International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

A ML and NLP based Framework for Sentiment Analysis on Bigdata

D. Krishna Madhuri, R. V. V. S. V PRASAD

Abstract: Big data as multiple sources and social media is one of them. Such data is rich in opinion of people and needs automated approach with Natural Language Processing (NLP) and Machine Learning (ML) to obtain and summarize social feedback. With ML as an integral part of Artificial Intelligence (AI), machines can demonstrate intelligence exhibited by humans. ML is widely used in different domains. With proliferation of Online Social Networks (OSNs), people of all walks of life exchange their views instantly. Thus they became platforms where opinions or people are available. In other words, social feedback on products and services are available. For instance, Twitter produces large volumes of such data which is of much use to enterprises to garner Business Intelligence (BI) useful to make expert decisions. In addition to the traditional feedback systems, the feedback (opinions) over social networks provide depth in the intelligence to revise strategies and policies. Sentiment analysis is the phenomenon which is employed to analyze opinions and classify them into positive, negative and neutral. Existing studies usually treated overall sentiment analysis and aspect-based sentiment analysis in isolation, and then introduce a variety of methods to analyse either overall sentiments or aspect-level sentiments, but not both. Usage of probabilistic topic model is a novel approach in sentiment analysis. In this paper, we proposed a framework for comprehensive analysis of overall and aspect-based sentiments. The framework is realized with aspect based topic modelling for sentiment analysis and ensemble learning algorithms. It also employs many ML algorithms with supervised learning approach. Benchmark datasets used in international SemEval conferences are used for empirical study. Experimental results revealed the efficiency of the proposed framework over the state of the art.

Index Terms –Big data, NLP, sentiment analysis, machine learning, artificial intelligence, ensemble learning, Twitter, aspect-based sentiment analysis

I. INTRODUCTION

Enterprises in the real world have their data warehouse for keeping track of business data. Such data assumes characteristics of big data and provides wealth of knowledge when discovered and interpreted using data mining techniques. Such technical knowhow is invariably used by enterprises to make strategic decisions for growth. However, the business intelligence extracted from data warehouse is considered inadequate in the contemporary era where Online Social Networks (OSNs) produce voluminous data having significant latent trends.

Revised Manuscript Received on February 28, 2020.
* Correspondence Author
D.Krishna Madhuri*, Assistant Professor,Dept of CSE, GRIET, Hyderabad, Telangana India. Email: krishnamadhuri.530@gmail.com
Dr. R. V. V. S. V Prasad, Professor & Head,Dept of IT, Swarandhra College of Engineering & Technology, Narsapur, India. Email: ramayanam.prasad@gmail.com

Twitter [1] is one such OSN which exhibits exponential growth of tweets every year. This data is actually goldmine to researchers and enterprises when exploited by using a phenomenon, which emerged of late, known as opinion mining or sentiment analysis. Many researchers contributed to exploit data of OSN and other sources of Internet where reviews are made available. In addition to classifying sentiments into Positive, Negative and Neutral, of late, aspect based sentiment analysis is given importance [1]. Moreover, previous studies usually treat overall sentiment analysis and aspect-based sentiment analysis in isolation, and then introduce a variety of methods to analyse either overall sentiments or aspect-level sentiments, but not both. Usage of probabilistic topic model is a novel approach in sentiment analysis.

Latent Dirichl*et al*location (LDA) is the generative process model used for processing documents in various applications [13], [27]. In fact, it is widely used in processing online reviews or opinions over Twitter tweets as reviewed in [20], [21]. There are many supervised learning approaches, unsupervised methods [19], [22] and semi-supervised approaches [24]. Neural Networks (NNs) [7] and Convolutional Neural Networks (CNNs) [25] are also used for sentiment analysis. There are ensemble classifiers used for sentiment analysis as studied in [26]. Feature selection is found given importance in sentiment analysis. Based on syntax models and context, it is employed appropriately [29]. Topic modelling is widely used based on LDA [3], [5], [6], [12] and [30]. Along with topic modelling, aspect based approaches are found in [4], [10], [31], [34] and [39]. Aspect based sentiment analysis could provide more useful knowledge due to its utility in making decisions.

Hai *et al*. [41] proposed a topic modelling approach for analysing sentiments. It was efficient when compared with the state of the art. However, it has the following drawbacks. It has no provision for spatio-temporal sentiment analysis of online reviews or Twitter tweets as part of semantic aspect detection and aspect-level sentiment identification. Estimating the number of latent topics for efficient probabilistic topic modelling is not included in their model. There is no provision for deep learning in their model which causes mediocre performance in sentiment analysis. In order to overcome these drawbacks, the aim of the proposed research is to develop a comprehensive framework that considers probabilistic topic modeling with both aspect level and overall sentiment analysis in sentiment identification. Our contributions in this paper are as follows.

1. Proposed a comprehensive framework that considers overall sentiment analysis and aspect based sentiment analysis with an effective training model.
2. We proposed algorithm for Aspect Based Topic Modelling for Sentiment Analysis (ABTM-SA).
3. An ensemble learning model is proposed and implemented.

Several machine learning algorithms are used for performance analysis.

4. We built a prototype application to evaluate the proposed framework and compare the results with the state of the art.

The remainder of the paper is structured as follows. Section 2 reviews relevant literature on different models for sentiment analysis. Section 3 presents the proposed methodology. Section 4 provides details of benchmark datasets used for empirical study. Section 5 showed experimental results while section 6 concludes the paper and provide directions for future work.

II. RELATED WORK

Sentiment classification has attracted many researchers and academia of late. Here is the review of various methods existing. Overview of sentiment analysis and its applications are provided in [1] while the Yang and Cardie [2] focused on the context based sentiment analysis. It is made at sentence level. They employed Conditional Random Field (CRF) model for learning sentence level sentiments. They found it to be useful with both supervised and unsupervised approaches. They intended to focus more on refined constraints pertaining to neutral sentiments. Lin *et al*. [3] proposed a joint sentiment topic detection model with weakly supervised learning approach. It is based on Latent Dirichl*et al*location (LDA). It discovers topics and sentiments simultaneously. They achieved cohesive and informative results. However, incremental learning of their method is not yet explored. Kim *et al*. [4] proposed a hierarchical aspect sentiment model (HASM) which discovers hierarchical structure of aspects and their sentiments. From consumer and technology standpoints, they understood the significance of hierarchical aspects. Still they intended to improve it to discover set of topics shared features in the hierarchical aspect structure. A joint model is proposed by Dermouche*et al*. [5] for evaluation of topic-sentiment over temporal domain. The model is based on LDA. Their model showed performance improvement. However, hyperparameter setting in their model is not yet addressed.

