International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp.3691-3702

ANALYSIS OF TWITTER DATA USING LOGISTIC REGRESSION CLASSIFICATION BASED MACHINE LEARNING METHOD

M Divyapushpalakshmi[1] and R Ramalakshmi*

Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamilnadu, India.

Abstract: Web innovation has shown great improvement and there is an enormous extent of information present in the web for web utilizers and many data is delivered also. Web has turned into a stage for internet getting the hang of, trading thoughts and imparting insights. Twitter is one of the eminent social networking site, which examine this work where people reveal their personal information, express their views and notions, interests, likes and dislikes in the site with or without knowledge. To identify the users opinion such as positive and negative comments the system required machine-learning techniques. For this purpose, this work concentrated on novel supervised machine learning based classification technique. To obtain high classification accuracy, the proposed system introduced with preprocessing and k-means clustering approach. parameters like F-Measure, accuracy, recall and precision are the most important factors for a social network analysis system. Our proposed system produced better results when compared with the existing approaches.

Keywords: Twitter, machine learning, preprocessing, clustering, Training and learning.

1. Introduction

Interpreting human language has attracted numerous researches in the recent times. The affluence of internet access had increased the social network penetration across the globe. Understanding human thoughts and deciphering the language has attracted a huge demand for various reasons. Natural language processing (NLP) serves as an important tool for intelligent systems that analyze, portray and decipher human language. The exploration of Natural language processing has formed through the group processing and time of punch cards, where the scrutiny for a single sentence might progress around 7 minutes. With technological surge and emergence of Google, limitless analysis and arrangements were made in few seconds [1]. With the help of NLP, computers or laptops processes a comprehensive assortment of natural language associated errands of any kind measurements to indulge the device by describing and linguistic structure naming. Fabricated intelligence structures and figuring have adequately made significant advances in fields for instance PC visualization and model acknowledgement. At present continuous NLP progressively concentrating on the utilization of new AI techniques [2]. Machine learning methodologies

ISSN: 2233 -7857 IJFGCN

are astounding decisions for direct conditions, differential conditions, and stochastic models to foresee possible destiny of time course of action.

In any case, there are still challenge to build precision from one side and presenting exhaustive arrangements from different sides. Artificial neural network as one of the amazing administered learning techniques connected numerous zones including time arrangement anticipating [3].

Sentiment arrangement implies the usage of customary language getting ready, content examination and computational semantics to recognize and remove dynamic data from resource equipment.Sentiment grouping intends to pick the demeanor of a speaker regarding some point or the all things considered coherent limit of a manuscript, for instance, 'positive' or 'negative' and 'endorsement' or ' dissatisfaction ' [14]. Lexicon and corpus methods preferred with the techniques of report sentiment characterization [15]. The measure of sentiment for substance dependent on sentiment expressions rely on the vocabulary based methodologies. The corpus-based methodologies incorporate an accurate characterization technique. Unsupervised learning uses corpus-based methods that beat the word reference based methodologies utilized in administered learning, semi regulated learning and unsupervised learning.With the consistent improvement of social systems administration sites, the volume of social media information has detonated and the client created content is ending up increasingly various. Accordingly, the methodology of gigantic social media information is never kept to the single text mode. In microblog sites, for instance, an ever-increasing number of clients are slanted to post multi model tweets, including an image in their tweets, which conveys new difficulties to social media examination in giving large-scale social media information and its multi model structure. Accept sentiment analysis for instance field of social media analysis. Breaking down client sentiments depends increasingly more on large measure of multi model substance, as opposed to the customary text-based sentiment analysis. Therefore, multi model sentiment analysis has turned into an inexorably significant research theme as of late, particularly in the context of social media huge information.

They can fill in as incredible online correspondence stages that enable a great many clients to deliver, spread, offer, or trade information whenever and wherever. Such information normally incorporates multimedia content, for example, content, picture, and video. The gigantic measure of multimedia data dispersion on shared medium infer prosperous learning and spread an extensive range of social elements happening over the world on a remarkable range and progressively. This marvel gives incredible chances to address significant issues, which appear to be difficult to settle in the past [4].

