A Comparative Analysis Of Classifiers Used For Detection of Clickbait In News Headlines

Aaryaman Bajaj[1], Himanshi Nimesh[2], Raghav Sareen[3], Dinesh Kumar Vishwakarma[4]
[1,2,3,4] Department of Information Technology, Delhi Technological University, Delhi, India
[1]aaryamancode@gmail.com, [2]nimeshhimanshi@gmail.com, [3]raghu99sareen@gmail.com, [4]dinesh@dtu.ac.in

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432123

Abstract— **Manipulative clickbait headlines in news articles represent one of the many ways in which misinformation is spread on the Internet. In times of increasing dependence on the Internet for keeping us updated about the happenings in the world, detection of clickbait has become a task of extreme relevance. This paper compares the performance of different classifiers in detecting the clickbait headlines of news articles by performing the extraction of new features from a multi-source dataset. Random Forest classifier yields a better accuracy than Naïve Bayes and Logistic Regression models in identifying headlines disseminating misleading information. Our research also illustrates that a certain set of minimal features is sufficient to achieve the given objective.**

Keywords— Clickbait, Precision, Accuracy, Naive Bayes, Random Forest, Logistic Regression.

I. Introduction

Clickbait takes advantage of the viewer's curiosity to get into details and know more about what the content exactly is. Generally, the page to which the viewer is forwarded is entirely different from what the headline was, or it is to get him to register for something, or it is merely an advertisement strategy. Clickbait websites exploit an innovative method known as Curiosity Gap [1], where the headlines are used to build curiosity in reader's minds to make them click on the links. It is made to create and capitalise on this information gap. With an increase in the number of people having access to the internet, this seems like a sly plan of action to increase web-counters and increase revenue. News outlets have adopted the online front to stay in business. They usually generate revenue by advertisements flashing on websites, or a model which require subscription for articles to make user's interest [1]. In addition to being a means of spreading fake news, the use of clickbait to entice users is also an unethical practice. Each day, the number of people having social media presence is increasing, and many of them do not see through the trap put in place and end up visiting a webpage that they did not mean to, just because they found a particular headline catchy and gave in to their temptation to know about it in depth.

With everything going digital these days, media outlets compete with one another to make their articles popular. The writers have begun employing various different ways to increase the number of people visiting their news articles. One way is to write clickbait-y headlines, which give cues to the user and not the entire content at once. This practice exploits the "curiosity gap" in such a way that the owner of the website generates revenue out of the user's curiosity.

It is required to detect clickbait because with an increase in the popularity of social media platforms and the number of people turning to digital means for information, the amount of data generated is growing along with its complexity. The dirty information may be in any form like fake review, fake news, satire, hoax, etc. that may affect the humanity in a wrong way [2]. It is becoming difficult to separate relevant data from manipulative misinformation. Such misinformation present on social media has the potential to influence judgement and disseminate fake news widely [3]. It is high time to develop a mechanism to deal with this issue to shield people from the adverse effects of such targeted manipulation. Identification of content which uses clickbait plays a major role in tackling the spread of misinformation via various media platforms and news outlets.

II. Literature Review

Identification and classification of clickbait content is an active field of research. Researchers are employing various approaches to combat the effect of clickbait on the overall experience of the users. Some of them used textual content only, whereas some also considered images. Deep learning can also be implied to as Deep Learning performs significantly better with the dataset which contains images, speeches, languages, etc. [4]

Potthast et al. [5][6] proposed the first method to tackle clickbait in 2016. 2992 tweets from the top users were collected, out of which 767 had clickbait content. 215 features were used to make the model. Random Forest Classifier scored 0.79 ROC-AUC with an accuracy of 0.76, performing better than Logistic Regression and Naïve Bayes in the test results.

According to a recent study by Rony et al [7], 166 million Facebook posts from 152 media organization accounts were taken into consideration. The authors trained their system using the above dataset and built their methods with the help of distributed sub-word embeddings instead of Bag-of-Word due to the popularity and accuracy of deep learning methods in text classification.