Yang *et al*. [6] focused on parametric and non-parametric models that are user-centric. They are known as User-aware Sentiment Topic Models (USTM). They considered demographic information into the modeling to improve performance in opinion mining. They intended to improve their models with machine learning and natural processing techniques based priors. Unlike the work in [2], Tang *et al*. [7] focused on document level classification by learning product and user representations. They employed a Neural Network (NN) approach. Their method could improve classification accuracy. Hu *et al*. [8] proposed a sentiment classification model based on social relations on microblogging sources. They found that social relations can help in sentiment discovery. They intended to explore further on sentiment diffusion process in social media.

Hai *et al*. [9] explored associations for mining opinion features. It is known as feature based opinion analysis. With a small set of feature seeds, they started and investigated further exploiting likelihood ratio tests (LRTs), Latent Semantic Analysis (LSA) and pair-wise associations. They found that it is possible to create domain features automatically. They intended to improve it with fine-grained topic modeling in future. Hai *et al*. [10] proposed a Supervised Joint Aspect and Sentiment Model (SJASM) which is a topic model and probabilistic in nature. It has ability to infer underlying sentiments. It could provide better performance over its predecessors. They are yet to address the problem of coarse-to-fine review selection.

Hai *et al*. [11] focused on identifying features pertaining to sentiment mining by using criteria like Intrinsic-Domain Relevance (IDR) and Extrinsic-Domain Relevance (EDR). They called it as interval thresholding approach. Their method showed better performance over state of the art. They intended to improve it further using fine-grained topic modeling approach. Cao *et al*. [14] proposed a model known as Visual Sentiment Topic Model (VSTM) for image sentiment analysis. Visual ontology features made it more intuitive. Tang *et al*. [15] investigated sentiment specific word embeddings. They used it for sentiment analysis applications using neural networks to improve Natural Language Processing (NLP). They verified it on word level, sentence level and lexical level. Lim *et al*. [16] proposed a topic model which makes use of Poisson-Dirichlet Processes (PDP) for modeling text for different applications like automatic topic labelling. It is an inference framework that can be used for sentiment analysis. They intended to employ posterior inference algorithm to their model in future.

Steinskog*et al*. [17] employed LDA with tweets aggregation technique for sentiment analysis. However, they are yet to explore other pooling techniques to overcome inherent limitations with short documents. Chen *et al*. [18] focused on the difference between analyzing sentiments with regular tweets and retweets. They could find the effect of tweet diffusion by employing a sentiment factor in their model. They planned to improve their model with gam theory approach. Babu and Pattani [19] proposed a clustering approach to improve sentiment analysis based on clusters. They used DBSCAN method for this. They intended to improve it further with event detection as well. Suresh and Raj [22] also employed a fuzzy clustering method for mining sentiments from tweets. It could lead to quality results when compared with Expectation Maximization (EM) and K-Means.

Gyori*et al*. [23] proposed a method for sentiment analysis which is meant for citizen-contributed urban planning. They intended to improve it further with location-specific features. Yan *et al*. [26] proposed two ensemble classifiers that are made up of many off the shelf classification techniques. They could observe the effectiveness of sentiment analysis with such prediction models. Yoon *et al*. [28] proposed a method for finding opinions about political issues. They employed LDA along with SVM for higher detection accuracy. At present their method does not consider relationship between terms. A set of feature selection methods for sentiment analysis are proposed by Duric and Song [29]. Those methods are based on context and syntax models. With feature selection, they were able to represent salient features of documents and thus improve quality in sentiment prediction. They intended to classify sentiments with finer level of granularity.

Aspect based sentiment analysis with SenticLDA [31], LDA based non-parametric model [32], AS-LDA [33], multi-aspect sentiment analysis [34], topic based mixture modeling [35] are other important contributions. A hybrid model based on LDA [36], probabilistic topic modeling [37], Concept Level Sentiment Analysis (CLSA) [38], probabilistic model based on syntax and topic for aspect based approach [39] and topic trends and user interests based topic model [40] are other useful approaches found for sentiment analysis. NSP is widely used for sentiment analysis as explored in [46] and [47]. From the literature it is found that aspect based and overall sentiment analysis were made individually. We proposed a comprehensive framework that uses joint approach based on generative process model and also we have made a training model to improve performance of sentiment analysis.

III. PROPOSED FRAMEWORK

Provided a set of online reviews or tweets from Twitter, finding sentiments and classify them into positive, negative and neutral is the fundamental problem considered. Besides, the proposed framework considers ML approaches and LDA based aspect-topic model for effectiveness in sentiment analysis.

A.The Framework

A framework is proposed to have a comprehensive sentiment analysis of tweets or online reviews that contain opinions that are valuable to businesses. In fact, they reflect social feedback that comes from different stakeholder

