CLICKBAIT DETECTION SYSTEM USING MACHINE LEARNING

[1]MEDAM SIVA SHANKAR, [2]D. RAMMOHAN REDDY

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate Professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: *Clickbait (headlines) make use of misleading titles that hide critical information from or exaggerate the content on the landing target pages to entice clicks. As clickbaits often use eye-catching wording to attract viewers, target contents are often of low quality. Clickbaits are especially widespread on social media such as Twitter, adversely impacting user experience by causing immense dissatisfaction. Hence, it has become increasingly important to put forward a widely applicable approach to identify and detect clickbaits. In this paper, we make use of a dataset from the clickbait challenge 2017 (this http URL) comprising of over 21,000 headlines/titles, each of which is annotated by at least five judgments from crowdsourcing on how clickbait it is. In this research we seek to construct an efficacious computational system primarily using ensemble learning. Several Machine learning algorithms such as Logistic regression, Stochastic Gradient Descent and Random Forest have been used. We attempt to build an effective computational clickbait detection model on this dataset.*

Keywords: *Clickbait, machine learning, Logistic regression, Stochastic Gradient Descent and Random Forest.*

I. INTRODUCTION

The media landscape is currently undergoing tremendous changes with news format moving from paper to online media. As such, online new headlines are being optimized in real-time, recasting teaser messages to maximize click-through. Such online headlines are different from the traditional printed frontpage headlines, in which feedbacks contributing to the newspaper sales is often indirect, delayed, and incomplete. Hence, online news providers are trying to make headlines to be more attractive. In particular, some headlines are carefully worded to be eye-catching and often misleading, resulting in unsatisfactory user experience on social media. In addition, the content on landing pages (by yellow journalism) are of low quality and significantly under-deliver the content promised by the exaggerated headline.

We informally consider this type of online news headlines as clickbait. According to the Oxford English Dictionary, clickbait is defined as “content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page." Clickbait usually leads to a site by withholding the promised “bait." Typical example clickbaits are as follows: “What This Person Did for That Will AMAZE You!” or “9 out of 10 Americans Are Completely Wrong About This Mind-Blowing Fact." This type of irresponsible reporting not only frustrates users, but also violates journalistic code of ethics. Scholars have argued that current trend towards a merging of commercial and editorial interests is detrimental to democratic values. In this paper, Couldry and Turow offer a preliminary discussion of clickbait as an example of false or misleading news, and review the identifying characteristics and potential methods to detect clickbaits. Due to the importance of the problem and its implication, in recent years, much research has investigated on the detection of clickbaits. For instance, Gianotto, implements a browser plug-in that transcribes clickbait teaser messages based on a rule set so that they convey a more truthful, or rather ironic meaning. Beckman, Mizrahi, Stempeck, Kempe manually re-share clickbait teaser messages, adding spoilers.

Click baits are used to increase the click throughs of a website for generating revenue. When going through such content people often end up jumping on the wrong conclusions. These developments have been a reason to cause enormous concern among many prominent authors and writers, because this can potentially shut down social media channels and it also defies ethical journalism. Click baits are basically those content on the internet which attract the attention of the viewer. Click baits may be graphical or textual in format, however, in this paper we’ll be limiting ourselves with textual click baits and their detection. Some popularly known websites that host such content are Upworthy and BuzzFeed. Besides these two there are several other hosting with similar kinds of content. Almost all click baits reside on two facts. First, they will always have incomplete or misleading information used to guarantee an emotional response which is often not served once the readers actually goes through the content. Second, they'll always try to exploit the curiosity gap in humans. Below is the list of some commonly used click bait.

To address the problem of clickbait detection, capitalizing on informative patterns found in previous works, we propose a set of new features extracted from post text (e.g., tweets), target paragraphs, keywords of target paragraphs, similarity between post text and target content. Experimental results show that our model achieves 0.035 in MSE, 0.82 in Accuracy, and 0.61 in F1-score on clickbait class, respectively. Our main contribution is the extraction of novel features for clickbait classification which have not been previously studied, and show that these features are among the most effective indicators for clickbait headlines. While still useful, however, much more research is needed toward the development of a detection method that is practical enough in real settings.