Anand et al. [8] used Recurrent Neural Networks to detect clickbait news by employing the dataset used by Chakraborty et al. The authors developed a chrome extension that gives a warning if the user's web page has clickbait content [9]. It also gives the option to obstruct the clickbait content on the next visit to the webpage.

Attention Mechanism has been used for different kinds of text classification tasks, such as aspect-based sentiment analysis and fake news identification [10]. The first example of clickbait detection in social platforms can be seen in hand-weaven linguistic features, making use of a reference dictionary of clickbait phrases, on a dataset of crowd-sourced tweets [7].

A Bauhaus-Universität Weima study used 2993 tweets on a model of 216 features and had an F1 score of 76% [6]. To develop an effective mechanism, data from WikiNews and various other websites which had clickbait content was used. 14 features were extracted such as the structure of the sentence, word patterns, etc. using Stanford CoreNLP tools. SVM outperformed Decision Tree and Random Forest in the test results having accuracy as 0.93. Recall and Precision were 0.9 and 0.95 respectively. Table 1 presents a brief overview of important literature on detection of clickbait.

TABLE I. A SHORT LIST OF CLICKBAIT LITERATURE REVIEW

	Method	**Dataset**	**Result**
Clickbait Detection	Random Forest with 215 features	Clickbait: 767	ROC-AUC: 0.79
		Non - Clickbait: 2225	Precision: 0.76
			Recall:0.76
Stop Clickbait: Avoiding Clickbait in Online News Media	Support Vector Machine with 14 features	Clickbait: 7500	ROC-AUC :0.93
		Non - Clickbait: 7500	Precision: 0.90
			Recall:0.95
Clickbait Detection Using Machine Learning	Convolution Neural Network (CNN)	Clickbait: 814	ROC-AUC: 0.90
		Non - Clickbait: 1574	Precision: 0.895
			Recall:0.90

III. METHODOLOGY

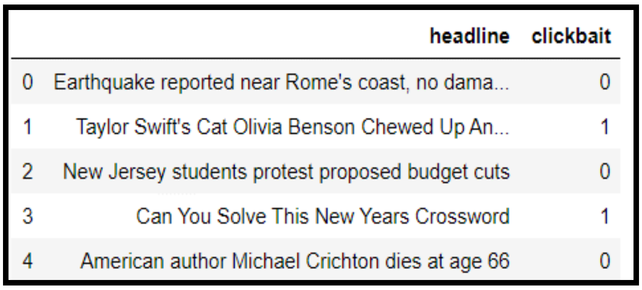
A. Dataset

The dataset employed contains headlines from various news sites such as 'Wikipedia', 'NY Times', 'The Guardian', 'The Hindu', 'BuzzFeed', 'Upworthy', 'ViralNova', 'ScoopWhoop' and 'ViralStories'. It has a dual column structure.

'Headline' column contains all the headlines from news sites in text format. 'Clickbait' columns contains numerical labels '1' or '0' indicating the presence or absence of clickbait in the corresponding headlines.

The dataset consists of 32000 rows of which 50% are clickbait and other 50% are non-clickbait.

Fig. 1 is a screenshot of the dataset, which shows a few of the headlines present in the dataset.



	headline	clickbait
0	Earthquake reported near Rome's coast, no dama...	0
1	Taylor Swift's Cat Olivia Benson Chewed Up An...	1
2	New Jersey students protest proposed budget cuts	0
3	Can You Solve This New Years Crossword	1
4	American author Michael Crichton dies at age 66	0

Fig. 1. View of the dataset

B. Steps involved in Data Cleaning

1) Data Cleaning: Unprocessed Data may have essential fields empty. This causes inconsistencies in the dataset. The data cleaning step takes care of all these issues.

