ISSN: 2096-3246
Volume 54, Issue 03, August, 2022

Combining Agglomerative Data Clustering And Anti-Dictionary To Improve Text Classification

Titu Singh Arora[1] and Amit Sharma[2]
Ph.D. Research Scholar, School of Engineering &Technology,
Career Point University, Kota, Rajasthan, India.[1]
Associate Professor, School of Engineering &Technology,
Career Point University, Kota, Rajasthan, India.[2]

Keywords:
Data Mining, Classification, Vectorization, Clustering

ABSTRACT

With expanding force of calculation and information stockpiling in this day and age, the idea of gathered information is likewise moving quickly from organized to unstructured information. The most well-known unstructured information gathered is the text information. These information can give significant experiences about the specific circumstance and help going with choices in light of the result of the bits of knowledge. Text order is one such piece of text information investigation where the named information is exposed to an AI model to recognize which text has a place with which class. Hence, this study comprehends the setting behind the text is to distinguish the associations of the words that show up regularly together. Sogrouping such words in bunches, it is feasible to have a relevant measurement and thus support text classification.This paper presents a clever methodology of consolidating message grouping and hostile to word reference word extraction as information pre-handling move toward further develop the characterization models. The dataset for this study contain is an assortment of news stories that are marked as 'phony' or 'genuine'. In this study the message information is exposed to TF-IDF vectorization which makes a meager framework andeach column in the lattice addresses a vectorized structure. In view of these vectors, agglomerative bunching is executed, and the information is appointed to two new groups as an additional component.

1. INTRODUCTION

By definition, text examination is the most common way of searching for huge bits of knowledge, patterns, and examples in unstructured and semi-organized text information. In above models, text information examination calculation would beat all people concerning effectiveness, nature of data and speed. The outcomes are less one-sided, and less asset concentrated. This makes it profoundly alluring to a few associations across different businesses to give a subjective as well as quantitative comprehension of their items, clients and the market. Data gathered from text information investigation is joined with information representation apparatuses to make an interpretation of data to significant bits of knowledge and in this manner, help dynamic cycles. Message investigation comprises of different strategies like quantitative message summarisation, opinion examination, record categorisation, directed arrangement, bunching, report highlight acknowledgment and regular language handling and so forth. One of them is text characterization where the goal of the calculation is to classify the text into at least one gave classifications. A portion of the instances of text characterization are:

- Utilizing online entertainment to check crowd opinion.
- Spam and non-spam email location.
- Client questions are naturally labeled.
- Reports are arranged into assigned points.
- Identify whether the distributed article is genuine or counterfeit.

The force of text order appears to be fascinating, however it accompanies expected issues. Unstructured information represents around 80% of the multitude of information with message being perhaps of the most well-known class. Besides, investigation of text information physically is troublesome and tedious because of its muddled nature and in this manner, many organizations neglect to use the data at its fullest. This is where AI and text order become possibly the most important factor. Organizations might utilize text classifiers to orchestrate a wide range of pertinent substance, including messages, authoritative records, web-based entertainment, chatbots, studies, and all the more rapidly and cost-successfully. Organizations can save time breaking down text information, computerize business cycles, and go with information driven business decisions because of this innovation.

1.2. Challenges in Text Classification

Customarily, text order can essentially be separated into four significant parts:

1. Feature Extraction
2. Dimensional decrease
3. Classification procedures
4. Evaluation

Different improvements have been finished in these four areas of text grouping and they likewise have given nice outcomes. In any case, these strategies are not 100 percent precise and represent specific subjective issue ready to go. Every one of these calculations, right off the bat, examine text grammatically. i.e., these calculations examine every text separately and considers no relationship among the texts that are in a similar class. This prompts a deficiency of significant data about the relationship among the records. Also, these calculations only sometimes search for the setting in the information or at the end of the day, "importance hidden therein" of text. By understanding the setting of the data in each class, we can give extra credits to grouping that will fortify the general result of the cycle.

