Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Design And Development Of Optimal Semantic Text Tokenization (OSTT) Method For Clickbait Pre-Processing

[1] S. S. Senthil Priya, [2] Dr. S. Manju Priya

[1] Research scholar and Assistant Professor
Department of Computer Science,
Karpagam Academy of Higher Education
[2] Professor,
Department of Computer Science,
Karpagam Academy of Higher Education

Abstract: Generally, the usage of social media is increasing. It leads to the increase in online advertisement to be more popular. But these advertisements accompanied with the disturbing clickbait headlines. It is spreading the headlines with irrelevant messages. But the users may get dissatisfaction because the content of the article doesn't match with their expectation**.** So, it is an important task to prediction of clickbaits in social network to fight this problem. In order to click the fake link and attract the user's attention the click bait uses good expression with good phrases It means that clickbait use false titles in order to obtain information about the hidden user from the target page. However, it is extremely difficult to foresee and recognize these headlines manually. As a result, there is a need to design an intelligent system for predicting clickbait in social networks. Before developing a method, pre-processing of text is playing key role to improve the prediction accuracy. This research work developed a new method to pre-process the text. There are four steps in pre-processing such as Tag and special characters removal, Tokenizing, stopwords removal, stemming and Lemmatization.
Keywords: Clickbait, Machine Learning, Stopwords, Stemming, Pre-processing

I INTRODUCTION

In the present-day world, social media platforms have become the chief domain to provide information to someone, to deliver news and to be in touch with people on the internet[1]. The majorly used social media platforms that are greatly in use to share our views and information about various sectors are Twitter, Facebook and Instagram. Due to this development of social networking, a colossal quantity of information in the form of text is being shared on such social networks which progressively has become a harder task to operate with man power[9]. Despite making it simple to express our opinions, these social media platforms are also being abused to spread rumours and announcements with inaccurate information.. Since all these false information are strong enough to rock the views of the users just by creating a supremacy among the people and sway their opinions, this has become a significant affair. Thus, evolving a well-grounded system to ascertain such false information is substantial in order to safeguard the users of social media from the stretch of these false information. Clickbait, with the purpose of attracting the consumer into beating on a button to information or promotions, whose caption are entirely different from the matter contained in them, is one of the forms of such false information. Clickbait is generally found as a title or a caption that is often constructed in such a way that it urges the users to beat on the chain, specifically when it directs to a suspicious content[10,11].

In light of Twitter's growth, editors have accepted a variety of methods for creating a "curiosity gap" between the data users want and the information that is available.. Users get into the link carried by the tweets due to this curiosity gad, and view the editor's profiles. The 20 top listed inventive and productive editors on Twitter are often found to use clickbait as a tool such that clickbait tweets have reached 26% compared to the other tweets that they've posted. Such tweets intentionally shut out the principal content, at the same time encompass amplified data that are often found to be deceptive[12,13]. Say for example,

- You should be knowing this if you know about Google Docs.
- Your character can be captured in a moment with this Tumblr account.
- Money for free.

Social networks, editors and users have pessimistic impacts on these eye-catching links. First off, it wastes the consumers' valuable time, abandons them, and causes them to become frustrated. Second, it goes beyond the bounds of journalism, costing the editors their reputation. Last but not least, if social media is restricted for using such poor and imitative link baiting subjects, it negatively affects the volume of users on the social media networks. [14,15,16].

II EXISTING WORK

Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016) [1]declare that click baits are the artefacts containing deceiving captions that lead to amplified information in the linked folio, with the goal of attracting the users to beat on the click so as to fabricate the alighting folio, where the data is crappy. The appearance of them between the other contents present in the homepage of the internet sites such as Yahoo news or Google news, seriously affects the readers' circumstance. Therefore, it is necessary to chunk such contents on the homepage. Through this study, would like to demonstrate a machine-learning model to point out the click baits. This research work have used many different characteristics to show that the casualness of an internet site can also be a reason for it to be a click bait. The proposed work have also performed a series of demonstrations to assess our procession and examine the characteristics of click bait and non-click bait reports. In forecasting click baits, our prototype has accomplished a pumped-up execution (74.9% F-1 score).