II. LITERATURE SURVEY

There have been extensive number of studies on evaluation of the authenticity of news or articles on the web, especially fake news. But clickbait does not necessarily have false content. It usually appears as misleading and exaggerating headline or title on top of a genuine article. In fact, the concept of clickbait is originated from the advent of tabloid journalism focusing on sensationalization of soft news, which was claimed by Rowe on the properties of making changes on professional journalism. In the field of psychology, information-gap theory was put forward to explain the curiosity arising as a gap in one’s knowledge, of which clickbait exploits to get more clicks. According to Loewenstein, riddles, events with unknown sequences, expectation violations or forgotten information are identified as stimuli that may spark involuntary curiosity. There have been also studies on analyzing structure of headlines which contain the properties such as sensationalism, luring, dramatization, emotionalism, etc.. Recently, there have been several literatures focusing on clickbait detection. Martin et al. claimed to introduce the first machine learning approach to the clickbait detection problem, especially in the social network context. They were collecting the first clickbait corpus of 2992 tweets, comprising of 767 clickbait tweets annotated by three assessors. Markus prevented links that relate to a fixed set of domains from appearing on the users’ news feeds. This rule-based approach is not scalable and may require continuous tunings accordingly with the emergence of new clickbait phrases. This might also block texts that are not necessarily clickbaits. Abhijnan et al. [1] achieved 93% accuracy in detecting clickbaits. However, they do feature analysis on the whole dataset instead of a separated training set. Md Main Uddin Rony et al. [2] only considered features extracted from the titles. However, our method considers the features from title, target, as well as the relation between them.

DATASET USED

The data used for this research consists of 24781 tuples and 4 attributes namely index, news id, title and text. We aim to use news title and new text to assign it one of the three classes – click bait, news and other. The testing data consists of 5647 tuples and 3 attributes. Fig 1 provide some insight

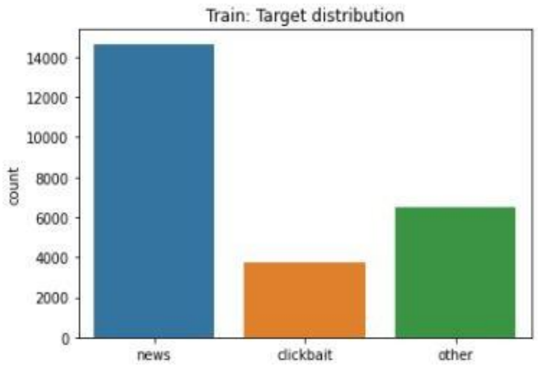


Fig.1 shows the overall diversity of the training data. In this research we have used supervised learning and count of each label is indicated in that graphic

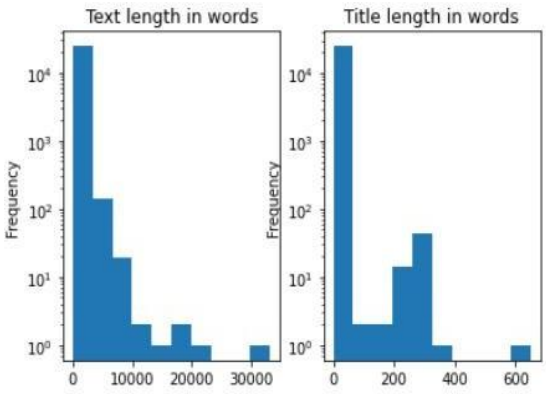


Fig.2 depicts the frequency of the words along with their length present in the given dataset

III. PROPOSED SYSTEM

In this research we have used Ensemble learning comprising of three machine learning algorithms namely logistic regression, stochastic gradient descent classifier and random forest. The following flowchart depicts the workflow of the research

