An Efficient Approach for Sentiment Classification using Logistic Regression

Mineeta Khanuja
Bharti College of Engineering & Technology, Durg

Leelkanth Dewangan
Bharti College of Engineering & Technology, Durg

***Abstract*— Sentiment Analysis (SA) or Mood Classification or Mood Classification is the mathematical modelling approach of person's sentiments, insolences and mood toward an object. The object can signify events, individuals, or issues. Persons express sentiments as part of day-by-day communiqué. Sentiments can be umpired by a mish mash of indications such as facial emotions, prosodies, shrugs, and actions. Emotions are also enunciated by written texts or from their handwriting style. Linguistic is a prevailing tool to converse and pass on the information. Automatic sentiment classification and examination methods are useful in numerous applications with psychosomatic basis. For instance, it can be magnificently applied to acquire user favorites and comforts from users' personal written text and vocalizations. These methods are frequently considered in the field of the domain of behavior or mood modeling and customer response analysis. In this paper we have applied logistic regression as classifier, and we have achieved 98% accuracy.**

***Keywords*— TPR,FNR,NLP,IR,F-Score**

I. INTRODUCTION

In this day and age, social networking sites like Hike, Twitter, Facebook, LinkedIn etc. have developed an exceptional platform to share opinion or sentiments frequently in the form of text. These social networking websites are used for trading assessments or opinions about a product, movies, and policymaking or about any user fascinated matters in the form of posting remarks, pictures and get feedback from other users. This kind of user generated text on social web forums about any products, people, and any events is very useful in business, government and individual. Data from social networking sites are aggressively mined for developments and patterns of interests.

In political discussions for example, we could access out commons sentiments on a convinced voting contenders or political parties. The voting outcomes can also be forecast from party-political posts. The micro-blogging and social network sites are painstaking a very good source of information because individuals share and converse their thoughts about a certain topic spontaneously. They are also cast-off as data sources in the SA progression.

Automatic sentiment classification and examination methods are useful in numerous applications with psychosomatic basis. For instance, it can be magnificently applied to acquire user favorites and comforts from users' personal written text and vocalizations. These methods are frequently considered in the field of the domain of behavior or mood modeling and customer response analysis. Likewise, e-learning systems can assistance from affective tutoring approaches.

What is Sentiment Analysis?

- Using NLP, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit.
- Sometimes referred to as opinion mining, although the emphasis in this case is on extraction.
- Opinions and reactions to ideas are relevant to adoption of new ideas.
- Analyzing sentiment reactions on blogs can give insight to this process.

Challenges in Sentiment Classification

- People express opinions in complex ways.
- In review texts, lexical content alone can be misleading.
- Intra-textual and sub-sentential reversals, negation, and topic change common
- Rhetorical devices/modes such as sarcasm, irony, implication, etc.

For example a customer write letter to hardware store as:

> "Dear <hardware store>
> Yesterday I had occasion to visit <your competitor>.
> They had an excellent selection, friendly and helpful salespeople, and the lowest prices in town. You guys suck.
>
> Sincerely,"

There are many possibilities for what we might want to classify:

- Users
- Texts
- Sentences (paragraphs, chunks of text)
- Predetermined descriptive phrases (<ADJ N>, <N N>, <ADV ADJ>, etc)
- Words
- There seems to be some relation between positive words and positive reviews

For sentiment classification we have used bag of word model for attribute selection and logistic regression as classification which is meant for prediction of sentiment from text.

Regression analysis is a form of predictive modelling technique which examines the association between a dependent (target) and independent variable (s) (predictor). This technique is used for predicting, time series modelling and finding the fundamental effect relationship between the variables. For example, relationship between impulsive driving and number of road calamities by a driver is best premeditated through regression. There are multiple benefits of using regression analysis. They are as follows:

- It designates the substantial relationships between dependent variable and independent variable.
- It signposts the metier of impact of multiple independent variables on a dependent variable.