1.3 What is Anti-Dictionary? In this undertaking, Anti-word reference is a term given to a bunch of words that doesn't show up in the Standard English Dictionary yet are found in the news stories. These words can be formal people, places or things(like the names of individuals, place, and so forth), abbreviations (short types of expressions, words and so on) or could be spelling mistakes. Out of these, the formal people, places or things and abbreviations are recognized and taken out by breaking down the linguistically right corpus. The excess words are recognized as against word reference of spell blunders that must be found in wrong or conflicting news sources.

2. LITERATURE REVIEW

The area of artistic substance mining attempts to eliminate supportive experiences from unstructured printed estimations through the ID and examination of captivating models. The philosophies used by and large in all actuality do at absolutely no point in the future contain significant semantic assessment or parsing, but depend on direct "pack of words" printed content depictions considering vector space. A couple of ways of managing the ID of styles are inspected, similar to dimensionality decline, motorized class, and grouping. Shloka Gilda presented thought generally on how NLP is relevant to stumble on fake information. They have used time frame repeat turn around record repeat (TFIDF) of bi-grams and probabilistic setting free accentuation (PCFG) recognizable proof. They have examined their dataset over more than one class of computations to sort out the staggering model. Mykhail organic proposed a fundamental technique for fake news disclosure the utilization of a guiltless Bayes classifier. They used BuzzFeed news for becoming mindful and offering a chance to the simple Bayes classifier. The dataset is taken from Facebook news disperse and completed accuracy up to 74% on the test set. Cody Buntain advanced a methodology for modernizing fake news disclosure on Twitter. They applied this methodology to Twitter content acquired from BuzzFeed's fake news Dataset.

Shivam B. Parikh hopes to present a piece of information on the depiction of reports in the high-level diaspora got together with the differential substance sorts of reports and their impact on perusers.

Himank Gupta et. al. [10] gave a design considering a different AI approach that plans with various issues including accuracy need, delay (BotMaker), and high taking care of time to manage extraordinary many tweets in 1 sec.

Maria Teresa Artese and Isabella Gagliardi (2022) The paper means to characterize models and devices for taking care of literary explanations, for our situation catchphrases of a logical library. The various strides of the pipeline are examined, executed, assessed, and analyzed, utilizing factual strategies, AI, and fake brain networks as fitting. Models are prepared on various datasets currently accessible or made specially appointed with normal attributes with the beginning dataset.

Joaquim Fernando Pinto da Costa and Manuel Cabral (2022) The significance of measurable strategies in finding examples and patterns in any case unstructured and complex enormous arrangements of information have developed over the course of the last 10 years. This paper is a thorough and orderly survey of these new improvements in the space of information mining. The information got from understanding information permits us to settle on speedy and informed choices that save time and give us an upper hand.

3. METHODOLOGY

Text information can be found in an overflow wherever on the web and is accessible either in unstructured or in different arrangements. By ordering the information into various classes, we can get patterns from the text information and get significant experiences. Message grouping helps in this and is profoundly productive and strong method for arranging message information under various classes that address various implications. In this part, we will examine different wordings relating to message characterization and afterward we will give data about the proposed calculation that has been carried out on the highest point of existing strategies.

3.1 Text Classification Pre-processing

In any analytical and Data Science project, data pre-processing is one of the major challenges and takes most of the time and effort. It is even common to assume that 80% of the project timeline is often dedicated to data collection and making it fit for use. In case of text data, text pre-processing is the method used to clean the data and make it organised to be fed to the data model. If we don’t perform this step on the unstructured text data, then the model will either provide poor results or it won’t even function as per requirement. Since the data is in the form of human language, which is different from what machines could understand, we need to convert them into numbers and in an efficient manner so that the machines could provide better analysis out of the data. To understand this step, first we will understand the scope of data pre-processing and then various methods that are commonly in use for the same.

3.1.1 Scope

Each row or vector of the text data available today is either in the form of document, paragraph, sentence, or even sub-sentence. Depending on the applications, these levels of text are analysed and categorised accordingly. For example, in a spam detection model, each email is considered as a document and the classifier is built such that it takes whole email text as input. Another example is sentiment analysis algorithm on customer feedback on a product. Here, each review is in the form of a paragraph and is fed into the model as such. Thus, depending on the applications, a document can be assigned into one or many classes while some application may assign different class to even every sentence. One may observe that the sentence level classification model may be a bit granular, but here we lose meaning because we may not know the context that is determined by other sentences near to the one under analysis. Therefore, it is important to first understand the scope of the problem and then define the classification model. This way, we will minimise the potential loss of context while building the classifier.