For instance, the utilization of common medium information has displayed an enormous prospective in different applications, for example, recognizing breaking news, spreading news, planning salvage endeavors, partaking in nearby events, following a game, and picking up situational mindfulness amid a

ISSN: 2233 -7857 IJFGCN

crisis. On the other hand, the viable utilization of social media information is testing the result of its quick changing rate, heterogeneity and huge volume [5].

The qualities of social media data present incredible difficulties to customary AI strategies, with, homogeneous, organized and scalable data. Enthusiasm for research on new strategies of AI with social media data is exponentially increasing. Analysts have led AI strategies considering the distributed gatherings and diaries crosswise over various fields, for example, perception social registering and data mining to comprehend data fetched from social media. The recent examinations on AI systems has resulted profound outcomes and discoveries, which proves the achievements in addition to the adequacy of utilizing AI strategies for managing complex social media data. Alongside these examinations, a wide range of visual investigation systems has been developed to deal with social media data [6]. Logistic Regression is a standout amongst the most dominant AI strategies with no meta-parameter is connected in this paper.

2. Literature survey

Shenghua Liu (2013) proposed a semi-supervised point versatile supposition order model (TASC), the model initiates with a classifier that depends on the mixed named information from several subjects and ordinary features. TASC learning calculation refreshes subject versatile highlights dependent on the collaborative determination of unlabeled information, which thus chooses increasingly solid tweets to support the exhibition. Besides, adjusting model along a course of events (TASC-t) for dynamic tweets is additionally planned. The proposed model TASC, outperforms the grasped oversaw and troupe classifiers. The system also performs better than the semi-managed learning methodologies that are without highlight adaptation. The model TASC-t for dynamic tweets results in accomplishing basic precision and F-score [7].

Sreekanth Madisetty et al (2018) proposed a troupe method for spam acknowledgment at tweet level. The proposed system is an emsemble of five CNNs and a feature based extraction model. To set up and train the model each CNN utilizes diverse word embedding techniques such as Word2vec, Glove and more. The n-gram highlights, content based and user-based features are used by the feature based model. The combination of the two models acts simillar to that of a meta classifier in the proposed multi layer neural network. Low dimension results in low computational speed in the hybrid system [8].

Tu manshu *et al*, (2019) proposed a various leveled consideration network with earlier learning data (HANP) for the cross-area opinion order (CDSC) task. The HANP can acquire both space free and area explicit highlights in the meantime by including earlier learning. Moreover, the HANP additionally incorporates a progressive portrayal layer with consideration system, therefore the HANP be able to grab

ISSN: 2233 -7857 IJFGCN

substantial disputes and verdicts linking towards approximation.Thus the examination of datasets determines that innovative methods might be worn-out by the proposed HANP [9].

Zhao Jianqiang *et al*, (2017) present a word embedding's gotten by unsupervised learning on huge twitter corpora that utilizes idle relevant semantic connections and co-event factual qualities between words in tweets. These word embedding has joined with n-grams highlights and word slant extremity score highlights to frame an assessment include set of tweets [10]. The list of capabilities is incorporated into a profound convolution neural network for preparing and foreseeing slant-grouping names. This model performs better on the exactness and F1-Measure for Twitter feeling arrangement.

Mondher bouazizi et al, (2017), nevertheless the previously mentioned errands of arrangements, proceeds from the data collected from writings of Twitter and commands the writings to numerous supposition modules [11] proposed a novel methodology. In multi-class grouping, the proposed methodology attains accuracy upto 60.2%. In any case, the methodology demonstrates to be exact in twofold arrangement and ternary grouping: in the previous case, we achieve an exactness of 81.3% for similar informational collection utilized in the wake of expelling unbiased tweets, and in the last case, we achieved a precision of order equivalent to 70.1%.