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line) fig 1 depicts the same as below:

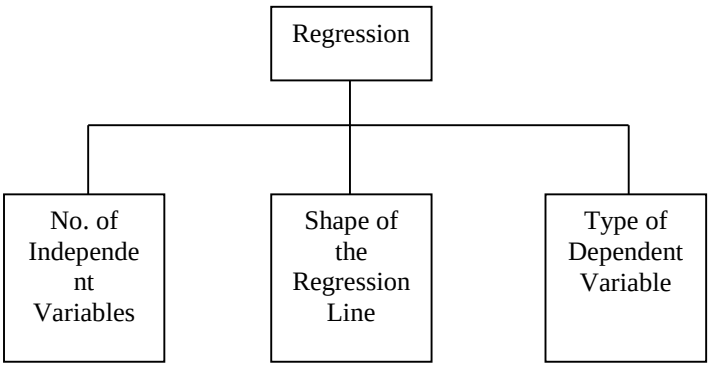


Fig.-1 Regression Technique

Logistic Regression: Logistic regression is castoff to catch the probability of event=Success and event=Failure. We can use logistic regression at situation when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence
ln(odds) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest.

Further in section-II we have gone through different literature and tabular comparison among literature, in section-III we will discuss bottle neck raised in earlier methods, in section IV we will discuss our proposed methodology, in section V implementation of proposed method discussed, at last we will conclude our research.

II. Literature Survey

Erik Cambria said that emotions play an important role in successful and effective human–human communication. In fact, in many situations, emotional intelligence is more important than IQ for successful interaction. There is also signif cant evidence that rational learning in humans is dependent on emotions. Affective computing and sentiment analysis, hence, are key for the advancement of AI and all the research f elds that stem from it. Moreover, they f nd applications in various scenarios and companies, large and small, that include the analysis of emotions and sentiments as part of their mission. Sentiment-mining techniques can be exploited for the creation and automated upkeep of review and opinion aggregation websites, in which opinionated text and videos are continuously gathered from the Web and not restricted to just product reviews, but also to wider topics such as political issues and brand perception [IEEE 2016].

Sufal Das et. al. concluded that Sentiment analysis is technically very challenging but more promising techniques are available, and it will become increasingly important as more people are buying and expressing their opinions on the web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers and has wide applications. Since people are interacting through internet, a huge data is being generated every second. Thus, a distributed parallel computing environment is very much needed to perform sentiment analysis efficiently [IJARCS 2015].

Basant Agarwal et. al. emphasis of paper is to discuss the research involved in applying machine learning methods mostly for sentiment classifcation at document level. Machine learning-based approaches work in the following phases, which are discussed in detail in this chapter for sentiment classifcation: (1) feature extraction, (2) feature weighting schemes, (3) feature selection, and (4) machine-learning methods [Speinger 2017].

Vivek Narayanan et. al. observed that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a significant improvement in accuracy. Which implies that a highly accurate and fast sentiment classifier can be built using a simple Naive Bayes model that has linear training and testing time complexities. Author achieved an accuracy of 88.80% on the popular IMDB movie reviews dataset. Author aslso ssaisd that The proposed method can be generalized to a number of text categorization problems for improving speed and accuracy [Springer 2013].

Soujanya Poria et. al. proposed a novel multimodal affective data analysis framework. It includes the extraction of salient features, development of unimodal classifiers, building feature- and decisionlevel fusion frameworks. The deep CNN-SVM -based textual sentiment analysis component is found to be the key element for outperforming the state-of-the-art model's accuracy. MKL has played a significant role in the fusion experiment. The novel decision-level fusion architecture is also an important contribution of this paper. In the case of the decision-level fusion experiment, the coupling of sentic patterns to determine the weight of textual modality has enriched the performance of the multimodal sentiment analysis framework considerably [Elsevier 2017].

David Zimbra et. al. present an approach to brand-related Twitter sentiment analysis using feature engineering and the Dynamic Architecture for Artificial Neural Networks (DAN2). The approach addresses challenges associated with the unique characteristics of the Twitter language, and the recall of mild sentiment expressions that are of interest to brand management practitioners. Author demonstrate the effectiveness of the approach on a Starbucks brand-related Twitter data set. The feature engineering produced a final tweet feature representation consisting of only seven dimensions, with greater feature density. Two sets of experiments were conducted in three-class and fiveclass tweet sentiment classification. Author compare the proposed approach to the performances of two stateof-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicate that the approach outperforms these state-ofthe-art systems in both three-class and five-class tweet sentiment classification by wide margins, with classification accuracies above 80% and excellent recall of mild sentiment tweets [IEEE 2016].