Pujahari A., & Sisodia D. S. (2021)[2] states a cross codification strategy to filter click bait and non-click bait reports by amalgamating various attributes, judgement construction and congregating. The titles are filtered using 11 attributes in the course of prefatory codification followed by recodifying the titles using judgement courtesy and acceptable corresponding estimations. The titles are recodified once more by appealing congregating with the use of term direction correspondence on the basis of "t-stochastic neighbourhood embedding" (t-SNE) procedure. Following the codification of the titles, machine-learning prototypes are appealed to the dataset for the estimation of the machine-learning blueprint. The acquired outcomes of the demonstrations stipulate that the suggested cross prototype is stronger, well-grounded and intensified than some discrete codification strategies for the dataset that has been used by us.

Dong et.al (2019)[3] says, Click bait is a kind of internet subject promotions that are constructed to attract the users into beating on the links that follow. Generally, these loops would direct to the reports which are either deceptive or useless, complicating the recognition of click baits necessary for the everyday survival. A proportionate latest experimentation theme is the Automated click bait diagnosis. Its recognition problem is handled by the newest operation with in-depth swotting suggestions in order to extract the qualities from the subject's contextual data. Nevertheless, the association betwixt the deceptive captions and the prey subject is a point to be concentrated on as it is believed to be a major key for the expansion of click bait diagnosis. This study suggested an intense resemblance-conscious watchful prototype to seize and speak for these resemblances with more eloquence. Being specific, this article propose the methods of either utilizing only the resemblance or combining them with further accessible attributes for the link bait recognition. Our prototype is assessed on a couple of word processing files, the outcomes of the demonstration exhibit the productiveness of our attitude by surpassing a sequence of fierce unconventional and beginning techniques.

Seddari, et.al (2022)[4] presented a unique composite sham information identification structure that merges semantic and empirical salutations and take over their perks, just by engaging a couple of distinctive with the attributes as: (1) lexemic attributes, that is, caption, word count, simple scanning, verbal variety and belief), and (2) a unique combination of empirical attributes, known as fact-corroboration attributes that agree with three kinds of particulars that are, (i) notoriety of the database where an announcement is being produced, (ii) range, that is, units of origin that produced the information, and (iii) examining the probe, that is, views of familiar addies about the information, that is, right or wrong. The presented method recruits only eight attributes which counts lesser than the unconventional techniques added that the outcomes of the assessment on hoax information text convey that the presented method recruiting the mentioned couple of features could achieve a 94.4% exactness which is finer in comparison to the one procured from individually recruiting semantic attributes (i.e., exactness=89.4%) and attribute of database notoriety (i.e., exactness=81.2%).

Jain, et.al (2021)[5] added that the evolution of an individual prototype to identify click bait on numerous forums that work based on pictures is still a chiefly trackless trouble. Thus, we would like to launch a unique

prototype which has the capability to identify discernible click baits that are being published on both Instagram and Twitter. A stockpile algorithm groundwork consisting six bottom examples (K-Nearest Neighbours, Support Vector Machine, XG Boost, Naive Bayes, Logistic Regression, and Multilayer Perceptron) and a trans-algorithm (Random Forest). The evolved algorithm has attained an exactness of 88.5% and 85% for Instagram and Twitter posts, respectively, which is an advancement beyond preceding individual advanced prototypes for the combo of the two platforms. Furthermore, trans-attributes (for example, likes count or number of followers) are not being used by this proposed algorithm for categorisation. This is useful in recognizing probable click baits instantly through which its relevance is intensified in instantaneous click bait recognition instances. Additionally, the proposed method have also extracted important inferences on the basis of our research about the significant properties of click baits.