Dong Deng *et al*, (2018) proposed a novel various leveled supervision theme model to develop a subject versatile feeling vocabulary (TaSL) for more elevated amount order task. Records are spoken to by different sets of points and estimations, in which individual pairs is labeled by a multinomial dissemination over words. In the meantime, this creating procedure is directed under various leveled supervision data of records and words. The principle-preferred standpoint of TaSL is the conclusion extremity of each word in various points caught adequately.It is advantageous to develop an area explicit opinion dictionary and afterward adequately improve the presentation of conclusion classification [12].

Meng Wang *et al*, (2017) projected a refined algorithm subject to significant learning and data geometry in which the dissemination of all availability tests in the space is treated as before information and is encoded by Deep Belief Networks (DBN). The geodesic parcel is prepared among the scatterings above the highlights on the view of data geometry. Thus, the committed observational datasets for supposition classification has been utilized [13].

3. Proposed methodology

This methodology concentrates of Logistic regression classification based supervised machine learning method. Firstly, we discuss the preprocessing of the tweets by using three important steps. These steps

ISSN: 2233 -7857 IJFGCN

make the tweets to be short and effective. Secondly, we proposed a novel classification method in supervised model.For better results K-means clustering has been used.

3.1 *Preprocessing*

In this stage, unstructured content information employed and processed data is utilized for element extraction**.** Consequently, information cleaning is must achieve a decent yield**.** This stage have been separated on numerous sub stages

1**.** Tokenization**-** Content report has a cluster of sentences, which is divided into individual words termed as tokens, accentuation characters, white spaces and other language preprocessing techniques**.** .Before tokenization the all the letters in the tweets are converted to lower case

2**.** Gathered remarks are sent for identification of estimation terms through part of speech tagging**.** Speech tagging is used to decide the estimation terms**.**

3**.** Word sense disambiguation **(**WSD**)** is a complex assignment in the domain of Regular language processing **(**NLP**).** It submits to the assignment that robotically allocates the suitable logic, chosen from an arrangement of pre**-**defined logic for a statement, according to an exacting content**.** Here is the theory of having the information appropriately pre**-**processed**:** to diminish the commotion in the content should help improve the exhibition of the classifier and accelerate the classification procedure, in this manner supporting continuously sentiment analysis**.**

Conditions to extracts the sentences from tweets are described below**:**

- Tweets contain information which are not clearly expressive or express significance and ought to be removed**.**
- Remove unfortunate accentuations**:** All accentuations which are a bit much, it has been cleared**.**
- Stop Word Removal**:** Some words used progressively additional time such words are called stop word**.** This pronouns, relational words, conjunctions have no specific significance**.**
- **"**i**","**a**","**an**","**is**","**are**","**as**","**at**","**from**","**in**","**this**","**on**","**or**","**to**","**was**","**what**","**will**","** with**"** , and so forth are case of stop word, so these kinds of words has been disappeared**.**
- Stemming**:** This pre**-**processing renovate opinion term word keen on its grammatical origin form**.** This procedure translate word similar to **"**process**"**, **"**processing**"**, **"**processed**"**, **"**procession**"**, **"**processor**"** to origin word process**.**
- Part of Speech Tagging**:** The Part**-**Of**-**Speech of a text is a linguistic that is related to a grammatical English language to be exact describe by morphological or syntactic activities**.** The general clusters