S.No.	Author/Paper title/Year	Description

1.	Erik Cambria/Affective Computing and Sentiment Analysis/IEEE 2016	Erik Cambria said that emotions play an important role in successful and effective human–human communication. Sentiment-mining techniques can be exploited for the creation and automated upkeep of review and opinion aggregation websites, in which opinionated text and videos are continuously gathered from the Web and not restricted to just product reviews, but also to wider topics such as political issues and brand perception.
2.	David Zimbra et. al./Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks/IEEE 2016	Author compare the proposed approach to the performances of two stateof-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicate that the approach outperforms these state-ofthe-art systems in both three-class and five-class tweet sentiment classification by wide margins, with classification **accuracies above 80%** and excellent recall of mild sentiment tweets.
3.	Vivek Narayanan et. al./Fast and accurate sentiment classification using an enhanced Naive Bayes model/Springer 2013	Naive Bayes model that has linear training and testing time complexities. We achieved an **accuracy of 88.80%** on the popular IMDB movie reviews dataset. The proposed method can be generalized to a number of text categorization problems for improving speed and accuracy.
4.	Soujanya Poria et. al./Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis/Elsevier 2017	The proposed framework outperforms the state-of-the-art model in multimodal sentiment analysis research with a margin of 10–13% and 3–5% accuracy on polarity detection and emotion recognition, respectively. The paper also proposes an extensive study on decision-level fusion, **Average accuracy of 81%.**
5.	Sufal Das et. al./Sentiment Analysis for Web-based Big Data: A Survey/ IJARCS 2017	Sentiment analysis is a very challenging and promising discipline which uses both intersection of information retrieval and computational linguistic techniques to deal with the reviews expressed in a source material. This work talks about the sentiment analysis process and focus on some machine learning techniques for sentiment classification and future challenges in opinion mining for big data.

Xilun Chen et. al. presented ADAN, an adversarial deep averaging network for cross-lingual sentiment classification, which, for the first time, applies adversarial training to cross-lingual NLP. ADAN leverages the abundant resources on English to help sentiment analysis on other languages where little or no annotated data exist. We validate our hypothesis by empirical experiments on Chinese and Arabic sentiment classification, where we have labeled English data and only unlabeled data in the target language. Experiments show that ADAN outperforms several baselines including domain adaptation models and a highly competitive MT baseline. We further show that even without any bilingual resources, ADAN trained with random initialized embeddings can still achieve meaningful cross-lingual performance. In addition, we show that in the presence of labeled data in the target language, ADAN can naturally incorporate this additional supervision and yields even more competitive results [arXiv 2017].

III. Problem Identification

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. Sentiment = feelings

- Attitudes
- Emotions
- Opinions

There are several application of sentiment analysis which motivated us towards this research.

- Based on a sample of tweets, how are people responding to this ad campaign/product release/news item?
- How have bloggers attitudes about the president changed since the election?
- Identifying child-suitability of videos based on comments.
- Identifying (in) appropriate content for ad placement.
- Use SA to

A) Search the web for opinions and reviews of this and competing laptops. Blogs, Opinions, amazon, tweets, etc.
B) Create condensed versions or a digest of consensus points

Earlier methods are not capable for following:

- Spam/Stop words present in input dataset.
- Redundant feature in input data.
- Less Accurate.

S. No.	Author	Method Used	Accuracy %
1.	Vivek Narayanan et. al. Springer 2013	Naive Bayes model	88%
2.	Soujanya Poria et. Al. Elsevier 2017	Multimodal Framework	81%
3.	David Zimbra et. al. IEEE 2016	Dynamic Artificial Neural Network	80%

Table-1 Accuracy of Algorithm

IV. Solution Methodology

After going through different literature we came across some problem which need to overcome in existing technique to improve the accuracy. Our proposed algorithm is divided into different phases.