Razaque et.al (2021) [6]social media users are frequently becoming prey to the click baits, which is generally being taken advantage by dodgers. An eye-catching caption is being generated by the dodger which entices most of the consumers and creates an urge in them to beat on the accompanied loop. The majority of scams target those who follow the loop's instructions because it's so simple to gather their personal information. A completely unique portal expansion called 'ClickBait Security' has been put forward in order to sort out this issue. This assists in assessing the link's safety. This unique expansion works on the basis of Legitimate and Illegitimate List Search (LILS) code and Domain Rating Check (DRC) code. Two of these codes include dual search attribute to recognize vindictive data at a very faster rate. Additionally, Clickbait Security influences the attributes of an intensified Recurrent Neutral Network (RNN). The suggested Clickbait Security is said to have more exactness in recognizing vindictive and secure loops, as compared to the other available findings.

III DATASET COLLECTION

This dataset contains headlines from many news sites and other social medias such as 'WikiNews', 'New York Times', 'The Guardian', 'The Hindu', 'BuzzFeed', 'Upworthy', 'ViralNova', 'Thatscoop', 'Scoopwhoop', 'Twitter', 'facebook' and 'ViralStories'. It has two columns: one with headlines and one with numerical labels of clickbait, where 1 indicates that the headline is clickbait and 0 indicates that it is not. The dataset comprises 32000 entries, half of which are clickbait and the other half are not.

IV PROPOSED WORK

OPTIMAL SEMANTIC TEXT TOKENIZATION(OSTT) METHOD

Accuracy of clickbait prediction depends on pre-processing data. It is an important step to increase the excellence of data and accuracy of detection. This step is used to improve the accuracy of clickbait classification also. Because data may have noise, extra characters and missing values. These may lead to affect the efficiency of classification. Therefore, this research introduced a new method to pre-processing. It is proceeded by removing all special characters (such as punctuation and Hash tag) so it keep only alphabets. Next to reduce errors in data interpretation the title characters are changed to lowercase letters. Then additional process have to perform to remove the following URL, hash tags (#), emojis, and punctuation (“”) . The figure 1 shows the framework of the proposed research. There are four steps are used in this proposed method such as

1. Removal of Tags and special characters
2. Tokenizing Document
3. Removal of Stop Words
4. Streaming & Lemmatization

The first is used to remove all the tags and special character of the text file. The clean text is sent as input to the tokenizing process in the second step after this cleaning phase. In this process the clean text is going to split up in to token. That is each word as divided into token.

After completing the tokenizing task, then third process focus on removal of unwanted words form token. According to the Oxford dictionary guidelines there are 29 stops words lists are available. Then remove the stop words form token as per the list and get the pure token. It is considered as noise free text. Then these pure texts are given as input to the streaming and Lemmatization process.

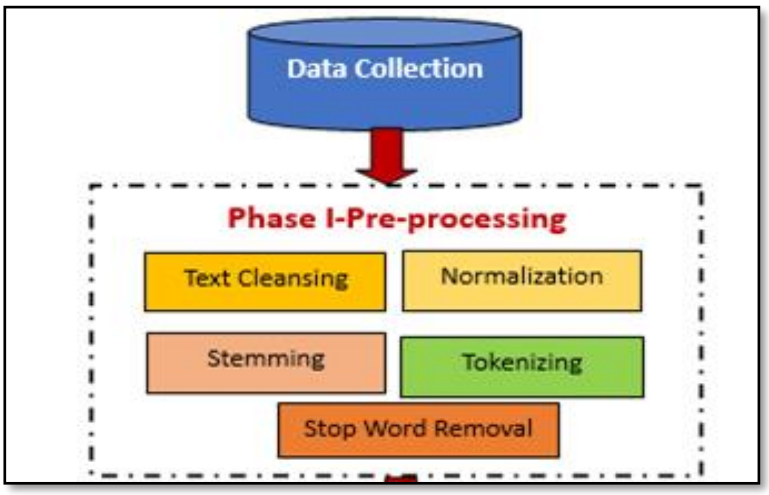


Figure 1. Flow of proposed method

a). Removal of Tags and special characters

This is the first step to remove the tags and special character from the text. It is very import to the next step. The raw data set is taken as input and remove the tag and special characters using library and non-library functions. Library function such as isalnum(), filter(),replace() and Translate(). Finally, it can get the tag free text as output. This output given as input for next step. Figure 2 shows the steps of removal of tags.