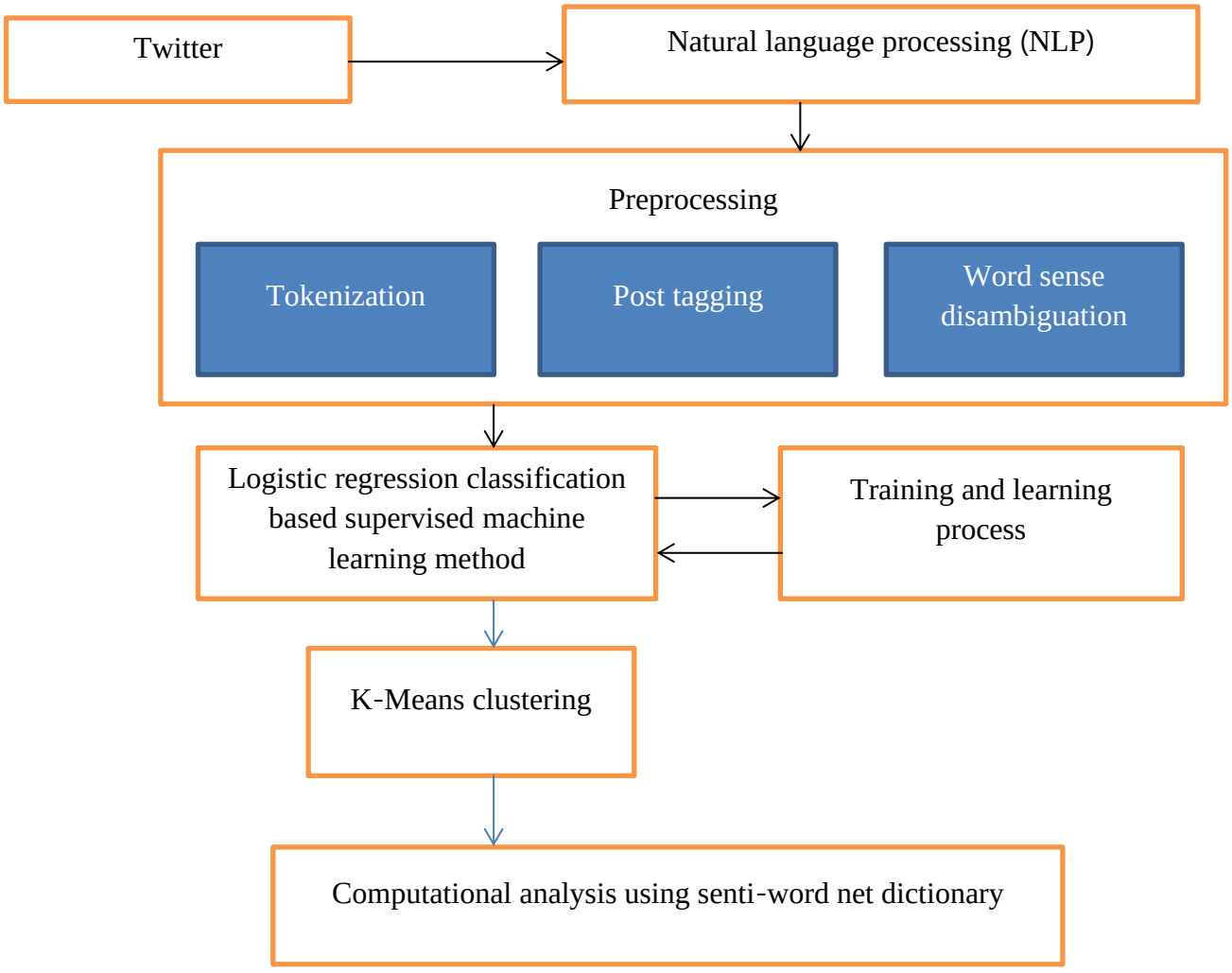
ISSN: 2233 -7857 IJFGCN

of Part of speech are Verb, pronoun, Noun, preposition, adjective, conjunction, interjection and adverb**.**

The Dimensionality diminution and Word Sense Disambiguation are executed by utilizing supervised machine learning. Training process of the proposed system or approach is accomplished as follows:

1. First step to change dispute a text word into unusual character $
2. In second step overwhelming the rest of text words in the given sentence.

This approach teaches a model, to assume and pick out a text word in a huge number of unlabelled prepared information with given neighbouring context. This enormous training dataset permit us to instruct a high-capacity model layer such as hidden or context. This directional approach has more rapidly and easy going to instruct our vast input dataset than a bidirectional process.



ISSN: 2233 -7857 IJFGCN

3.2 Training and learning process

In regression analysis, free factor is otherwise called relapse or indicator while the needy variable is known as relapsed or clarified variable. Logistic regression analysis is a system used for assessing the dark estimation of a destitute variable from the known estimation of self-ruling variable. By the day's end, X and Y are two related variables, by then direct relapse methodology gauges the estimation of Y for a given estimation of X. Likewise, gauge the estimation of X for given estimation of Y.