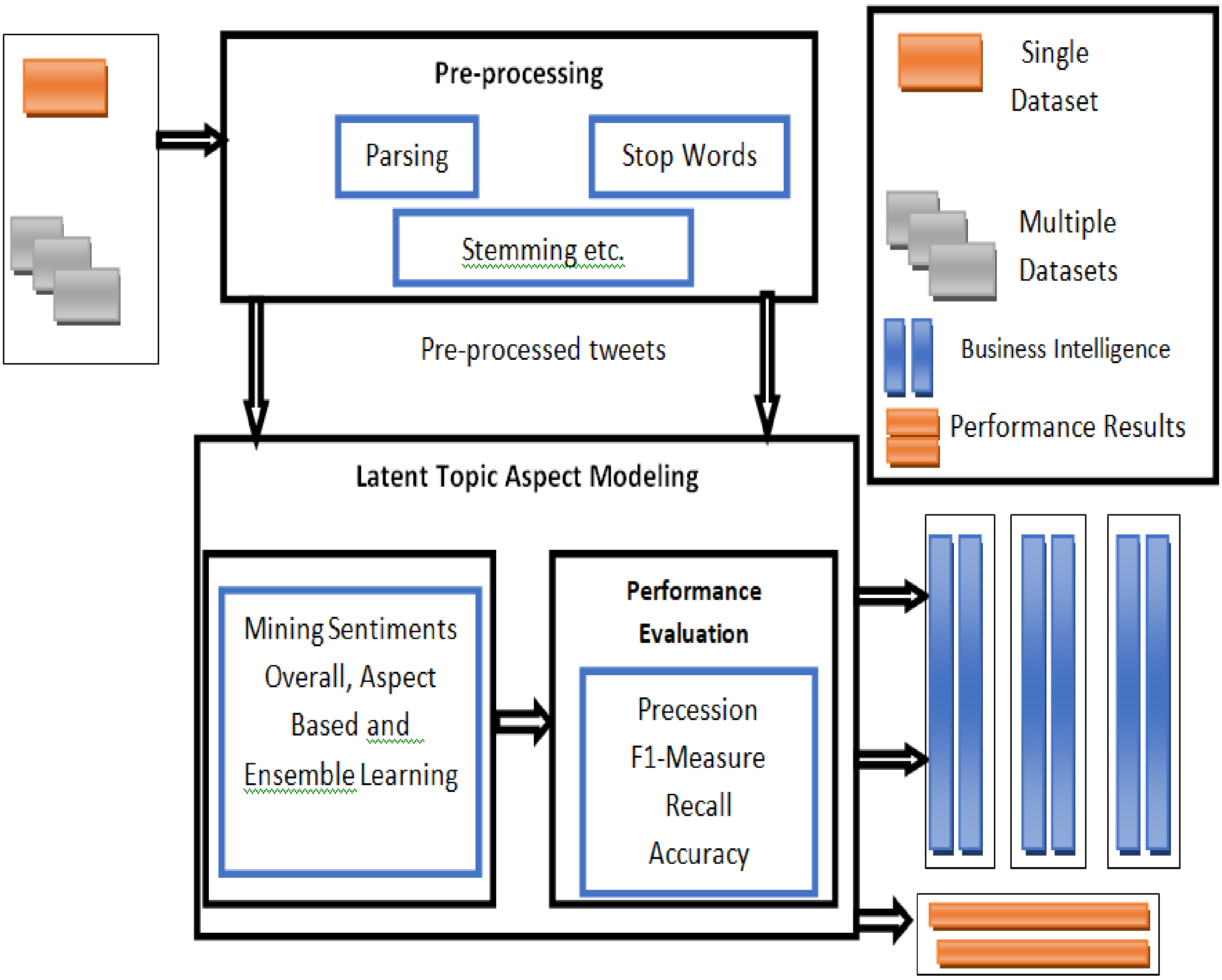


Figure 1: Proposed framework

As shown in Figure 1, after preprocessing, latent topic aspect modeling is used based on the generative process model named Latent DirichletAllocation (LDA). This model takes care of overall sentiment analysis and sentiments based on aspects. It is evaluated with performance metrics like precision, recall, accuracy and F1-measure.

B. Overall Prediction of Sentiments

Overall prediction of sentiments is made with the methodology illustrated in Figure 2. It has a strong training phase which helps in efficient learning process. Thus it provides higher level of accuracy in sentiment prediction.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

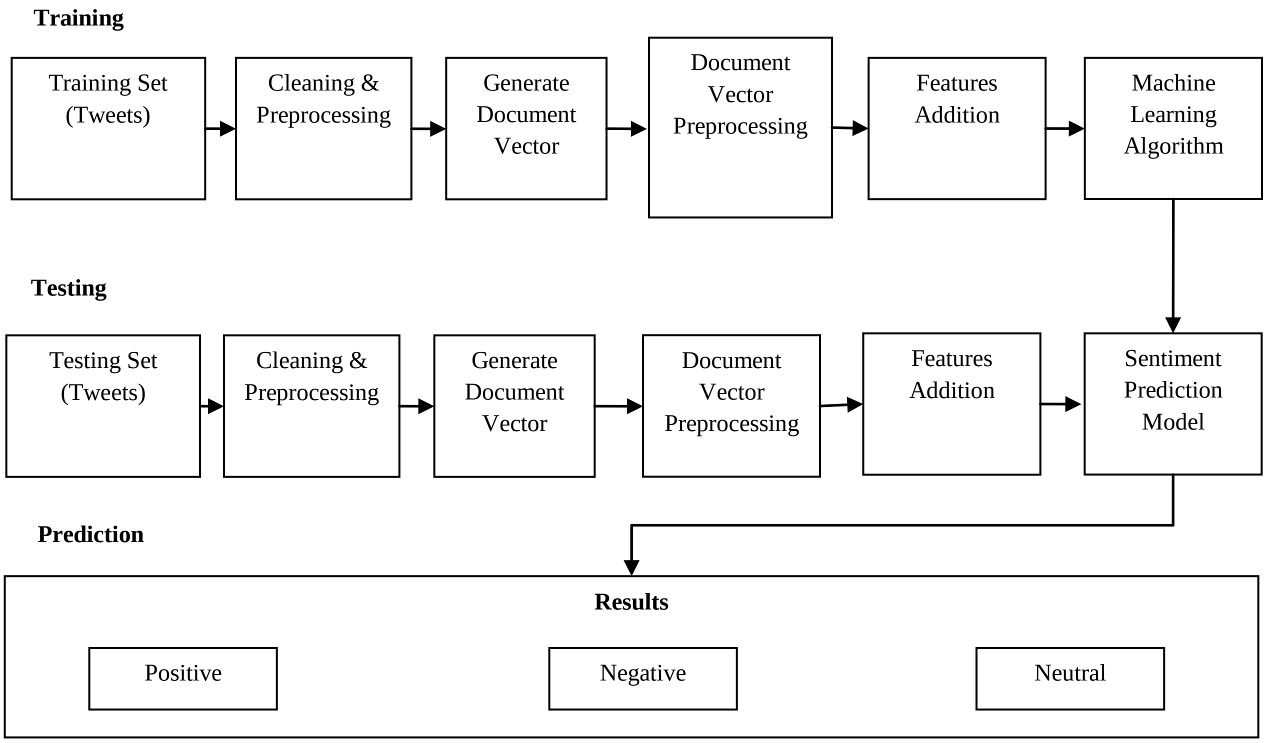


Figure 2: Methodology for overall prediction of sentiments

As presented in Figure 2, methodology is provided which includes two phases such as training phase and prediction phase. In the training phase training set is given as input. The training phase has many steps involved. They include cleaning and pre-processing, generation of document vectors, pre-processing of document vectors followed by features addition. Then the features are provided to a machine learning algorithm which learns and results in a sentiment prediction model. Once this model is built, it is used for prediction. A testing set is given as input to the prediction phase which has many steps similar to that of training phase. The steps include cleaning and processing, generation of document vectors, pre-processing of document vectors followed by features addition. Afterwards, the sentiment prediction model built in training phase predicts class labels for training set. The resultant classes will be positive, negative and neutral.

