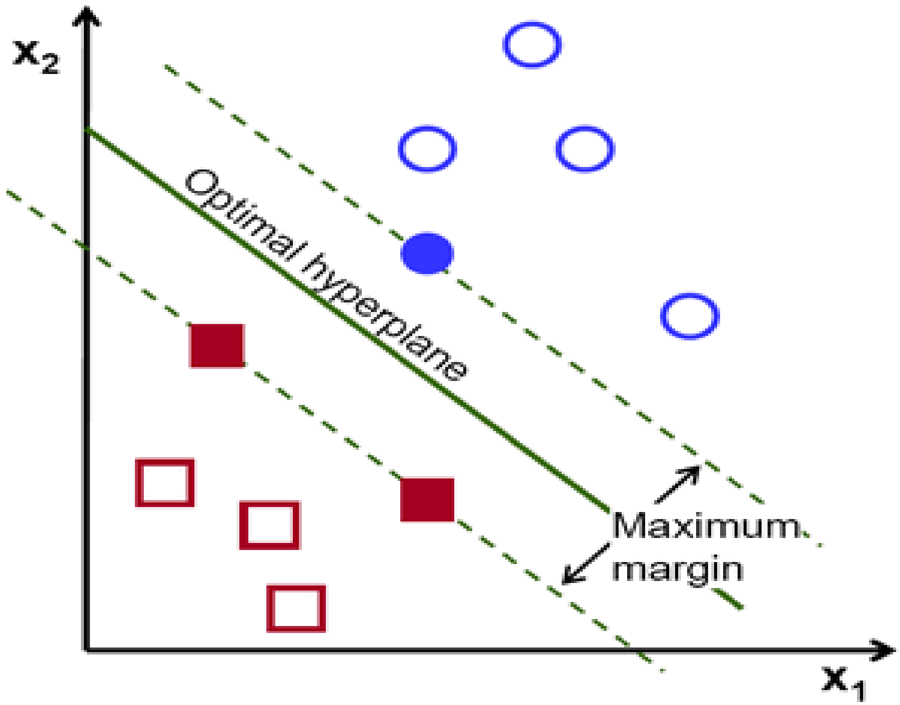


Fig. 3: Classification method of SVM[4]

Implementation

Dataset Description:

Movie review [5] and Yelp Restaurant Review [6] datasets are used.

- Those corpus includes **1,000 positive** and **1,000 negative** reviews.
- All text converted to lowercase and lemmatized, and HTML tags removed.

Hardware and Software Configuration:

- Operating System: Linux
- Language: Python 2.7
- Processor: Intel® Core™ i5-3317U CPU @ 1.70GHz
- RAM: 4.00 GB

Cross Validation:

- 10 fold cross validation has been used.

Result Analysis

TABLE 1: CONFUSION MATRIX WITHOUT DIMENSION REDUCTION

Movie Review Dataset					
		Actual			
Predicted		Negative		Positive	
	Classifier	SVM	LR	SVM	LR
	False	850	864	150	136
	True	154	137	846	863

Yelp Restaurant Review Dataset					
		Actual			
Predicted		Negative		Positive	
	Classifier	SVM	LR	SVM	LR
	False	901	907	93	93
	True	117	106	883	894

Result Analysis

Table 2: Result Analysis Without Dimension Reduction

Movie Review Dataset			
Classifier	**Accuracy (%)**	**#Features**	**Training Time**
SVM	84.4	12209	8.1166726
LR	86.35	12209	0.0763286
Yelp Restaurant Review Dataset			
Classifier	**Accuracy (%)**	**#Features**	**Training Time**
SVM	89.5	12209	1.4578088
LR	90.05	12209	0.0175273

Result Analysis

TABLE 3: CONFUSION MATRIX AFTER DIMENSION REDUCTION

Movie Review Dataset					
		Actual			
Predicted		Negative		Positive	
	Classifier	SVM	LR	SVM	LR
	False	818	849	182	151
	True	208	135	792	865

Yelp Restaurant Review Dataset					
		Actual			
Predicted		Negative		Positive	
	Classifier	SVM	LR	SVM	LR
	False	846	901	154	99
	True	103	115	897	885

Result Analysis

Table 4: Result Analysis After Dimension Reduction

Movie Review Dataset			
Classifier	**Accuracy (%)**	**#Features**	**Training Time**
SVM	80.5	100	0.5638779
LR	85.7	100	0.0437935
Yelp Restaurant Review Dataset			
Classifier	**Accuracy (%)**	**#Features**	**Training Time**
SVM	88.65	100	0.5866717
LR	89.3	100	0.0398641