In the event that the class-restrictive densities are Gaussian and share a typical covariance framework, the log probability proportion can be spoken to as a direct capacity of test vector X [2]. For the instance of two classes, we have,

$$\text{Log } (p(x/c1))/ (p(x/c2)) = w(t).x + m \quad (1)$$

Where, w (t) refer to a weight vector and m refer to a bias

By using the bayes rule,

$$\text{Log } (p(c1/x)) /1\text{-}p(c1/x)) = \text{Log } p(c1/x)\ p(c2/x)$$
$$= \text{Log}(p(x/c1))/(p(x/c2)) + \text{Log}(p(c1)/(p(C2)) \quad (2)$$

Now,

$$p(c1) = 1/1+e^{\wedge}\text{-}\ (w(t).x+w(0)) \quad (3)$$

It can be generalized to k>2 classes. Taking one of the classes, for example, c(k) as the reference class. Then it is given as,

$$\text{Log } (p(ci/x))/ (1\text{-}p(cj/x)) = \text{Log } p(ci/x)/ (p(ck/x)) \quad (4)$$

The strategic relapse of multi-target arrangement utilizes this equation to ascertain the likelihood of test x having a place with classification c_i. The weight lattice and the predisposition vector are the parameters of this model

4. Results and Discussion

The proposed approach is implemented using the programming language Python. All data-sets are trained to scikit-learn's compatible ML model using a Core i7 processor with extreme pipelining along with 16 Giga Bytes of RAM and 64 bit processor (64-bit) is important due to floating point precision and it's memory space is quad-word) along with 64-bit Python. For a system with lower configuration use pre-built, model available over the internet as open source. In this simulation, 80% of the twitter dataset are taken for training the proposed classifier and 20% of the dataset are taken for testing the classifier. Using the training twitter dataset, pre-processing is done and the classifiers supervised learning with logistic regression, decision tree and random forest are trained.

ISSN: 2233 -7857 IJFGCN

4.1. *Performance metrics*

Depend on the positive, negative and neutral sentiment classes, the performance metrics accuracy, F-measure, recall and precision are evaluated. In addition, these metrics are defined with the False positive (FP), True Positive (TP), False Negative (FN) and True Negative (TN). The performance metrics are defined as follows.

a) ***Accuracy***

The number of text retrieved exactly in given datasets. It is represented as,

$$\text{Accuracy} = TN + TP / TN + TP + FN + FP * 100 \quad (5)$$

b) ***Precision***

The precision is calculated as follows:

$$\text{Precision} = TP / FP + TP \quad (6)$$

Precision is characterized as a calculation of accuracy or feature, though review is a calculation of culmination or capacity. In addition, high accuracy demonstrates that the methodologies returned altogether more applicable outcomes than insignificant.

c) ***Recall***

The count of the recall esteem is done as pursues:

$$\text{Recall} = TP / FN + TP \quad (7)$$

Recall is depicted as the quantity of important reports recouped through a pursuit isolated by the all-out number of available applicable archives. Recall is additionally the quantity of genuine positives isolated through the complete number of components that successfully have a place with the positive class.

d) ***F-measure***

It processes the consolidated estimation of accuracy and review as the mean constant of exactness and the review. The value of f-measure esteem is acquired as pursues

$$\text{F - Measure} = (\text{Recall} * \text{Precision})^2 / (\text{Recall} + \text{Precision}) \quad (8)$$

4.2. *Analysis of performance of the proposed system*

ISSN: 2233 -7857 IJFGCN

This section enlightens the performance of the logistic regression based sentiment analysis in comparison with that of the traditional classifiers such as Random Forest and Decision tree driven sentiment analysis used for varying iterations. Figures 2-5 show the evaluation of the performance metrics for different approaches. Figure 2 shows the comparison of precision of different approaches. As shown in the figure, the proposed logistic regression attains 64% of precision while decision tree and random forest attain 63.25% and 62.25% of precision respectively. The comparison of recall of different approaches is shown in figure 3. As shown in the figure, the proposed logistic regression obtains 55.98% of recall than the existing classifiers. Figure 4 shows the comparison of F-measure of the different classifiers. As the proposed logistic regression attains better precision and recall, F-measure of logistic regression is also increased. Namely, the proposed logistic regression attains 59.4% of F-measure while decision tree and random forest attain 59.98% and 55.5% of F-measure respectively.