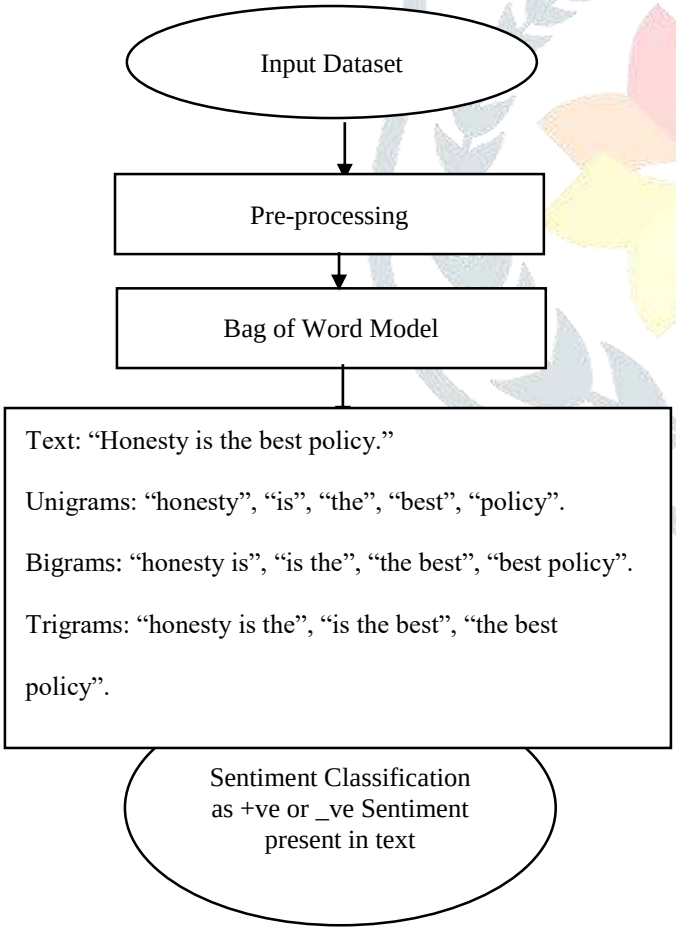


Fig.-2 Proposed Flow

Proposed Algorithm

Step-1. Read Training Data
Step-2. Read Test Data
Step-3. Remove non letters from test data.
Step-4. Tokenize training data and test data.
Step-5. For each, print the vocabulary word and the number of times which appears in the training set.
Step-6. Remove unlabeled data from both training and test dataset.
Step-7. Train the training dataset using logistic regression.
Step-8. Get +ve and –ve as prediction of test dataset.

Attribute selection

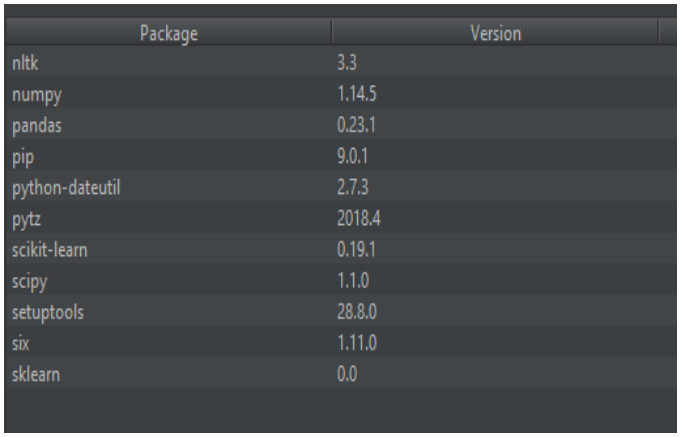
After going through different literature we came across conclusion that applying N-gram feature vector method will eventually help to improve classification accuracy.

N-gram: An n-gram model is a type of probabilistic language model for predicting the next word conditioned on a sequence of previous words using Markov models. N-gram of size 1 is referred to as unigram, size 2 as bigram, and size 3 as trigram. Since n-grams are used for capturing dependencies between single words that stay in a text sequentially, the combination of words does not necessarily have syntactical or semantic relations. Unigrams performed much better than bigrams when used as features for feature spaces in Pang et al. (2002), while bigrams and trigrams contributed higher performance than unigrams in (Dave et al., 2003; Ng et al., 21 2006). In Pang et al.(2002), unigrams also outperformed adjectives when treated as features.