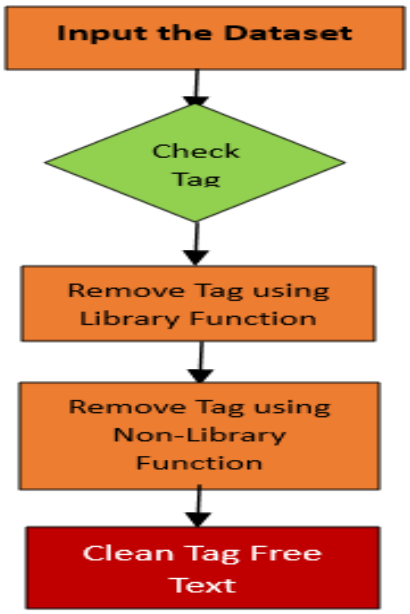


Figure 2 Removing of Tag and Special character

b). Tokenizing Document

After the removing of tags form the text the clean text as given as input to tokenizing. the proposed system going to split the text into the token. It is the process of splitting the sentence in to words. After the tokenizing conversion of mixed case such as uppercase letter with lower case and lower case sentence or words may mixed with upper case. Finally, this can get the noise free token. Figure 3 shows the process of tokenizing.

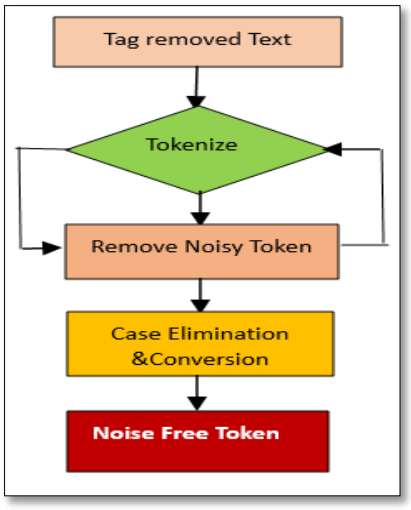


Figure 3 Process of Tokenizing

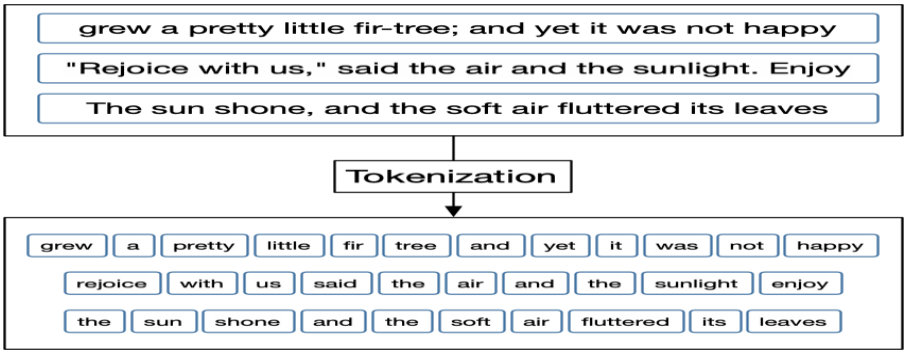


Figure 4 sample process of Tokenizing

c). Removal of Stop Words

A Stopwords are frequently used English words such as “a”, “an”, “in” i’, ‘me’, ‘my’, ‘myself’, ‘we’, ‘our’, ‘ours’, ‘ourselves’, etc.,. These words don’t add more meaning to the sentences. They must be eliminated without considering the sentence’s meaning. These words will take up space in the database or it will take a more processing time. For this reason, in this method which can remove them easily by storing a list words. The special python-based package Natural Language Toolkit (NLTK) is used to eliminate stop words from noise free token. Figure 5 shows the steps of stopwords removal.