The comparison of accuracy of different classifiers is shown in figure 5. As shown in the figure, accuracy of the different approaches is increased when the iterations increase. Namely, the proposed logistic regression attains 59% of accuracy while decision tree and random forest attain 57.5% and 54.5% of accuracy respectively.

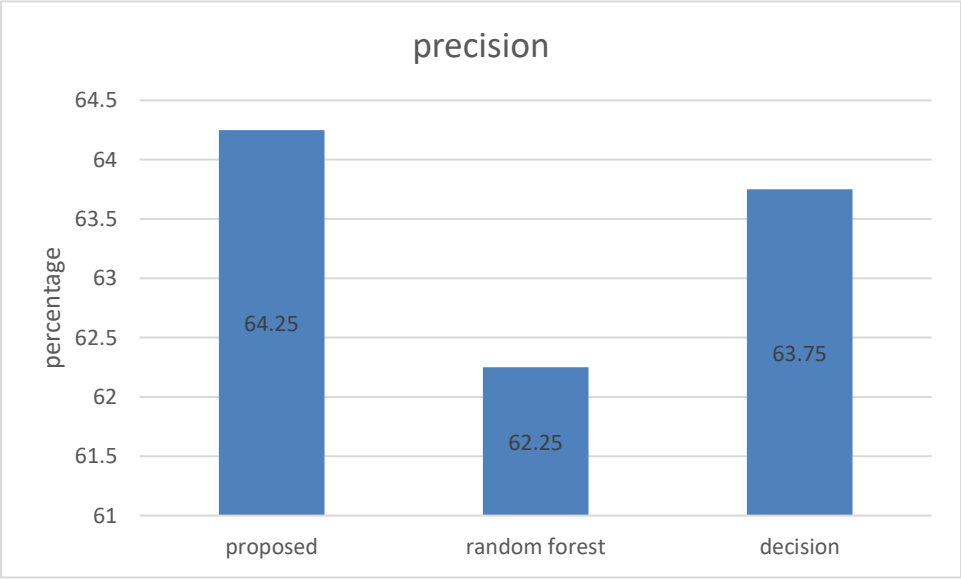


Figure 2: Comparison of Precision with benchmark classifiers

ISSN: 2233 -7857 IJFGCN

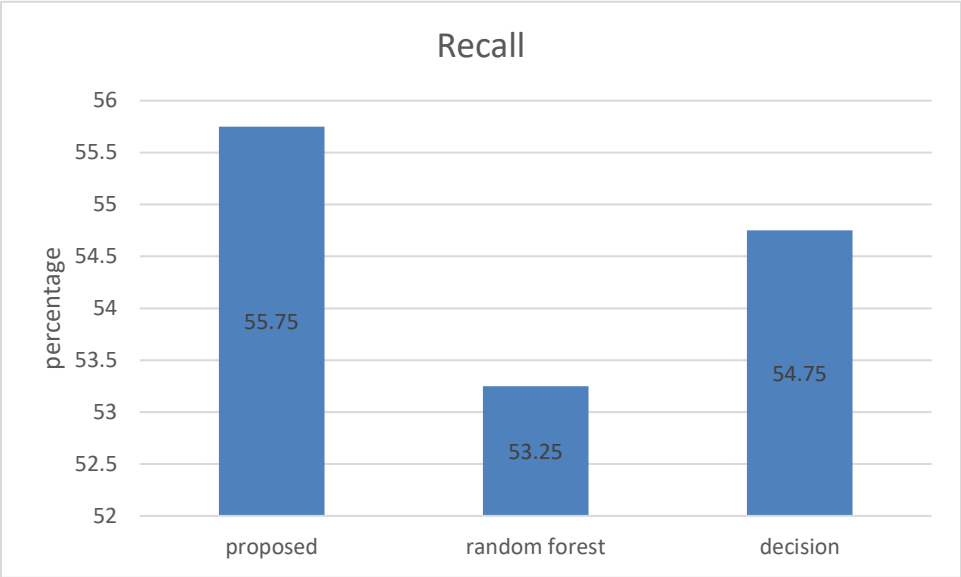


Figure 3: Comparison of Recall with benchmark classifiers

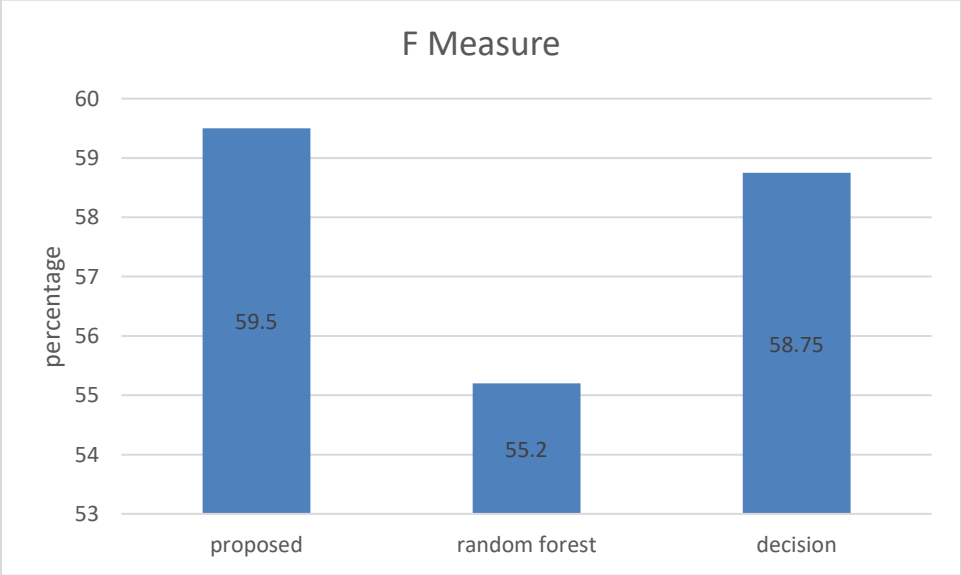


Figure 4: Comparison of F-Measure with benchmark classifiers

ISSN: 2233 -7857 IJFGCN

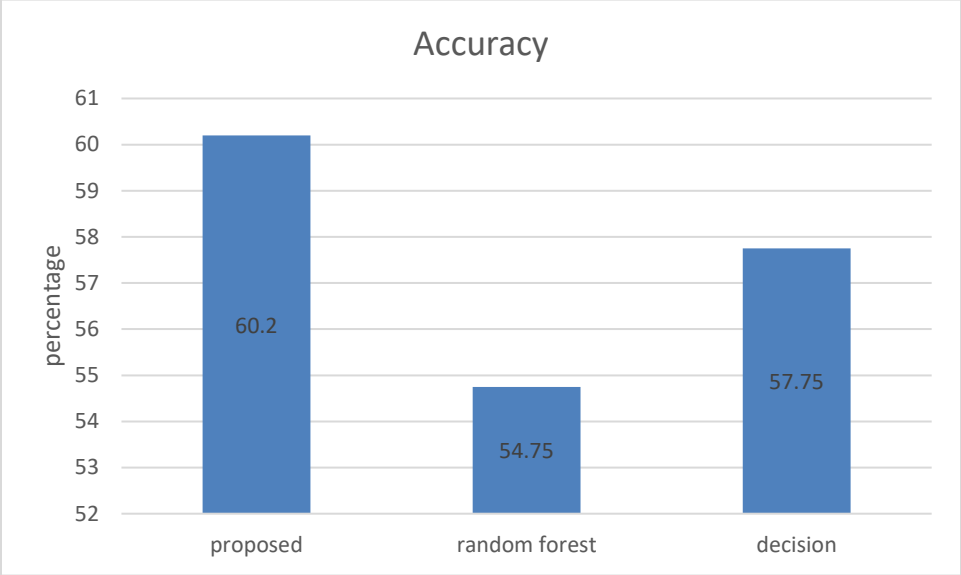


Figure 5: Comparison of Accuracy with benchmark classifiers

5 Conclusion

The approach utilized Natural Language Processing (NLP) for internet based life knowledge recovery.This exploration is done by using k-means clustering with logistic regression method to examine the tweets establishment**.** In this work, procedure of Natural Language Processing **(**NLP**)** and online networking, content characterization is broken down**.** Utilization of the online life insight by utilizing NLP is demonstrating really a distinct advantage for the managing an account establishments**.** The expectation exactness of the calculation is just more about 95**%.** The K-Means clustering based opinion examination approach bunch positive and negative comments independently**.**The proposed model attained the precision of 60.2% with logistic regression classification**.**