Unigrams presents the simplest model for the n-gram approach. It consists of all the individual words present in the text. The bigram model defines a pair of adjacent words. Each pair of words forms a single bigram. The higher order grams can be formed in the similar way by taking together the n adjacent words. Higher order n-grams are more efficient in capturing the context as they provide better understanding of the word position. An n-gram defines a subsequence of n items from a given sequence. It is used in various fields of natural language processing and genetic sequence analysis. An n-gram model defines a method for finding a set of n-gram words from a given document. The commonly used models include unigrams (n=1), bigrams (n=2) and trigrams (n=3). However the value of n can be extended to higher level grams. The ngram model can be better explained with the following examples:

V. Result and Discussion

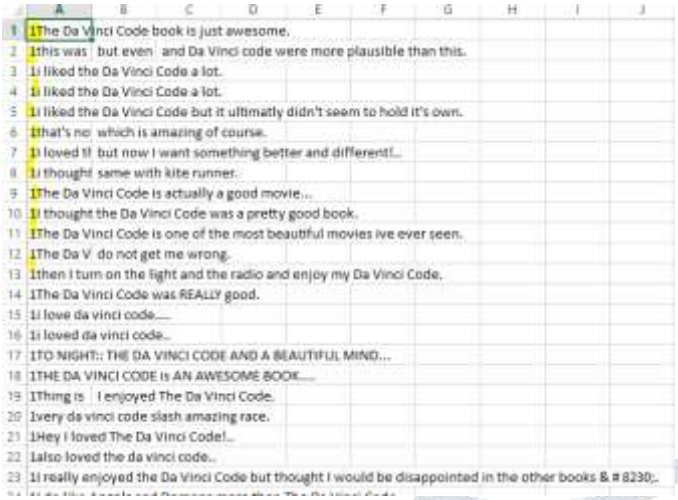
For implementation of proposed work we have used Python 2.6 with Pycharm environment and following libraries are used:



Package	Version
nltk	3.3
numpy	1.14.5
pandas	0.23.1
pip	9.0.1
python-dateutil	2.7.3
pytz	2018.4
scikit-learn	0.19.1
scipy	1.1.0
setuptools	28.8.0
six	1.11.0
sklearn	0.0

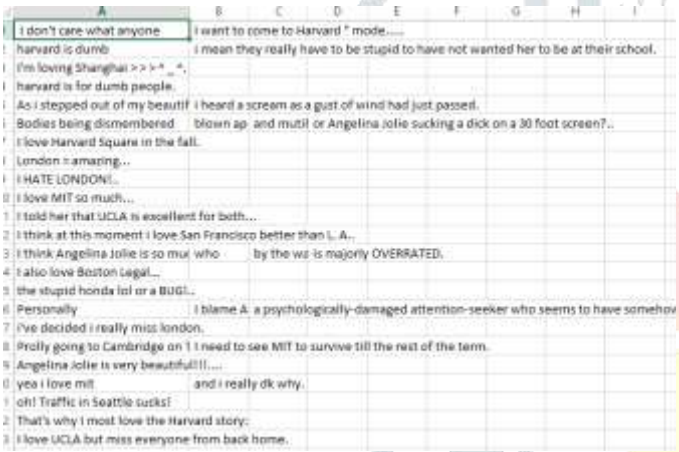
Fig.-3 Packages used in Implementation

Fig.-4 shows the snippet of training dataset, highlighted portion shows the shows sentiment 1 as +ve sentiment and 0 as –ve sentiment.