3.1.2 Document Representation – Bag of Words

The key idea behind any analytics is to convert the data into numbers and then numbers in turn provide structure to the whole thing. In text analytics as well, it is essential for pre-processing the documents to convert them into numbers so that we can properly structure the text documents into information that can easily be taken up by machine learning algorithms to identify patterns. [4] One of the common approaches for document representation is Bag-of-Words. In this process, each document vector is analysed by counting the frequency of each word that appear in the document. This creates a long vector in which each cell of the vector represents how many times a particular word has appeared in that document. This can be understood by the following example.

1. Ram plays cricket very well.
2. Cricket is Ram's favourite sport.
3. Ram is a good swimmer.
4. Cricket is a good sport.

When we convert the above four sentences into Bag-of-words model, we get something like below:

Words	Ram	Plays	Cricket	Very	Well	Is	Favourite	Sport	Good	swimmer	A
1.	1	1	1	1	1	0	0	0	0	0	0
2.	1	0	1	0	0	1	1	1	0	0	0
3.	1	0	0	0	0	0	0	0	1	1	1
4.	0	0	1	0	0	1	0	1	1	0	1

By addressing the words in mathematical arrangement, it is not difficult to dissect and look at the records. This is truly useful with solo text examinations like relationship, grouping, and correlation. In this model, sentences 1,2,3 discuss Ram, and 4 don't, and that actually intends to get data about Ram, we really want not to concentrate on sentence 4. Additionally, sentences 1,2, and 4 notice cricket, and accordingly, these sentences are near one another in the setting of the word 'cricket'.

Sack of words model diminishes the dataset size fundamentally changing text over completely to a mathematical organization that is simpler to peruse by machines and requires fewer calculation prerequisites. In any case, it has its own faults too. It, right off the bat, is influenced by a revile of dimensionality. By addressing each word independently in the report, there are chances that they might contain phrases that are not really utilized. This prompts a superfluous expansion in vector size. Besides, there are stop words like 'is', 'what', 'are', 'like', and a lot more that happen in high recurrence in each record however hold practically no importance concerning sets. Besides, when we address a record in a vector design, we break its semantic honesty totally and all significance of adjoining sentences is lost. This prompts the deficiency of setting in the passage designs.

3.1.3 TF IDF

In Bag of words model, the significant downside is that it simply counts the words and gives a vector portrayal of them. This prompts falsehood since well-known words that happen every now and again in the English language get more word counts than those which the subject is about. To take care of this issue, another calculation is presented. TF-IDF represents Term Frequency - Inverse Document Frequency. A mathematical measurement furnishes the significance of each word concerning the record as well as the assortment of reports. For any word, it is proportionate to the recurrence of that word showing up in a report and is contrarily corresponding to the recurrence of that word showing up in different records. This prompts a particular measure that unequivocally gives how significant is a word regarding a specific record and subsequently disposing of the downside of well known words that happen clinched of words model.

The TF-IDF statistic consists of 2 parts

TERM FREQUENCY:

Term recurrence is distinguished by working out the recurrence of the appearance of a word in a specific report. it is determined as:

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

Where, the numerator ft,d is a recurrence of word t showing up in the archived and the denominator is the complete number of words in the report d.

Converse Document Frequency:

Inverse Document Frequency:

Converse Document Frequency is the proportion of significance of a word that shows up in the record. Numerically it is characterized as:

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where N is the absolute number of records in the dataset and the term $|\{d \in D : t \in d\}|$is the number of reports wherein the word shows up no less than once.

number of documents in which the word appears at least once.

Consequently, the metric TF-IDF is determined as:

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Consequently, TF-IDF offers a higher benefit to the terms that are uncommon in the dataset yet show up regularly in a solitary record.

3.3.3 Ensemble techniques

To dispense with the dangers presented by choice tree classifiers, many advances have been taken. The most widely recognized and successful methodology is Ensemble procedures where many models with various boundaries are constructed. Out of these many models, the aggregate aftereffects of the classifiers are considered for navigation. Irregular timberland classifier is one such group classifier that pursues numerous choice trees by utilizing approaches like stowing and Boosting and afterward giving the outcomes in view of the consolidated result of all the choice trees. This strategy makes it profoundly effective and hearty.