Reference

1. Tom Young, Devamanyu Hazarika, Soujanya Poria," Recent Trends in Deep Learning Based Natural Language Processing," IEEE transaction Computational intelligence magazine, august 2018.pp.55-75
2. R. Mourtada and F. Salem, "Citizen Engagement and Public Services in the Arab World: The Potential of Social Media," SSRN Electronic Journal, Arab Social Media Report, Dubai: Governance and Innovation Program, MBR School of Government, vol. 6, Jun. 2014

ISSN: 2233 -7857 IJFGCN

3. Pan, S. J. and Yang, Q. “A Survey on transfer learning," Knowledge and Data Engineering, IEEE Transactions , vol.22, no.10, pp.1345,1359, Oct. 2010.
4. Yingcai Wu, Nan Cao, David Gotz, Yap-Peng Tan, and Daniel A. Keim,” A Survey on Visual Analytics of Social Media Data,” IEEE TRANSACTIONS ON MULTIMEDIA, July 2016, pp.1-14
5. M. Dork, D. M. Gruen, C. Williamson, and M. S. T. Carpendale, ¨ “A visual backchannel for large-scale events,” IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1129–1138, 2010.
6. A. Zubiaga, H. Ji, and K. Knight, “Curating and contextualizing twitter stories to assist with social newsgathering,” IEEE transactions on Intelligent User Interfaces, 2013, pp. 213– 224
7. Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li,” TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets,” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, August 2013, pp.1-14