The cleaning and pre-processing involves various Natural Language Processing (NLP) functions. Lowercasing is used to convert text into lowercase. PorterStemmer is used for stemming to reduce search space. WordNetLemmatizer is employed for lemmatization. There are functions for stop word removal, removal of emoticons from tweets, removal of Unicode characters, removal of numbers, removal of URLs, removal of hashtags, removal of '@' sign, removal of punctuations and replacement of punctuations and contractions. Repeated exclamation marks, repeated question marks and repeated stop marks are replacing by words like MultiExclamationMarks, MultiQuestionMarks and MultiStopMarks respectively. Slang words and abbreviations are replaced. There are other functions for elongated word replacement, replacement of negations and word correction.

Three models are used for document vector generation. They are known as count model, TF-IDF model and word embeddings model. Word embeddings model is known as GoogleNews-vectors-negative300. For document vector pre-processing is done with dimensionality reduction and normalization. The former is made with two algorithms like Chi2 and Principal Component Analysis (PCA). An important phase which is used as part of training and prediction phases is features addition. Made with many features. Number of all words in a tweets (words), number of elongated words in a given tweet such as 'boooooring', number of exclamation marks, number of question marks, number of dots, number of capitalized words like TOMORROW and number of positive words. Lexicon of positive words is provided as input to the program. For negative words also, lexicon is provided. Slang lexicon is given as input for identifying slang words. List of positive and negative emoticons are provided as input to program. The features also include hash tags, URLs, '@' signs, nouns, verbs, adjectives, adverbs and Vader sentiment analysis vector. POS tagging is used to for finding number of nouns, verbs, adjectives and adverbs. After completion of features addition, the classifiers are learned. For this supervised learning purpose 10 widely used classifiers are employed. They are known as Naïve Bayes (NB), Random Tree (RT), Random Search (RS), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), XGB, SVM, Boosted Tree (BT) and ensemble learning classifier.

C.Ensemble Classifier

An ensemble classifier is built which is made up of multiple machine learning classifiers. This ensemble model is used for sentiment classification. The class labels are chosen based on the majority voting concept.

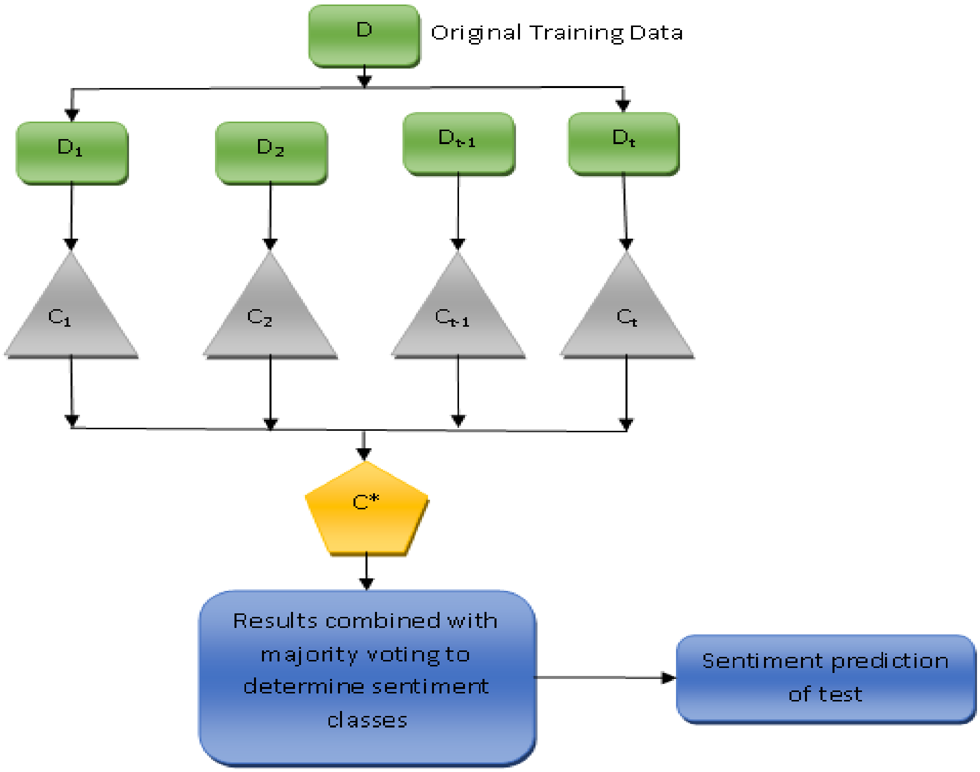


Figure 3: Ensemble model for sentiment prediction

As shown in Figure 3 the prediction model is made up of multiple ML models in AI domain. The tweets training set is divided into many parts randomly. Each portion of data is provided to a separate classification method to learn a classifier. The training data D is thus split into d1, d2, d3, ... and so on. These are assigned to classifiers like c1, c2, c3… and so on. The classifiers provide their decisions. The voting approach is used to finally combine results. The test data is assigned class labels based on the ensemble classification model. As set of negative words are used for the purpose of prediction.

> Abnormal, abolish, abominable, abominably, abominate, abomination, abort, aborted, aborts, abrade, abrasive, abrupt, abruptly, abscond, absence, absent-minded, absentee, absurd, absurdity, absurdly, absurdness, abuse, abused, abuses, abusive, abysmal, abysmally, abyss, accidental, accost, accursed, …

Listing 1: An excerpt from list of negative words used

As can be seen in Listing 1, the negative words are used as part of training and prediction. In the same fashion, a list of positive words is used for training and prediction purpose.