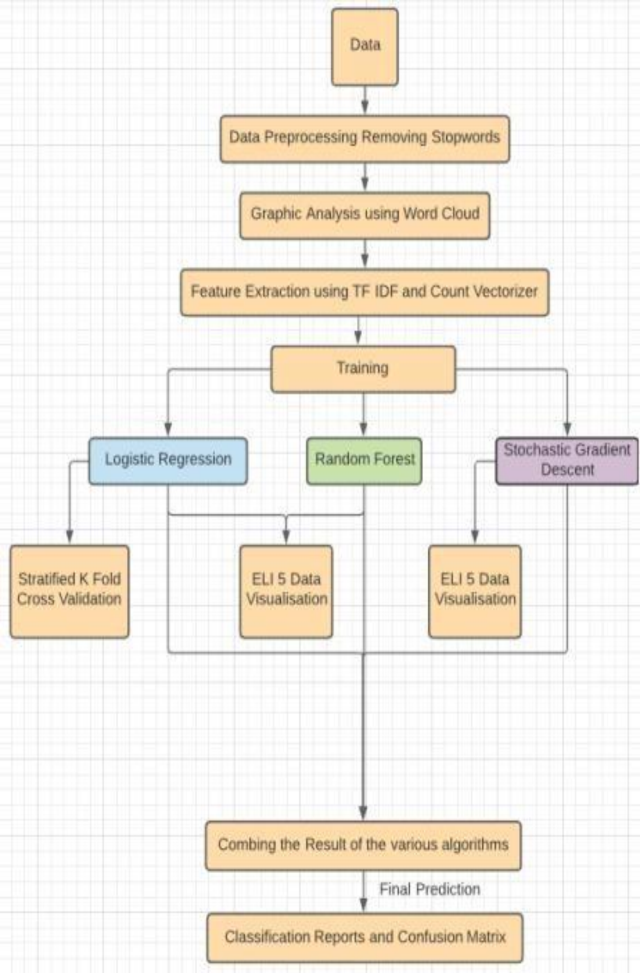
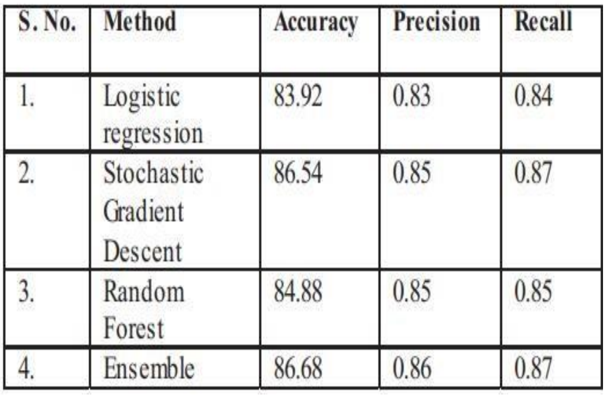

Data
Data Preprocessing Removing Stopwords
Graphic Analysis using Word Cloud
Feature Extraction using TF IDF and Count Vectorizer
Training
Logistic Regression
Random Forest
Stochastic Gradient Descent
Stratified K Fold Cross Validation
ELI 5 Data Visualisation
ELI 5 Data Visualisation
Combing the Result of the various algorithms
Final Prediction
Classification Reports and Confusion Matrix

Fig.3 System architecture

The various algorithms used in this research and their individual results along with the result of ensemble of all of them are depicted in the following table

TABLE I. COMPARSION OF VARIOUS ALGORITHMS USED



S. No.	Method	Accuracy	Precision	Recall
1.	Logistic regression	83.92	0.83	0.84
2.	Stochastic Gradient Descent	86.54	0.85	0.87
3.	Random Forest	84.88	0.85	0.85
4.	Ensemble	86.68	0.86	0.87

FEATURE ENGINEERING

Feature Engineering is a process that involves usage of knowledge inferred from data. The knowledge obtained is used to synthesize explanatory parameters as well as features which in turn help us build a predictive model specifically a classifier in this case. We've several motivations when talking about feature engineering. When we identify essential parameters the overall predictive ability of the model improves significantly. Another fundamental goal of producing less complex and computationally less expensive models with impressive predictive ability is also achieved. As in the case of any other machine learning project, we always first get to deal with raw data which is a complete mess and haywire. The whole process commences with cleaning of this data which involves a lot of steps like removing missing values and also changing the data kind / types of the provided input. Further we go ahead removing some outliers or some useless features that hold little importance to our model. Once the data is cleaned, we can explore the data well and even go ahead create new features that better help in analysing the given problem statement, facilitating a smooth learning process for our machine learning model and improving the performance / accuracy.

DATA PREPROCESSING

Removing Stop words Before we discuss their removal first let's understand stop words. Stop words are some commonly occurring words in any given language and these words don't help much in determining the nature of text or text analyses. Whenever dealing with machine learning models it becomes important to get rid of them. Generally, the most common words used in a text are "a", "I", "while", "for", "where", "when", "to", "at" etc. Consider the following example "There is a pencil in the pencil box". Now, the words "is", "a", "in", and "the" add no meaning to the sentence. Whereas words like "there", "pencil", and "pencil box" are the unique keywords and on parsing them we get useful results. NLTK generally provides us with a bunch of stop words from 16 different languages and we did use the same in the research. Just to be more accurate we even added more stop words from our side which would be evident later.