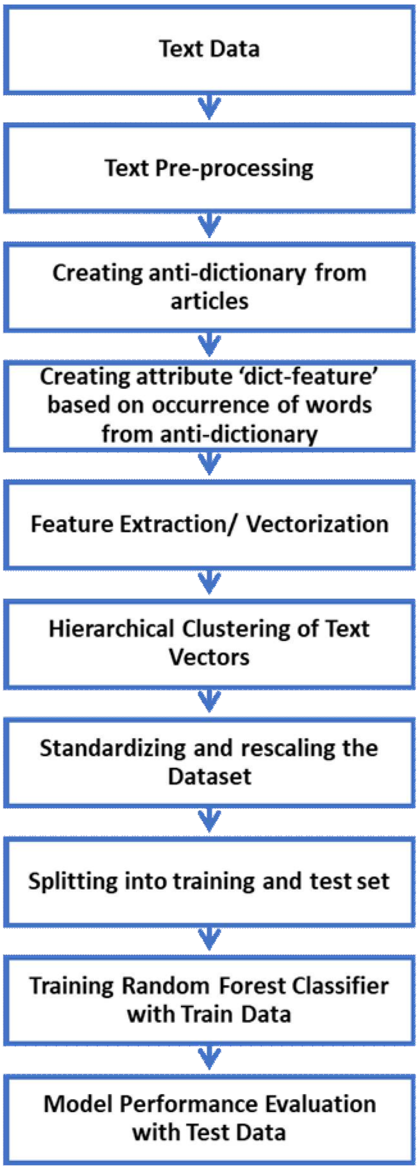

Text Data
Text Pre-processing
Creating anti-dictionary from articles
Creating attribute 'dict-feature' based on occurrence of words from anti-dictionary
Feature Extraction/ Vectorization
Hierarchical Clustering of Text Vectors
Standardizing and rescaling the Dataset
Splitting into training and test set
Training Random Forest Classifier with Train Data
Model Performance Evaluation with Test Data

Fig. 1 Flow of Algorithms process

Model

This figure provides a pictorial representation of the proposed algorithm in the project. First, the text data is loaded in the environment. Next, a few basic pre-processing steps are performed on the data. These include separating text attribute from the labels, removing punctuations, decapitalizing, dropping nulls, removing stop words and special characters,etc. Then, we perform TF-IDF vectorization, where each row of text data is converted into the vector format. Here, each vector represents the frequency of a word that appears in the text. After vectorization, the data is subjected to K-means clustering where the clusters are formed based on the vectors obtained in the previous step. The cluster label for each vector is used as a new attribute and is added as a new feature in the dataset. The final modified dataset is split into training and test set. Training set is used to train the ensemble classification model (Random Forest). Validation set (Taken from the training set) is used to fine tune the model. Once the model is trained successfully, the test data is used to evaluate the performance of the model. This performance is compared with the base algorithm (without clustering) and the observations are done.