8. Sreekanth Madisetty, Maunendra Sankar Desarkar,” A Neural Network-Based Ensemble Approach for Spam Detection in Twitter,” IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, March 2018,pp.1-12
9. Tu manshu, D Wang bing,” Adding Prior Knowledge in Hierarchical Attention Neural Network for Cross Domain Sentiment Classification,” volume 7,March 2019,pp 32578-32587
10. Zhao Jianqiang, Gui Xiaolin,” Deep Convolution Neural Networks for Twitter Sentiment Analysis,” IEEE Transactions on Dependable and Secure Computing, March 2017,pp 23253 – 23260
11. Mondher bouazizi AND Tomoaki ohtsuk,” A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter,”IEEE Transactions on Knowledge and Data Engineering, October 2017,pp. 20617- 20639
12. Dong Deng, Liping Jing*, Jian Yu , Shaolong Sun, and Michael K. Ng,” Sentiment Lexicon Construction with Hierarchical Supervision Topic Model”, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 8, pp 704 - 718, APR. 2018
13. Meng Wang , Zhen-Hu Ning, Chuangbai Xiao and Tong L.,” Sentiment Classification Based on Information Geometry and Deep Belief Networks”, IEEE Computational Intelligence Magazine, Volume 9, april 2017 pp 26 – 36
14. L. Shoushan, S. Y. M. Lee, Y. Chen, C. Huang, and G. Zhou, "Sentiment classification and polarity shifting,"IEEE transactions on Computational Linguistics, march 2010, pp. 635–643.
15. X. Wan, "Bilingual co-training for sentiment classification of Chinese product reviews," Computational Linguistics, vol. 37, no. 3, pp. 587–616, Jan. 2011

ISSN: 2233 -7857 IJFGCN