Logistic Regression in Click bait detection

Logistic function is basically used to obtain the probability of a certain class. We can also use it to get the probability of events like whether something will pass or fail, someone will win or lose, something is dead or alive or maybe healthy or sick. It is also known as the Logit Model. Logistic regression derives its root from the Logit Model or logistic function. It's a statistical model which has one or many independent variables of various kinds but essentially a dependent binary variable although more complex alterations do exist. We can also have regression analysis here; logistic regression estimates the parameters of a logit model. This task is similar to a form of binary regression. Mathematically, a binary logistic model will always have a single dependent variable with two possible values and we generally label them as "True" and "False" or sometimes as 0 or 1. When dealing with the logistic model, the logarithm of the odds for the value labelled as true / 1 comprises a linear combination of one or more variables which are independent and are also called predictors and these variables can each be a binary kind variable or a continuous kind variable. Figure 6 presents the confusion matrix and the accuracy obtained.

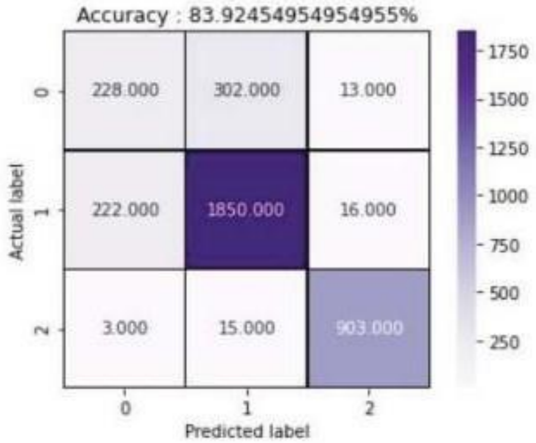


Fig.4 shows the confusion matrix for Logistic regression

Stochastic Gradient Descent in click bait detection

Stochastic Gradient Descent (SGD) is a machine learning algorithm commonly used to improve performance of the standard gradient descent algorithm [10]. This algorithm is commonly used in other machine learning algorithms and it also underpins neural networks. Gradient means slope or the slant of a surface at a given point. So gradient descent as the name suggests means descending slope to reach the lowest point on the surface. Stochastic Gradient descent is similar to Gradient descent except the fact that it randomly chooses a point and then starts executing in an iterative fashion trying to find the lowest point. We decided to use SGD because there are a few drawbacks of the gradient descent algorithm and that would become evident once we take an example to show the number of computations involved in each iteration of the gradient descent algorithm. Let us assume that we have 5,000 data points and 10 features. Now Gradient descent algorithm demands computation of derivative of this function with respect to each and every feature, so in total we end up doing 5000 * 10 = 50,000 computations and that is just for 1 iteration. Generally, we end up with 1000 iterations and that will be an absolutely safe and realistic assumption so in total we have 50,000 * 1000 = 50000000 computations to complete the algorithm. The number of computations involved are too large and thus gradient descent is computationally expensive whereas SGD reduces all these computations by a large amount. Figure 7 presents the confusion matrix and the accuracy obtained

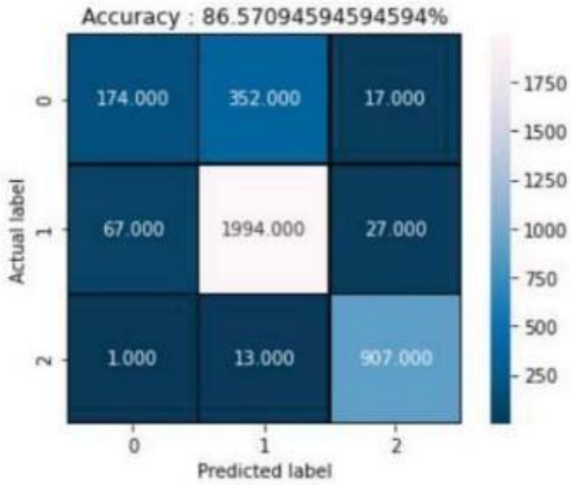


Fig.5 shows the confusion matrix for stochastic gradient descent

Random Forest in click bait detection

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A Decision tree is a rooted tree used as a predictive model. We express a decision tree by partitioning the feature space recursively into many subspaces and each of these subspaces form a basis of prediction. In a decision tree a set of features are used in a hierarchical fashion in such a way that after each split the entropy of the system minimizes. We can have a split criterion for the internal nodes. Finally, each leaf is assigned to one class or its probability. Decision tree suffers from Overfitting as even small variation in the dataset lead to large errors and to deal with the same methods like pruning of trees is adopted. A Random forest basically constitutes an ensemble of a large number of these decision trees. Every Decision tree in random forest provides an output to the given input data utilizing all the available features and finally a voting is done and class occurring the most frequently becomes the final output of the random forest.