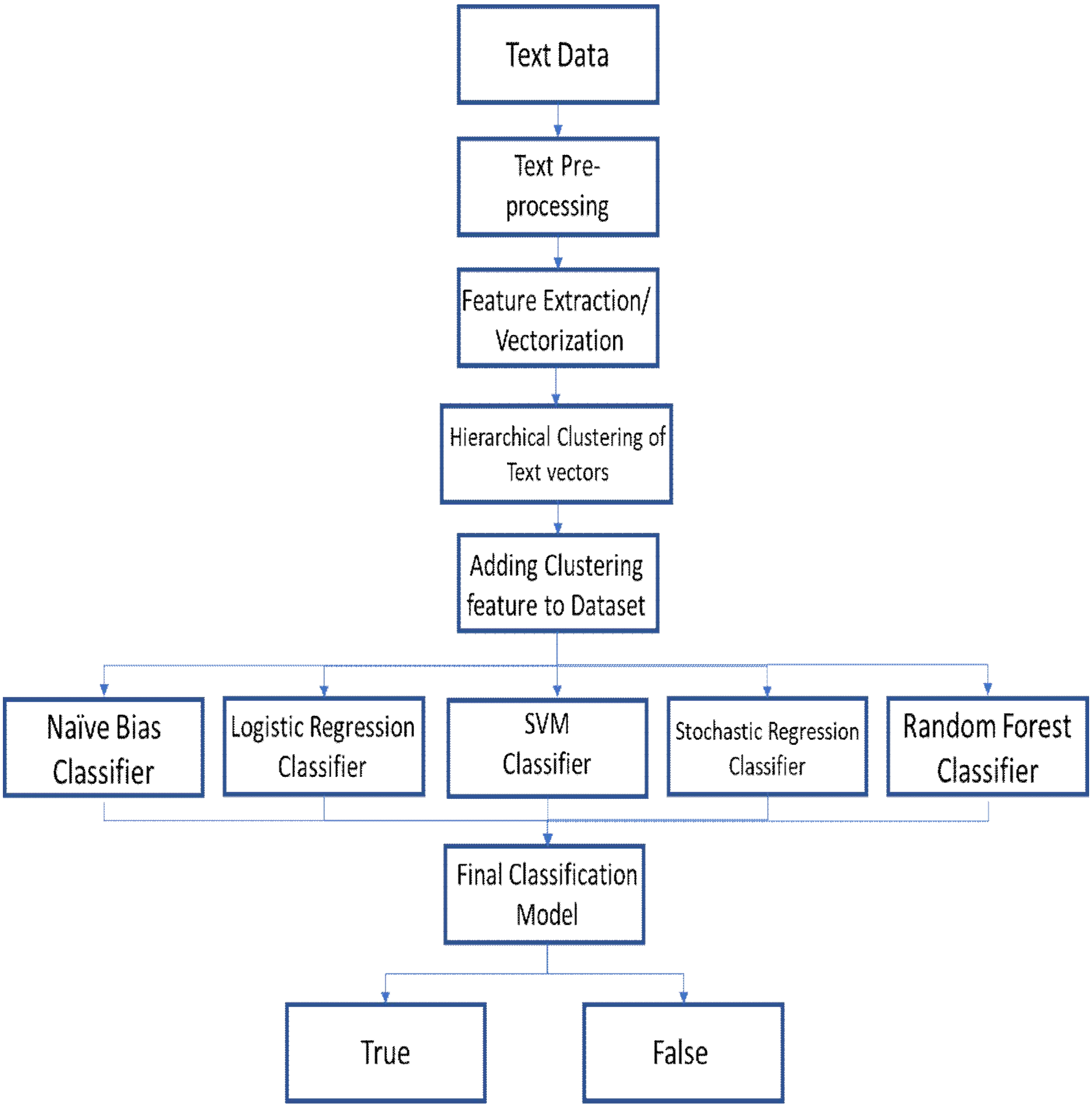

Text Data
Text Pre-processing
Feature Extraction/ Vectorization
Hierarchical Clustering of Text vectors
Adding Clustering feature to Dataset
Naïve Bias Classifier
Logistic Regression Classifier
SVM Classifier
Stochastic Regression Classifier
Random Forest Classifier
Final Classification Model
True
False

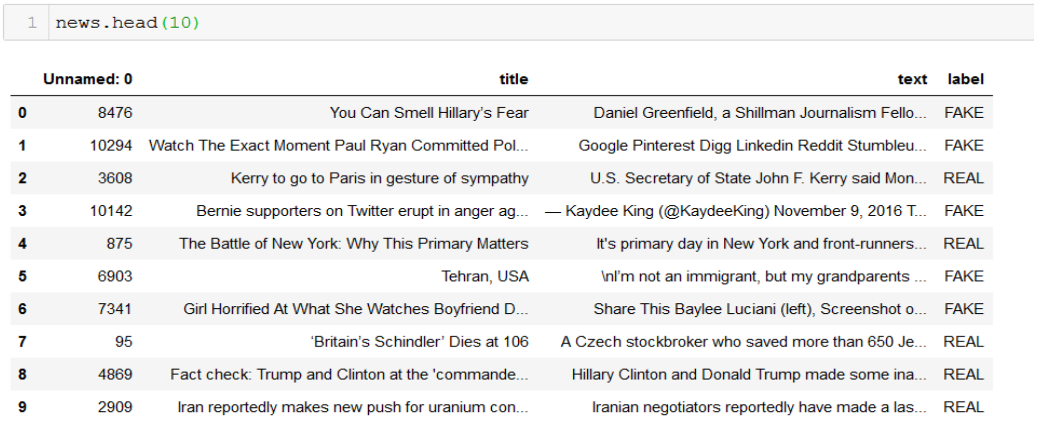
Fig. 2 Proposed Flow of Algorithms process