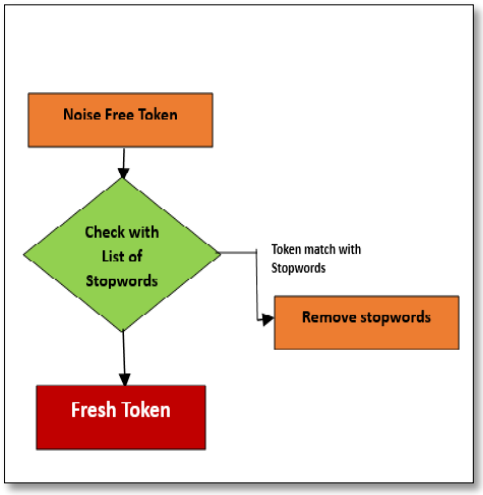


Figure 5 Removal of Stopwords

d). Stemming & Lemmatization

Stemming is a technique for obtaining the base or root/stem form of words by removing affixes. It's the same as pruning a tree's branches down to the trunk. The stem of the terms writing, writes, and written, for example, is write. It will provide the root of a word. In this study, the wordnet technique is applied to the steaming process.. After the token process it has been send the list of tokens to WordNet. Then it will get the synonym of the token. This research work implement three kinds of stemmer functions such as PorterStemmer, LancasterStemmer, RegexpStemmer, converting plural to singular and translating its past tense to present tense

1.Converting the plurals of a token to its singular form

2.PorterStemmer will remove the postfix in the token for example writing to write,

3.LancasterStemmer will remove the postfix in the token for example eats to eat

4.RegexpStemmer will remove postfix and suffix in the token for example ingeat to eat, eating to eat, eats to eat.

5.Translating the past tense of a token to its present tense

After the stemming process a Lemmatization process is there. It is also a stemming process but it will give a root word rather than root stem. It gives a meaning full word finally. It is also called as lemma. For example believes to believ. Lemmatization is the final process to get the stemmed token.

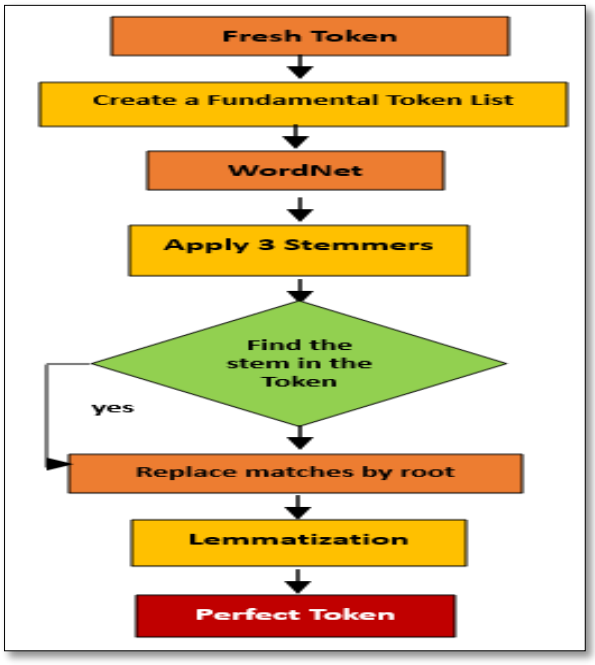


Figure 6. Process of Stemming and Lemmatization

V EXPERIMENT AND RESULT

This research used Python programming to implement the above all process. Python is a dynamic and opensource language. It used for machine learning because of its huge library packages in it. It is very easy to learn that is the point make us to implement. It supports to the Natural Language Processing (NLP). It has lot of tools for NLP.

After the implementation this need some sample data to test. Then this have to test massive amount of data. In this research data has been collected from various news medias and social medias. Using the sample data as input, this operation first removes tags and tokenizes each text file. There are 15 text files in this folder. Table 1 shows token reports after removing tags and tokenizing 15 files.