> Abound, abounds, abundance, abundant, accessible, accessible, acclaim, acclaimed, acclamation, accolade, accolades, accommodative, accommodative, accomplish, accomplished, accomplishment, accomplishments, accurate, accurately, achievable, achievement, achievements, achievable, acumen, adaptable, adaptive, adequate, adjustable, admirable, …

Listing 2: An excerpt from list of negative words used

As shown in Listing 2, a part of list of positive words are used in the empirical study. In Tweets there are many slang words coming frequently. In order to handle this kind of tweet, slang lexicon is used for effectiveness. A part of slang lexicon used for the study is as shown in Listing 3.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

2F4U	Too Fast For You
4YEO FYEO	For Your Eyes Only
AAMOF	As a Matter of Fact
ACK	Acknowledgment
AFAIK	As far as I know
AFAIR	As far as I remember / recall
AFK	Away from Keyboard
AKA	Also known as
B2K BTK	Back to Keyboard
BTT	Back to Topic
BTW	By the Way
B/C	Because
C&P	Copy and Paste
CU	See you
CYS	Check your Settings

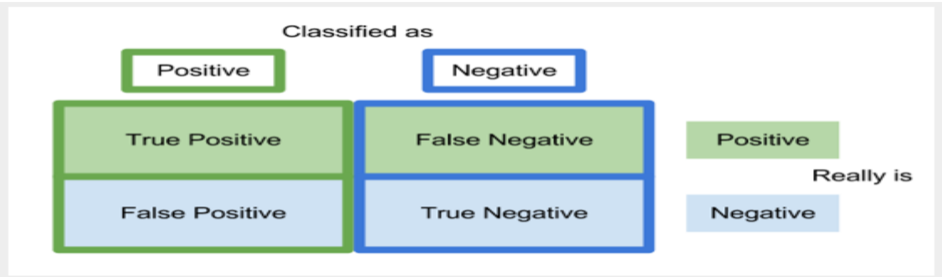
Listing 3: Part of slang lexicon used

As presented in Listing 3, there are many slang words used in the tweets. The list of slang lexicon is used to comprehend tweets in the learning process.

D. Evaluation Procedure

The proposed algorithms are evaluated using a standard approach based on confusion matrix. Confusion matrix helps in deriving multiple metrics in machine learning approaches. They are used in process mining for prediction problems as well.

Table 1: Shows confusion matrix



As shown in Table 1, confusion matrix provides different cases like TP, FP, FN and TN. These are used to derive performance metrics like precision, recall, F1 score and accuracy as in Eq. (1), Eq. (2), Eq. (3) and Eq. (4) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2.\frac{\text{precision} \cdot \text{recall}}{(\text{precision}) + \text{recall}} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

IV. DATASET DETAILS

Benchmark datasets are collected from the "International Workshop on Semantic Evaluation (SemEval)" on 2013, 2015, and 2016 [42]. Three datasets from events conducted in 2013, 2015 and 2016 respectively are merged together to form a bigger dataset for training. Table 1 shows details of datasets from which training data is generated.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 1: Training dataset details

SL. NO.	DATASET NAME	NO. OF INSTANCES	NO. OF ATTRIBUTES
1	SemEval 2013	9682	3
2	SemEval 2015	488	3
3	SemEval 2016	5831	3
Total number of instances after merging : 16061			

Each dataset contains three attributes. They are known as Tweet ID, Tweet's Polarity and Tweet. The polarity is qualitative in nature denoted by the strings such as Positive, Negative and Neutral. Highest number of instances (9682) are found in 2013 SemEval dataset. Total number of instances after merging is 16061.

```
628949369883000832      negative dear @Microsoft the newOoffice for Mac is great and all, but no Lync upd
C'mon.

628976607420645377      negative @Microsoft how about you make a system that doesn't eat my friggin discs.
is the 2nd time this has happened and I am so sick of it!

629023169169518592      negative I may be ignorant on this issue but... should we celebrate @Microsoft's par
leave changes? Doesn't the gender divide suggest... (1/2)

629179223232479232      negative Thanks to @microsoft, I just may be switching over to @apple.

629186282179153920      neutral   If I make a game as a #windows10 Universal App. Will #xboxone owner
able to download and play it in November? @majornelson @Microsoft

629226490152914944      positive  Microsoft, I may not prefer your gaming branch of business. But, you do m
a damn fine operating system. #Windows10 @Microsoft

629345637155360768      negative @MikeWolf1980 @Microsoft I will be downgrading and let #Windows1
out for almost the 1st yr b4 trying it again. #Windows10fail
```

Listing 4: Shows an excerpt of 2016 SemEval dataset

As presented in Listing 4, every tweet has an id, polarity such as positive, negative and neutral and actual tweet.

V. EXPERIMENTAL RESULTS

Experiments are made with the datasets described in Section 4. The proposed methodology is evaluated with different prediction models including the one based on the LDA for aspect orientation. The overall sentiment analysis, aspect based sentiment analysis and ensemble based model are evaluated and compared with the state of the art. Comparison is made with state of the art models like ASUM [43], JST [44], SLDA [45], Lexicon [41], SVM [41], Pooling [41] and SJASM [41].

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Ground Truth Results

id	sentiment
2.18775E+17	positive
2.58965E+17	neutral
2.62926E+17	negative
2.54949E+17	neutral
1.71874E+17	neutral
2.5601E+17	positive
2.61777E+17	neutral
2.64144E+17	neutral
2.23053E+17	neutral
2.64089E+17	positive
2.64034E+17	neutral
2.5864E+17	positive
2.50981E+17	neutral
1.95555E+17	neutral
2.60536E+17	neutral
2.63182E+17	neutral
2.62435E+17	neutral
2.56514E+17	neutral
2.63734E+17	positive
2.41403E+17	neutral

Predicted Results

id	sentiment
2.18775E+17	positive
2.58965E+17	neutral
2.62926E+17	negative
2.54949E+17	neutral
1.71874E+17	neutral
2.5601E+17	positive
2.61777E+17	neutral
2.64144E+17	neutral
2.23053E+17	neutral
2.64089E+17	positive
2.64034E+17	neutral
2.5864E+17	positive
2.50981E+17	neutral
1.95555E+17	neutral
2.60536E+17	neutral
2.63182E+17	neutral
2.62435E+17	neutral
2.56514E+17	neutral
2.63734E+17	positive
2.41403E+17	neutral

Figure 4: Expert from results comparing ground truth and predicted values

As shown in Figure 4, a part of the predicted results and corresponding ground truth values are used for comparison. In the first 20 training tweets, the prediction results are provided.