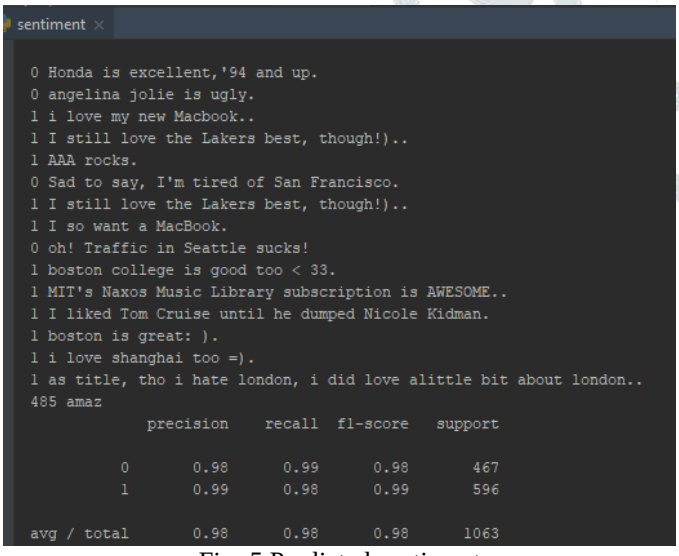
	A	B	C	D	E	F	G	H	I	J
1	1	The Da Vinci Code book is just awesome.								
2	1	this was	but even	and Da Vinci code were more plausible than this.						
3	1	i liked the Da Vinci Code a lot.								
4	1	i liked the Da Vinci Code a lot.								
5	1	i liked the Da Vinci Code but it ultimatly didn't seem to hold it's own.								
6	1	that's no	which is amazing of course.							
7	1	i loved t	but now I want something better and different!..							
8	1	i thought	same with kite runner.							
9	1	The Da Vinci Code is actually a good movie...								
10	1	i thought the Da Vinci Code was a pretty good book.								
11	1	The Da Vinci Code is one of the most beautiful movies ive ever seen.								
12	1	The Da V	do not get me wrong.							
13	1	then i turn on the light and the radio and enjoy my Da Vinci Code.								
14	1	The Da Vinci Code was REALLY good.								
15	1	i love da vinci code.....								
16	1	i loved da vinci code...								
17	1	TO NIGHT:: THE DA VINCI CODE AND A BEAUTIFUL MIND...								
18	1	THE DA VINCI CODE is AN AWESOME BOOK.....								
19	1	Thing is	I enjoyed The Da Vinci Code.							
20	1	very da vinci code slash amazing race.								
21	1	Hey I loved The Da Vinci Code!...								
22	1	also loved the da vinci code..								
23	1	i really enjoyed the Da Vinci Code but thought I would be disappointed in the other books & # 8230;..								

Fig.-4 Training Dataset



	A	B	C	D	E	F	G	H	I
1	I don't care what anyone	i want to come to Harvard " mode.....							
2	harvard is dumb	i mean they really have to be stupid to have not wanted her to be at their school.							
3	i'm loving Shanghai >> >.^ _ ^.								
4	harvard is for dumb people.								
5	As i stepped out of my beautif	i heard a scream as a gust of wind had just passed.							
6	Bodies being dismembered	blown ap	and mutil	or Angelina jolie sucking a dick on a 30 foot screen?..					
7	i love Harvard Square in the fall.								
8	London = amazing...								
9	i HATE LONDON!..								
10	i love MIT so much...								
11	i told her that UCLA is excellent for both...								
12	i think at this moment i love San Francisco better than L. A..								
13	i think Angelina Jolie is so mu	who	by the wa	is majorly OVERRATED.					
14	i also love Boston Legal...								
15	the stupid honda lol or a BUG!..								
16	Personally	i blame A	a psychologically-damaged attention-seeker who seems to have somehov						
17	i've decided i really miss london.								
18	Prolly going to Cambridge on 1	i need to see MIT to survive till the rest of the term.							
19	Angelina Jolie is very beautiful!!!.....								
20	yea i love mit	and i really dk why.							
21	oh! Traffic in Seattle sucks!								
22	That's why I most love the Harvard story:								
23	i love UCLA but miss everyone from back home.								