Table 1: Sample text file report after tokenize

Name of File	Number of Token in File
Sample 1	613
Sample 2	557
Sample 3	527
Sample 4	504
Sample 5	551
Sample 6	616
Sample 7	508
Sample 8	559
Sample 9	651
Sample 10	379
Sample 11	149
Sample 12	386
Sample 13	254
Sample 14	400
Sample 15	132

After tokenizing send these files to remove the stop words. Here it send 15 files to remove the stopwords. It removes all of the stop words from our 15 sample files. Oxford dictionary service was used to compile this list of stop words. Although there are just 529 words, their usage frequency is substantially higher. After eliminating the stop word from our token report, we may remove the token numbers listed in Table 2.

Table 2 Report on number of stopword removed

Name of Source File	Number of Removed Stopwords	Token after Remove Stop word
Sample 1	86	527
Sample 2	105	452
Sample 3	71	456
Sample 4	86	418
Sample 5	213	338
Sample 6	103	513
Sample 7	61	447
Sample 8	128	431
Sample 9	62	589
Sample 10	34	345
Sample 11	60	89
Sample 12	49	337
Sample 13	74	180
Sample 14	47	353
Sample 15	66	66

This step uses the WordNet and synset. The synset is synonyms of the word and the words can be replaced the by synonyms. The result of these process is replaced by the root word. After this process it focused on lemmatization. It will help to normalize the token.

This experiment contains 15 samples drawn from ten various sources, including social media and news. When the example text was tokenized after the tags were removed, it had 6786 tokens. Then proceed to the

transaction involving those 6786 tokens. After receiving the token from the previous phase, proceed to clean up the stop words from those tokens, and it tends to find that 1524 token is from stopwords. It tends to subtract those tokens from the total number of tokens. The token's range can be extremely broad. That is 46.29% of the total token. Once removed, there are 1784 tokens. Those 1784 tokens are sent independently to WordNet to obtain synsets (synonyms). The word was then changed to synsets. The term was then substituted by its signifier using the synsets. Then it proceeds to the lemma procedure. This system transmits the original word to be lemmatized. When using the complete outset margining and lemmatization procedure, it returns 3674 tokens to the system. The system receives 6786 tokens from the input and currently has just 3674 tokens.

VI IMPLEMENTATION RESULTS

Figure 7 shows the sample 1 files it has 20 headlines with special characters like #,Emoji, smiley, @, numeric and other characters. These special characters have to be removed by using library and non-library functions.

1	**Head lines**
2	Facebook Is Eating the World!!
3	15 Resolutions To Make Good On In 2016#
4	What New Thing Should You Try In 2016🙃🙃
5	Zoo Animals Around The World Are Opening Their Christmas Presents Early
6	Tell Us About Yourself(ie): Erica Ash))
7	9 Times I Cried$$$
8	21 Vegetarian Dump Dinners For The Crock Pot()
9	This Goat Has Been Bullying His Tiger Friend
10	8 Fall Shows To Be Excited About, 10 To Give A Chance, And 6 To Avoid
11	Another Round, Episode 25: Stop Telling Women To Smile!
12	16 Signs You Are Too Stubborn To Live 😎😎
13	This Country Singer Makes Music On His Game Boy In His Spare Time
14	When You Realize Every Guy Is Taken%%
15	An Awesome Look At The Behind-The-Scenes Concept Art Of "Aladdin"
16	Sarah Jessica Parker Talks Being A Carrie, "Hocus Pocus," And Her Love Of New York
17	Which Type Of Swearer Are You@
18	A Dad Recorded All The Adorable Questions His Son Asked When Watching Star Wars For The First Time
19	29 Impossibly Stylish Cat Gifts, In Order Of Awesomeness**
20	13 Of The Most Glorious Made-Up Words From Literature

Figure 7 Sample 1- Head lines

```
Output
---------
Facebook Is Eating the World
Resolutions To Make Good On In
What New Thing Should You Try In
Zoo Animals Around The World Are Opening Their Christmas Presents Early
Tell Us About Yourself Erica Ash
Times I Cried
Vegetarian Dump Dinners For The Crock Pot
This Goat Has Been Bullying His Tiger Friend
Fall Shows To Be Excited About To Give A Chance And  To Avoid
Another Round Episode Stop Telling Women To Smile!
Signs You Are Too Stubborn To Live
This Country Singer Makes Music On His Game Boy In His Spare Time
When You Realize Every Guy Is Taken
An Awesome Look At The Behind-The-Scenes Concept Art Of Aladdin
Sarah Jessica Parker Talks Being A Carrie Hocus Pocus And Her Love Of New York
Which Type Of Swearer Are You
A Dad Recorded All The Adorable Questions His Son Asked When Watching Star Wars For The First Time
Impossibly Stylish Cat Gifts In Order Of Awesomeness
Of The Most Glorious Made-Up Words From Literature
```