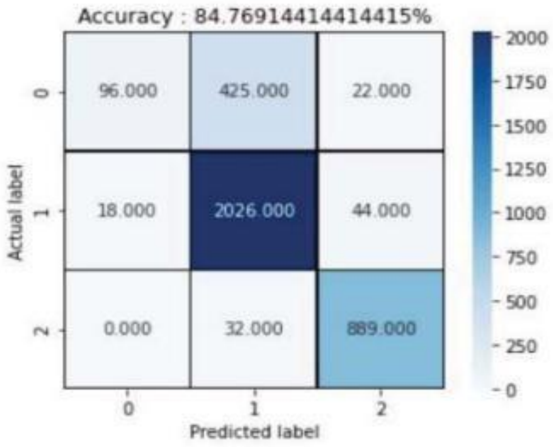


Fig.6 shows the confusion matrix for random forest

IV. CONCLUSION

In this paper we used Ensemble learning to detect click baits. This paper provides an intuitive method of the combining the results by giving preference to the model with highest accuracy in case of ambiguity. First, we resorted to hard voting but in order to avoid clashes we have more preference to Stochastic Gradient Descent classifier due to its higher accuracy owing to the fact that it implements mini batch technique to find the optimal parameters. Feature extraction was done using TF IDF vectorizer as well as Count vectorizer. Various machine learning algorithms were used and a combine function was defined to create ensemble using hard voting mechanism. Model Validation has been done with Stratified K Fold cross validation. This research paper focuses on ensemble learning to ensure that algorithms are faster and can be applied in various social media platforms as well as news articles. The final accuracy so obtained was 86.68%

REFERENCES

[1] Potthast M., Köpsel S., Stein B., Hagen M. (2016) Click bait Detection. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_72

[2] A. Agrawal, "Click bait detection using deep learning," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2016, pp. 268-272, doi: 10.1109/NGCT.2016.7877426.

[3] K. Shu, S. Wang, T. Le, D. Lee and H. Liu, "Deep Headline Generation for Click bait Detection," 2018 IEEE International Conference on Data Mining (ICDM),

Singapore, 2018, pp. 467-476, doi: 10.1109/ICDM.2018.00062.

[4] H.-T. Zheng, J.-Y. Chen, X. Yao, A. Sangaiah, Y. Jiang, and C.-Z. Zhao, “Click bait Convolutional Neural Network,” Symmetry, vol. 10, no. 5, p. 138, May 2018.

[5] A. Geçkil, A. A. Müngen, E. Gündogan and M. Kaya, "A Click bait Detection Method on News Sites," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, 2018, pp. 932-937, doi: 10.1109/ASONAM.2018.8508452.

[6] S. Chawda, A. Patil, A. Singh and A. Save, "A Novel Approach for Click bait Detection," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 1318-1321, doi: 10.1109/ICOEI.2019.8862781.

[7] A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, "Stop Click bait: Detecting and preventing click baits in online news media," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 9-16,doi: 10.1109/ASONAM.2016.7752207.

[8] Chen, Y. and Rubin V. L. (2017). Perceptions of Click bait: A QMethodology Approach. In the Proceedings of the 45th Annual Conference of The Canadian Association for Information

[9] Prasadu Peddi (2022), A Hybrid-Method Neighbor-Node DetectionArchitecture for Wireless Sensor Networks, ADVANCED INFORMATION TECHNOLOGY JOURNAL ISSN 1879-8136, volume XV, issue II.

[10] Prasadu Peddi (2023), Using a Wide Range of Residuals Densely, a Deep Learning Approach to the Detection of Abnormal Driving Behaviour in Videos, ADVANCED INFORMATION TECHNOLOGY JOURNAL, ISSN 1879-8136, volume XV, issue II, pp 11-18.