Table 1: Performance comparison in terms of overall sentiment prediction

Prediction Model	**Performance (%)**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
Naive Bayes	62.7	55.6	43.3	55.6
Random Tree	62.54	61.55	54.7	61.5
Random Search	51	46	54	60
Logistic Regression	65.7	66.66	62.29	66.6
Random Forest	76	80	70	80
Decision Tree	55.9	56	55.3	56
XGB	75	74	71	80
SVM	63.14	64.63	58.89	64.6
Boosted Tree	65	65.6	60.3	65.6
Ensemble Learning Voting Classifier	64.6	58	58	65.7

As presented in Table 1, many prediction models such as NB, RT, RS, LR, RF, DT, XGB, SVM, BT and ensemble learning classifier are evaluated. The performance is observed in terms of precision, recall, F-measure and accuracy.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

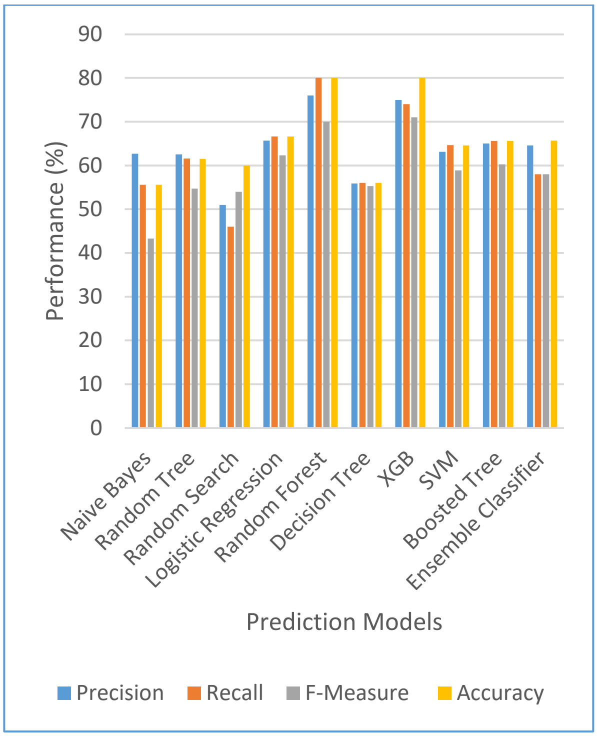


Figure 5: Overall sentiment prediction models based on proposed training model

As presented in Figure 5, the results of prediction models that were trained by the proposed training phase are provided. The 10 prediction models aforementioned in this paper are provided in horizontal axis. Vertical axis shows performance in terms of precision, recall, F-measure and accuracy. The results are obtained with a standard 10-fold cross validation using stratified sampling. The prediction models showed varied performance. Performance metrics described in section 4 are used to obtain results. Highest precision is shown by RF with 76%. Least value for LR is exhibited by RS with 51%. Ensemble learning method showed better performance over all other models except the RF, XGB and LR. Highest recall is exhibited by RF with 80%. Least recall percentage is shown by RS with 46%. Highest F-score is exhibited by XGB with 70%. Naïve Bayes showed least performance with 433% F-score. Highest accuracy is shown by RF and XGB with 80% while the least accuracy is exhibited by NB with 55.6%.

Table 2: Computation time

Prediction Models (%)	Computation Time (seconds)
Naive Bayes	312
Random Tree	334
Random Search	345
Logistic Regression	354
Random Forest	342
Decision Tree	335
XGB	338
SVM	356
Boosted Tree	354
Ensemble Learning Voting Classifier	365

As shown in Table 2, the computation time of all prediction models that exploit the proposed training model is provided.

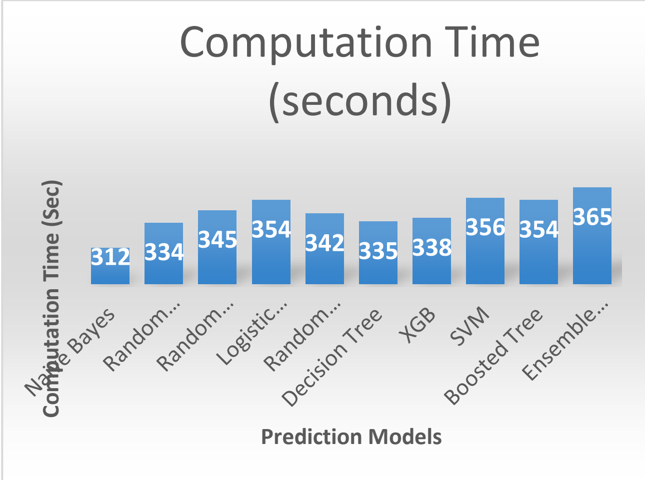


Figure 6: Comparison of prediction models in terms of computation time

As presented in Figure 6, the computation time is observed as presented in vertical axis. The horizontal axis shows different techniques used for sentiment prediction. Naïve Bayes showed least execution time that is 312 seconds while the ensemble learning model has taken highest execution time with 365 seconds.

Table 3: Shows overall sentiment prediction accuracy vs. aspect number

Prediction Model	Overall Sentiment Prediction Accuracy (%) against Aspect Number							
	AN 5	**AN 10**	**AN 15**	**AN 20**	**AN 25**	**AN 30**	**AN 35**	**AN 40**
Lexicon	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
ASUM	0.76	0.78	0.77	0.74	0.74	0.73	0.73	0.73
JST	0.78	0.78	0.8	0.78	0.77	0.79	0.77	0.77
SVM	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Pooling	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
SLDA	0.82	0.84	0.83	0.82	0.82	0.81	0.79	0.79
SJASM	0.88	0.863	0.861	0.862	0.86	0.86	0.85	0.85
ABTM-SA	**0.94**	**0.92**	**0.91**	**0.89**	**0.89**	**0.89**	**0.87**	**0.87**

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

As presented in Table 3, the overall sentiment prediction accuracy of different prediction models against various aspect numbers is provided. The accuracy is measured in percentage.