Figure 8 Special characters Removed

```
Output
---------
[ Facebook Is Eating the World,
  Resolutions To Make Good On In,
  What New Thing Should You Try In,
  Zoo Animals Around The World Are Opening Their Christmas Presents Early,
  Tell Us About Yourself Erica Ash,
  Times I Cried,
  Vegetarian Dump Dinners For The Crock Pot,
  This Goat Has Been Bullying His Tiger Friend,
  Fall Shows To Be Excited About To Give A Chance And  To Avoid,
  Another Round Episode Stop Telling Women To Smile,
  Signs You Are Too Stubborn To Live,
  This Country Singer Makes Music On His Game Boy In His Spare Time,
  When You Realize Every Guy Is Taken,
  An Awesome Look At The Behind-The-Scenes Concept Art Of Aladdin,
  Sarah Jessica Parker Talks Being A Carrie Hocus Pocus And Her Love Of New York,
  Which Type Of Swearer Are You,
  A Dad Recorded All The Adorable Questions His Son Asked When Watching Star Wars For The First Time,
  Impossibly Stylish Cat Gifts In Order Of Awesomeness,
  Of The Most Glorious Made-Up Words From Literature ]
```

Figure 9 Tokenizing

VI Performance Metrics

The performance of proposed method is analysed with precision, recall, F-measure and accuracy.

Performances are analysed with classification stage of text mining with precision, recall, F measure and accuracy. The calculations are performed based on the accurate text with feature extracted text.

Precision:

Ratio of relevant token retrieved to the total number of tokens

Precision = (TP / TP + FP)→(1)

Recall ratio:

Ratio of correctly predicted Token to actually retrieved token

Recall ratio = (TP / TP + FN)->(2)

F-Measure:

The F-Measure computes some average of the information retrieval precision and recall metrics. Root means the squared method is used to calculate the RMS error.

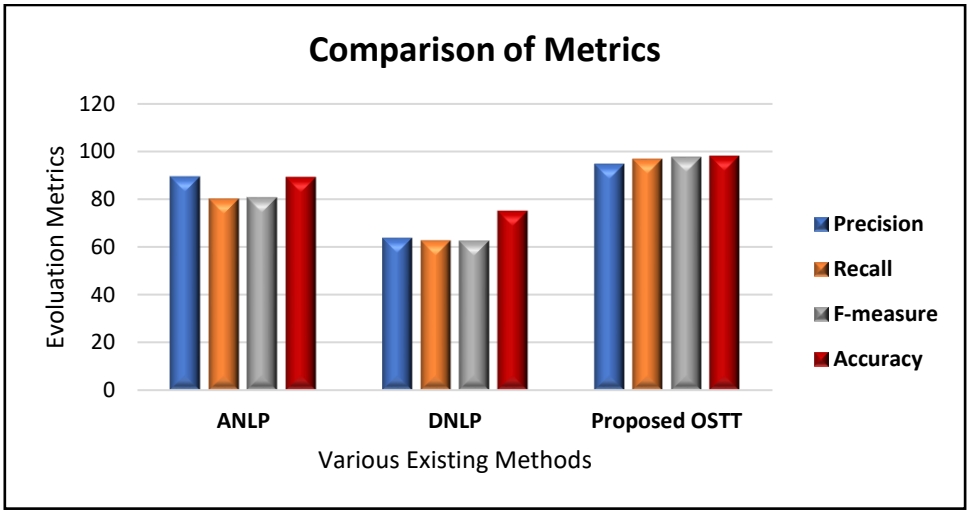
F-measure =$\sqrt{(\text{True negatives}-\text{True Positives})^2+(\text{False Negatives}-\text{False Positives})^2}$ →(3)

Accuracy(AC):

Accuracy is the most intuitive performance measure, and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = (TP + TN / TP + FP + FN + TN)→(4)

The performance of proposed pre-processing techniques evaluated and compared with existing algorithm such as ANLP, DNLP. The proposed method gave better result than others.