4. RESULTS AND DISCUSSION

Description of Datasets

The dataset comprises of two sorts of articles namely True and Fake. These articles are mostly related to the US political news and are sourced from Reuters.com which is a news blog website. The fake articles are collected from other websites such as Politifact and Wikipedia which are used as fact checking websites in the United States. There are around 25000 articles with almost equal proportions of real and fake news dataset. The real articles are stored as “True.csv” while the false articles are stored in the file named “Fake.csv”. The datatype of these dataset is text format with attributes: article title, text, type and the date the article was published on. The dataset is mostly clean and no missing values are present.

A

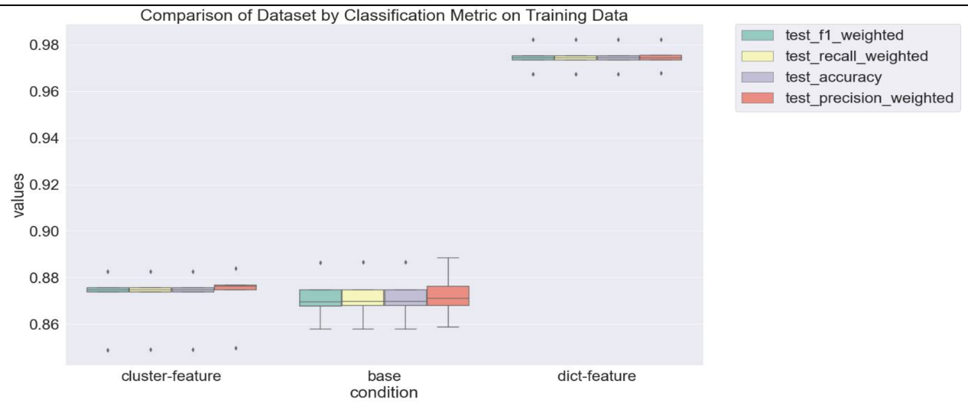


```
news.head(10)
```

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL
5	6903	Tehran, USA	\nI'm not an immigrant, but my grandparents ...	FAKE
6	7341	Girl Horrified At What She Watches Boyfriend D...	Share This Baylee Luciani (left), Screenshot o...	FAKE
7	95	‘Britain's Schindler’ Dies at 106	A Czech stockbroker who saved more than 650 Je...	REAL
8	4869	Fact check: Trump and Clinton at the 'commande...	Hillary Clinton and Donald Trump made some ina...	REAL
9	2909	Iran reportedly makes new push for uranium con...	Iranian negotiators reportedly have made a las...	REAL

This study is not intended to compare different classification models. Instead it compares the effect of text clustering as a pre-processing step.

B



C

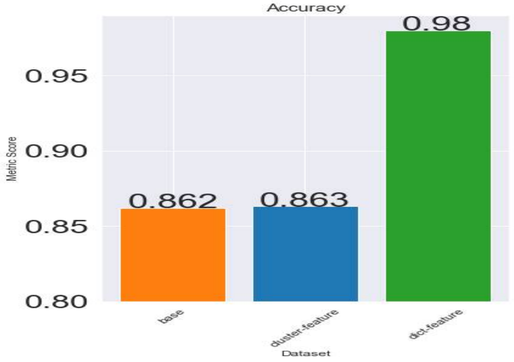


Fig. 3 A B C is Combine Output of the Research work that is to be given the satisfactory about the scope of outcome.