Fig.-4 Test Dataset



```
sentiment
0 Honda is excellent,'94 and up.
0 angelina jolie is ugly.
1 i love my new Macbook..
1 I still love the Lakers best, though!)..
1 AAA rocks.
0 Sad to say, I'm tired of San Francisco.
1 I still love the Lakers best, though!)..
1 I so want a MacBook.
0 oh! Traffic in Seattle sucks!
1 boston college is good too < 33.
1 MIT's Naxos Music Library subscription is AWESOME..
1 I liked Tom Cruise until he dumped Nicole Kidman.
1 boston is great: ).
1 i love shanghai too =).
1 as title, tho i hate london, i did love alittle bit about london..
485 amaz
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       467
           1       0.99      0.98      0.99       596

avg / total       0.98      0.98      0.98      1063
```

Fig.-5 Predicted sentiemnt

Proposed	Precision	Recall	F-score	support
0 (-ve Sentiment)	0.98	0.99	0.98	467
1 (+ve Sentiment)	0.99	0.98	0.99	596
Average	0.98	0.98	0.98	1063

Table-3 Proposed algorithm accuracy

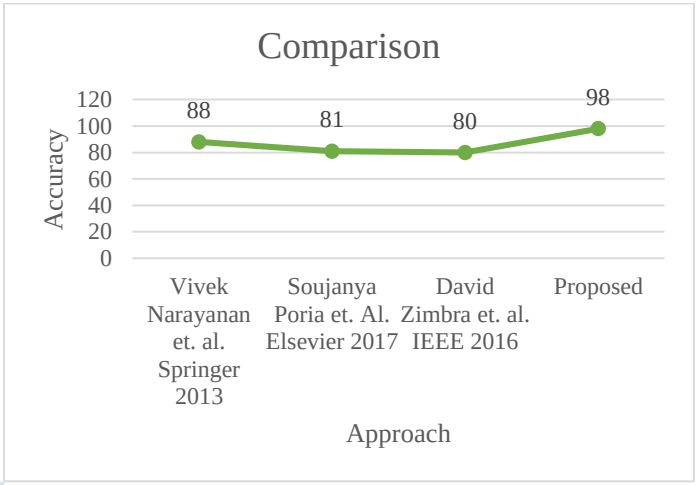


Fig.-6 Graphical Representation of Accuracy Comparison

VI. CONCLUSION

In the context of sentiment analysis, for example, the products and services recommended to a person should be those that have been positively evaluated by other users with a similar personality type. After going through implementation, our algorithm performed well and achieved average 98% accuracy. In future we can apply deep learning to improve the accuracy over large dataset and further we can apply better preprocessing technique to reduce computation time.

REFERENCES

[1] Navonil Majumder et. al. Deep Learning-Based Document Modeling for Personality Detection from Text IEEE 2017.
[2] Erik cambria et.al. Affective Computing and Sentiment Analysis IEEE 2016
[3] Giulio Angiani et. al. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter Springer 2016
[4] I.Hemalatha et. al. Preprocessing the Informal Text for efficient Sentiment Analysis IJETTCS 2012
[5] Ahmed Hassan et. al. Sentiment analysis algorithms and applications: A survey Elsevier 2014
[6] Daoud Clarke et. al. Developing Robust Models for Favourability Analysis ACLWEB 2012
[7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passoneau, R.: Sentiment Analysis of Twitter Data. Computer Science - Columbia University (New York, USA) (2011).
[8] Balahur, A.: Sentiment Analysis in Social Media Texts. European Commission Joint Research Center (Varese, Italy) (2013).
[9] Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. In: International Conference on Intelligent Computing. pp. 615{624. Springer (2014).
[10] Cambria, E., Olsher, D., Rajagopal, D.: Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Twenty-eighth AAAI conference on artificial intelligence (2014) [5] Duncan.
[11] B., Zhang, Y.: Neural networks for sentiment analysis on twitter. In: Cognitive Informatics & Cognitive Computing

(ICCI* CC), 2015 IEEE 14th International Conference on. pp. 275{278. IEEE (2015)
[12] Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC. vol. 6, pp. 417{422. Citeseer (2006)
[13] Fornacciari, P., Mordonini, M., Tomaiuolo, M.: A case-study for sentiment analysis on twitter. In: Proceedings of the 16th Workshop "From Objects to Agents"-WOA (2015).
[14] E. Cambria, "Affective Computing and Sentiment Analysis," IEEE Intelligent Systems, vol. 31, no. 2, 2016, pp. 102–107.