The accuracy of the proposed method is 98.02. when compared with other methods it is better accuracy. The precision value of OSTT is 94.56. The recall is 96.78 then F measure is 97.67.

VI CONCLUSION

This paper has proposed a comprehensive approach for pre-processing that includes four main steps for pre-processing: data collection, Tag Removal, Tokenizing, Stopword removal and Stemming & Lamitization steps. In this experiment, there are 15 samples from 10 different sources such as social medias and news. The system gets 6786 token as input initially after all processing it has 3674 fresh tokens. These tokens can be taken for further process.

REFEENCES

[1]Biyani, P., Tsioutsiouliklis, K., & Blackmer, J. (2016, February). " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).

[2]Pujahari, A., & Sisodia, D. S. (2021). Clickbait detection using multiple categorisation techniques. *Journal of Information Science*, *47*(1), 118-128.

[3]Dong, M., Yao, L., Wang, X., Benatallah, B., & Huang, C. (2019). Similarity-aware deep attentive model for clickbait detection. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23* (pp. 56-69). Springer International Publishing.

[4]Seddari, N., Derhab, A., Belaoued, M., Halboob, W., Al-Muhtadi, J., & Bouras, A. (2022). A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, *10*, 62097-62109.

[5]Jain, M., Mowar, P., Goel, R., & Vishwakarma, D. K. (2021, March). Clickbait in social media: detection and analysis of the bait. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)* (pp. 1-6). IEEE.

[6]Lischka, J. A., & Garz, M. (2021). Clickbait news and algorithmic curation: A game theory framework of the relation between journalism, users, and platforms. *New Media & Society*, 14614448211027174.

[7]Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). Bert, xlnet or roberta: the best transfer learning model to detect clickbaits. *IEEE Access*, *9*, 154704-154716.

[8]Razaque, A., Alotaibi, B., Alotaibi, M., Hussain, S., Alotaibi, A., & Jotsov, V. (2022). Clickbait detection using deep recurrent neural network. *Applied Sciences*, *12*(1), 504.

[9]H. Huan, J. Yan, Y. Xie, Y. Chen, P. Li et al., “Feature-enhanced nonequilibrium bidirectional long short-term memory model for Chinese text classification,” IEEE Access, vol. 8, pp. 199629–199637, 2020.

[10] V. Kuppili, M. Biswas, D. R. Edla, K. R. Prasad and J. S. Suri, “A mechanics-based similarity measure

fortext classification in machine learning paradigm," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 2, pp. 180–200, 2018.
[11] S. Ahmad, M. Z. Asghar, F. M. Alotaibi and S. Khan, "Classification of poetry text into the emotional states using deep learning technique," IEEE Access, vol. 8, pp. 73865–73878, 2020.
[12] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood et al., "Document-level text classification using single-layer multisize filters convolutional neural network," IEEE Access, vol. 8, pp.42689–42707,2020.
[13] H. Calvo, A. P. Rocha-Ramirez, M. A. Moreno-Armendáriz and C. A. Duchanoy, "Toward universal wordsense disambiguation using deep neural networks," IEEE Access, vol. 7, pp. 60264–60275, 2019.
[14] Q. P. Nguyen, A. D. Vo, J. C. Shin and C. Y. Ock, "Effect of word sense disambiguation on neural machine translation: A case study in Korean," IEEE Access, vol. 6, pp. 38512–38523, 2018
[15] https://github.com/bhargaviparanjape/clickbait/blob/master/dataset/non_clickbait_data.gz
[16]https://www.kaggle.com/datasets/amananandrai/clickbait-dataset?resource=download