5. CONCLUSION

As a human behavior the fake are more attractive then a true term because fake text impact on human thought process. Machine learning techniques with the support of innovative human insights are getting more effective in text labeling so that they are now being used to classify text and this case Here with the help of artificial intelligence it is possible to control and minimize the flow of wrong text. As the proposes process the deep learning used verbalization and holding dataset for checking words for prediction and correction.With the help of text clustering and anti-dictionary feature, this project tries to improve text classification methods. These two are novel approaches in text pre-processing steps that significantly improves data quality of text classification and hence, provide improved results.

REFERENCES

[1] Mamitsuka, N.A.H. Query learning strategies using boosting and bagging. In Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98); Morgan Kaufmann Pub.: Burlington, MA, USA, 1998; Volume 1.

[2] Kim, Y.H.; Hahn, S.Y.; Zhang, B.T. Text filtering by boosting naive Bayes classifiers. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; pp. 168–175.

[3] Dr. Amit Sharma " Online K-means clustering with adaptive dual cost functions " 2018 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) , DOI: 10.1109/ICICICT1.2017.8342665 , INSPEC Accession Number: 17720894

[4] Dr. Amit Sharma " KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering" IEEE, 2018 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) DOI: 10.1109/ICICICT1.2017.8342664, INSPEC Accession Number: 17732058

[5] Vijay Malav, Sarita Naruka, Dr. Amit Sharma. (2020). Improve Prediction Analysis on Cloud Through Optimization Soft Hybrid Classification. International Journal of Advanced Science and Technology, 29(12s), 3272 - 3277. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/35936

[6] *Sivic, Josef (April 2009).* "Efficient visual search of videos cast as text retrievalIEEE Transactions on Pattern Analysis and Machine Intelligence, 2004. IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(4), pp.0_2-0_2.

[7] Robertson, S. *(2004). "Understanding inverse document frequency: On theoretical arguments for IDF". Journal of Documentation.* ***60*** *(5): 503–520.*

[8] Annie Syrien and M. Hanumanthappaat.al **"**Evaluation of Supervised Classification Techniques on Twitter Data using R**"** International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-10 Issue-8, June 2021

[9] SikhaBagui and Debarghya Nandi at. al. "Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding" Journal of Computer Science 2021, 17 (7): 610.623 , DOI: 10.3844/jcssp.2021.610.623

[10] P. Arumugam and V. Kadhirveni at. al. "Prediction, Cross Validation and Classification in the Presence COVID-19 of Indian States and Union Territories using Machine Learning Algorithms" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-10 Issue-1, May 2021

[11] Vijay and Dr. Pushpneel Verma "Linear Discriminant Analysis for Hate Speech Text Classification" International Journal of Engineering Development and Research ISSN: 2321-9939 , Volume 9, Issue 2 ©IJEDR 2021 Year 2021

[12] Hao Zhang and Yanchun Liang at. al “Deep Feature-Based Text Clustering and Its Explanation” Journal of Latex Class Files, vol. 14, No. 8, Jul 2020, DOI 10.1109/TKDE.2020. 3028943, IEEE

[13] Anastasiu, D. C., Tagarelli, A., Karypis, G. (2014). Document Clustering: The Next Frontier. In Aggarwal, C. C. & Reddy, C. K. (Eds.), Data Clustering, Algorithms and Applications (pp. 305–338). Minneapolis: Chapman & Hall.

[14] Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 42(5), 2785–2797.

[15] Wang, G., Zhang, X., Tang, S., Zheng, H., & Zhao, B. (2016). Unsupervised clickstream clustering for user behavior analysis. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (p. 225–236).

[16] Jernite, Y., Bowman, S. R., ,& Sontag, D. (2017). Discourse-based objectives for fast unsupervised sentence representation learning. CoRR, 2(2), 758-786.

[17] Agrawal, A., & Gupta, U. (2014). Extraction based approach for text summarization using
k-means clustering. In Proceedings of the international conference on information and knowledge management (Vol. 4, p. 9–12).

[18] Mikolo, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of wordrepresentations in vector space. In In proceedings of workshop at iclr (Vol. 1, p. 89–152).

[19] Anita Kumari Singh, Mogalla Shashi (2019) Vectorization of Text Documents for IdentifyingUnifiable News Articles , IJACSA (Vol. 10, No. 7, P 305-310)