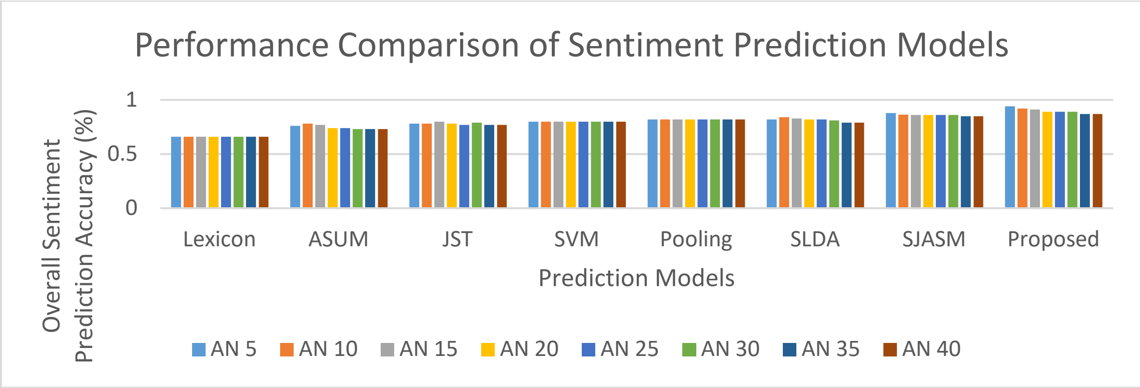


Figure 7: Overall sentiment analysis vs. aspect number

As shown in Figure 7, various prediction models are presented in horizontal axis while the vertical axis shows the overall prediction accuracy (%) against different aspect numbers (shown as legend). The aspect numbers at which observations are captured include 5, 10, 15, 20, 25, 30 and 35. The aspect number has no influence on the prediction models such as Lexicon, SVM and Pooling. The prediction accuracy of these models is 0.66, 0.8 and 0.82 respectively for all given aspect numbers. The least performance is exhibited by Lexicon while the proposed method showed highest performance. The values in bold indicate highest accuracy for a given aspect number. SJASM showed highest performance then all other models expect the proposed model. SJASM exhibits prediction accuracy values 0.88, 0.863, 0.861, 0.862, 0.86, 0.86, 0.85, and 0.85 respectively for AN 5, AN 10, AN 15, AN 20, AN 25, AN 30, AN 35 and AN 40. The proposed prediction model showed accuracy performance values 0.94, 0.92, 0.91, 0.89, 0.89, 0.89, 0.87 and 0.87 respectively for different aspect numbers.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a comprehensive framework for intelligent sentiment analysis on big data. It is NLP ad machine learning based with supervised learning models and also generative process model such as LDA based approach. It focuses on overall sentient prediction and also a joint topic model that extracts sentiments and also aspects simultaneously. It also has provision for an ensemble method that uses multiple classifiers' results with voting. An algorithm by name Aspect Based Topic Modelling for Sentiment Analysis (ABTM-SA) is proposed and the results are compared with the state of the art such as ASUM, JST, SLDA, Lexicon, SVM, Pooling [41] and SJASM [41]. The proposed model showed better performance over these models. The rationale behind this is that it exploits the proposed training model for improving overall sentiment prediction accuracy against aspect number. The proposed framework also exploits the training model to train machine learning based prediction models such as NB, RT, RS, LR, RF, DT, XGB, SVM, BT and ensemble learning classifier. The steps in training phase helps to improve performance of these classifiers and the proposed topic-aspect based model as well. The empirical study revealed that highest accuracy is exhibited by Random Forest model with 80% while the least accuracy 55.6% is exhibited by Naïve Bayes. These observations are related to overall sentiment analysis without aspects. The proposed ABTM-SA algorithm which simultaneously uses aspects and analyze overall sentiments showed highest performance with 0.94 and 0.87 when aspect number is minimum (5) and maximum (40) respectively. Against all aspect numbers, the proposed algorithm showed better performance over the state of the art methods such as Lexicon, ASUM, JST, SVM, Pooling, SLDA and SJASM. In future we intend to model meta data for achieving spatio-temporal sentiment analysis. Another direction for our future work is to explore deep learning for improving performance of sentiment analysis with large volumes of data or big data.

REFERENCES

1. B. Liu, (2012) "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167.
2. B. Yang and C. Cardie, (2014) "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, p325–335.
3. Lin, Changhua, He, Yulan, Everson, Richard and Ruger, Stefan (2012). Weakly-supervised joint sentiment-topic detection from text. IEEE Transactions on Knowledge and Data Engineering, 24(6) p1134–1145.
4. S. Kim, J. Zhang, Z. Chen, A. Oh, and S. Liu, (2013) "A hierarchical aspect-sentiment model for online reviews," in Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, ser, p526–533.
5. M. Dermouche, J. Velcin, L. Khouas, and S. Loudcher, "A joint model for topic-sentiment evolution over time," in 2014 IEEE International Conference on Data Mining, p773–778.
6. Z. Yang, A. Kotov, A. Mohan, and S. Lu, "Parametric and nonparametric user-aware sentiment topic models," in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, p1-10.
7. D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joi. P1014–1023.
8. X. Hu, L. Tang, J. Tang, and H. Liu, (2013) "Exploiting social relations for sentiment analysis in microblogging," in Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, p537-546.
9. Z. Hai, K. Chang, and G. Cong, (2012) "One seed to find them all: mining opinion features via association," in Proceedings of the 21st ACM international conference on Information and knowledge management, , p255-264.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