Comparison

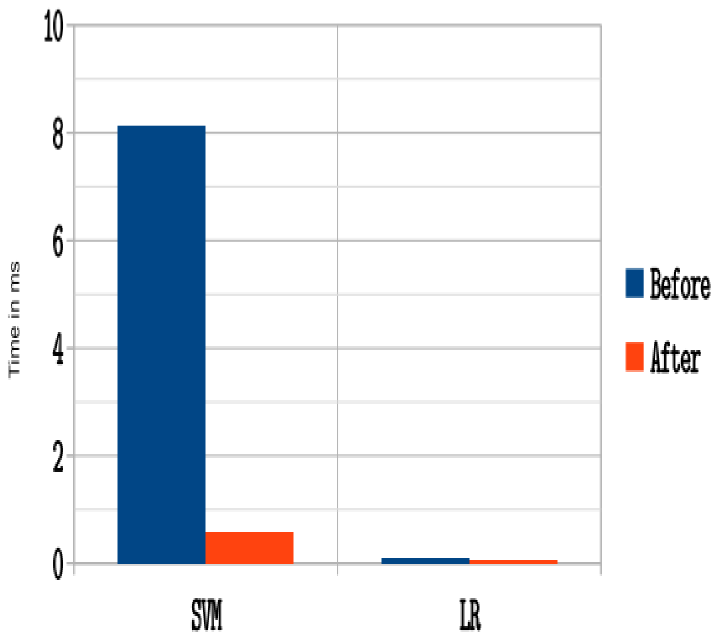


Fig. 4: Comparison of Training Time

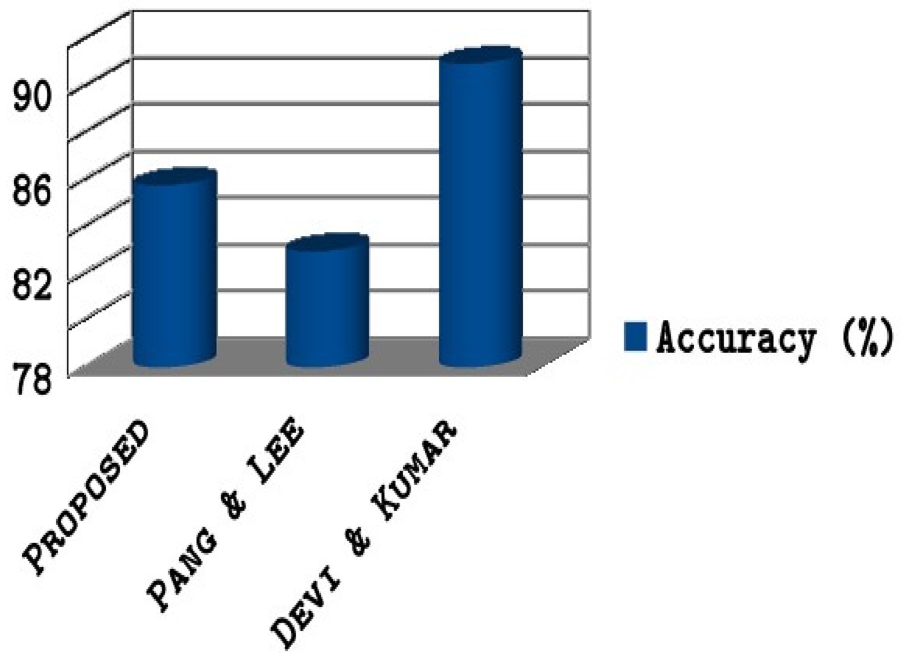


Fig. 5: Accuracy Comparison

Conclusion, Limitation and Future Work

Conclusion:

After reducing dimension, the accuracy are closest before and after dimension reduction. But in terms of training time, dimension reduction outperforms (42.63% Faster Training Time).

Limitation:

- Unigram TF-IDF is used for converting text into vector.

Future Work:

- Bigram TF-IDF will be used for better accuracy.
- Deep Neural Network for further improvements.

References

[1] E. Cambria, P. Gastaldo, F. Bisio, and R. Zunino, "An elm-based model for affective analogical reasoning," *Neurocomputing*, vol. 149, pp. 443–455, February 2015.

[2] B. Pang, L. Lee, and S. Vaithyanathan,"Thumbs up? sentiment classification using machine learning techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP), July 2002, vol. 10, pp. 79-86.

[3] D. V. N. Devi, C. K. Kumar, and S. Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine," *IEEE 6th International Conference on Advanced Computing*, February 2016.

[4] "Introduction to Support Vector Machine",https://docs.opencv.org/2.4/doc/tutorials /ml /introduction_to_svm/introduction_to_svm.html.

References

[5] B. Pang, L. Lee, “Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales,” *43rd Annual Meeting on Association for Computational Linguistics*, pp. 115–124, June 2005.

[6] Yelp.com. (2019). Yelp Dataset. [online] Available at: https://www.yelp.com/dataset [Accessed 15 Apr. 2019].

THANK YOU