10. Z. Hai, G. Cong, K. Chang, W. Liu, and P. Cheng, (2014) "Coarse-to-fine review selection via supervised joint aspect and sentiment model," in Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. P1-10.
11. Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, (2013) "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Transactions on Knowledge and Data Engineering, 99, p1-12.
12. Mohamed Dermouche, Leila Kouas, Julien Velcin and Sabine Loudcher. (2015) A Joint Model for Topic Sentiment Modeling from Text. 30th ACM/SIGAPP Symposium On Applied Computing, Salamanca, Spain. P1-7.
13. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications. P1-43.
14. Cao, D., Ji, R., Lin, D., & Li, S. (2014). Visual sentiment topic model based microblog image sentiment analysis. Multimedia Tools and Applications, 75(15), P8955–8968.
15. Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., & Zhou, M. (2016). Sentiment Embeddings with Applications to Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering, 28(2), P496–509.
16. KarWai Lim, Changyou Chen and Wray Buntine. (2016). Twitter-Network Topic Model a Full Bayesian Treatment for Social Network and Text Modeling, p1-6.
17. AsbjørnOttesenSteinskog Jonas FoynTherkelsenBjörnGambäck. (2017). Twitter Topic Modeling by Tweet Aggregation. Proceedings of the 21st Nordic Conference of Computational Linguistics, p77–86.
18. Chen, J., Li, H., Wu, Z., & Hossain, M. S. (2017). Sentiment analysis of the correlation between regular tweets and retweets. 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA). P1-5.
19. AnumolBabu and Rose V Pattani. (2016). Efficient Density Based Clustering of Tweets and Sentimental Analysis Based on Segmentation. International Journal of Computer Techniques. 3 (3), p53-57.
20. Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis. ACM Transactions on Management Information Systems, 9(2), 1–29.
21. Yanchang Zhao. (2016). Twitter Data Analysis with R – Text Mining and Social Network Analysis, p1-40.
22. Suresh, H., &Gladston Raj S. (2016). An unsupervised fuzzy clustering method for twitter sentiment analysis. 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). P1-6.
23. Kovacs-Gyori, A., Ristea, A., Havas, C., Resch, B., And Cabrera-Barona, P. (2018). #London2012: Towards Citizen Contributed Urban Planning Through Sentiment Analysis of Twitter Data. Urban Planning, 3(1), P75-99.
24. [24] Silva, N. F. F. D., Coletta, L. F. S., &Hruschka, E. R. (2016). A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning. ACM Computing Surveys, 49(1), P1–26.
25. Lan, M., Zhang, Z., Lu, Y., & Wu, J. (2016). Three Convolutional Neural Network-based models for learning Sentiment Word Vectors towards sentiment analysis. 2016 International Joint Conference on Neural Networks (IJCNN). P1-8.
26. Yan, Y., Yang, H., & Wang, H.-M. (2017). Two simple and effective ensemble classifiers for Twitter sentiment analysis. 2017 Computing Conference. P1-8.
27. Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G., and Pranata, R. (2018). Twitter Topic Modeling on Football News. 2018 3rd International Conference on Computer and Communication Systems (ICCCS). P1-5.
28. Yoon, H. G., Kim, H., Kim, C. O., And Song, M. (2016). Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. Journal of Informatics, 10(2), 634–644.
29. Duric, A., And Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. Decision Support Systems, 53(4), 704–711.
30. Thien Hai Nguyen and KiyoakiShirai. (2015). Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. International Joint Conference on Natural Language Processing, p1354–1364.
31. Poria, S., Chaturvedi, I., Cambria, E., &Bisio, F. (2016). Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. 2016 International Joint Conference on Neural Networks (IJCNN). P1-9.
32. Shams, M., Shakery, A., And Faili, H. (2012). A non-parametric LDA-based induction method for sentiment analysis. The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012). P1-6.
33. Liang, J., Liu, P., Tan, J., And Bai, S. (2014). Sentiment Classification Based on AS-LDA Model. Procedia Computer Science, 31, P511–516.
34. Lu, B., Ott, M., Cardie, C., And Tsou, B. K. (2011). Multi-aspect Sentiment Analysis with Topic Models. 2011 IEEE 11th International Conference on Data Mining Workshops. P1-8.
35. Bing Xiang and Liang Zhou. (2014). Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, p434–439.
36. Ding, W., Song, X., Guo, L., Xiong, Z., And Hu, X. (2013). A Novel Hybrid HDP-LDA Model for Sentiment Analysis. 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). P1-8.
37. Changhua Lin. (2011). Probabilistic topic models for sentiment analysis on the Web. Department of Computer Science, p1-135.
38. Cambria, E., Poria, S., Bisio, F., Bajpai, R., And Chaturvedi, I. (2015). The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis. Lecture Notes in Computer Science, P3–22.
39. Haochen Zhou and Fei Song. (2015). Aspect-Level Sentiment Analysis Based on a Generalized Probabilistic Topic and Syntax Model. proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, p241-244.
40. Kentaro Sasaki, Tomohiro Yoshikawa and Takeshi Furuhashi. (2014). Online Topic Model for Twitter Considering Dynamics of User Interests and Topic Trends. Conference on Empirical Methods in Natural Language Processing, p1977–1985.
41. Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng and Chunyan Miao. (2017). Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach. IEEE Transactions on Knowledge and Data Engineering. 29 (6), p1-14.
42. SemEval (2016). International Workshop of Semantic Evaluation (SemEval-2016). Available at: http://alt.qcri.org/semeval2016/
43. S. Moghaddam and M. Ester, "Ilda: Interdependent lda model for learning latent aspects and their ratings from online product reviews," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR'11, 2011, pp. 665–674.
44. M. M. Rahman and H. Wang, "Hidden topic sentiment model," in Proceedings of the 25th International Conference on World Wide Web, ser. WWW '16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 155–165.
45. G. Paltoglou and M. Thelwall, "Twitter, myspace, digg: Unsupervised sentiment analysis in social media," ACM Trans. Intell. Syst. Technol., vol. 3, no. 4, pp. 66:1–66:19, Sep. 2012.
46. Nafees Akhter Farooqui, Ritika, Aayush Saini. (2019). Sentiment Analysis of Twitter Accounts using Natural Language Processing. International Journal of Engineering and Advanced Technology. 8 (3), p473-479.
47. Abdullah Alsaeedi, Mohammad Zubair Khan. (2019). A Study on Sentiment Analysis Techniques of Twitter Data. International Journal of Advanced Computer Science and Applications,. 10 (2), p363-374

AUTHORS PROFILE

D. Krishna Madhuri Received her B.Tech degree in computer Science and engineering from Swarnandhra college of Engineering & Technology, Narsapur, Andhra Pradesh in 2009, and M.Tech Degree in Computer Science and Engineering from Sridevi Women's Engineering College Hyderabad, Telangana in 2012. She is currently pursuing her Ph.D degree from Sri Satyasai University, Sehore, Madhya Pradesh. Currently she is working as Assistant professor in the department of Computer Science and Engineering of GokarajuRangaraju Institute of Engineering and Technology, Hyderabad, Telangana India. Her research interest includes Data Bases, Data warehouse and Data Mining and Big Data.

Dr. RVVSV. Prasad received his Master's degree in Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada and obtained doctoral degree from Department of Computer Science & Engineering, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, India. He is working as a Professor and Head in the Department of Information Technology, Swarnanadhra College of Engineering & Technology, Seetharampuram, Narsapur, India. His research interests include machine learning, big data analytics, recommender Systems, P2P networks, Security, Privacy and trust in mobile agents.

Retrieval Number: D9062019420/2020©BEIESP
DOI: 10.35940/ijitee.D9062.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)