By following proper criteria, these missing values are filled and the dummy values are removed.

2) Creating Test and Training Sets: We divide our dataset into three parts: - training (80%), testing (10%), and validation (10%).

3) Feature Scaling: One variable starts dominating other variables when the independent variables are measured at different scales. Feature scaling allows changing all data values according to a certain range. By doing so, all the variables contribute equally in training the dataset.

There are two majorly used methods of feature scaling:

- Standardization:

$$X_{stand} = \frac{(X - X_{mean})}{(X_{standard\ deviation})} \quad (1)$$

- Normalization:

$$X_{Norm} = \frac{(X - X_{min})}{(X_{max} - X_{\mathrm{min}})} \quad (2)$$

C. Steps involved in Model Development

1) Preprocessing and Analysis

We implemented the following operations on each string:

- conversion to lower-case
- expansion of contractions
- removal of punctuation
- lemmatization of words

For each headline, we also counted the number of contractions as Clickbait headlines have a higher ratio of contracted words. We defined all the parsing functions we need.

2) Bag of Words

It contains words which are commonly used in Clickbait titles, and these make no reference to what the article is actually about. After removing stop words, we picked out 100 words which were used repeatedly in Clickbait headlines and the 200 frequently appearing words in non-Clickbait titles.

Table 2 shows the most common words used in headlines. Some of the frequent words used are ‘you’, ‘this’ and ‘people’.

TABLE II. 14 MOST COMMON CLICKBAIT WORDS (IN ALPHABETICAL ORDER)

S.No.	No Clickbait	Clickbait	Full
0	after	about	and
1	and	and	at
2	as	be	be
3	at	do	do
4	be	for	for
5	by	have	from
6	for	in	have
7	from	make	in
8	in	sum	num
9	kill	of	of
10	new	on	on
11	num	that	that
12	of	the	the
13	on	thing	this

3) Pipeline Creation

We then make a pipeline in which all the features are generated. Along with the “Count of Words”, we introduce 6 more features, mentioned as follows:

- Length of Headline
- Word Length
- Ratio of Stopwords
- Ratio of Contractions
- Use of Possessives, Subjects and Determiners.
- A flag indicating if a number starts a headline.

a) Headline Length: It is generally seen that the length of clickbait headlines is larger than non-clickbait ones. In our dataset, it is seen that clickbait sentences have an average length of 9.94, whereas non-clickbait sentences have an average length of 8.18.

Clickbait headlines contain more function words, and have well-formed sentences, such as “These pictures portray Donald Trump like you’ve never seen him before!”. Traditional headlines do not have such function words in them, and have a short length. These sentences contain specific locations or people, such as “Coronavirus-induced lockdown brings country to grinding halt”.

b) Length of Words: It is seen that though clickbait headlines are longer in length, the size of the word used in it are shorter. The main reason for shorter length of words in clickbait headlines is the use of contraction words. In our dataset the mean length of word came out to be 6 for non-clickbait and 4.5 for clickbait headlines.

c) Stopwords: These words can be identfied as the words which are repetitive in nature in a particular language. They come more often in clickbait headlines. On the contrary, genuine headlines contain more content words. Due to this clear demarcation, it's easy to figure out a clickbait out of the bag because of the anomalous proportion of stopwords' contribution to sentence semantics. This renders stopword detection to be a crucial step in identifying clickbait titles.

d) Contraction Ratio: Contractions are short forms of words that omit certain letters or sounds. The most extensively used symbol in place of missing words is an apostrophe. For example: I will = I'll, She would = She'd. It has been identified that clickbaits are guilty of using more contractions than regular content. So, this gives us another feature to identify clickbait headlines.

e) Use of Possessives, Subjects and Determiners: Subject words such as 'I', 'We', and 'You' are used to address second and first persons in a clickbait headline whereas non-clickbait titles are referenced to in third person [6]. Determiners are used to make users curious about the article being referred and to persuade them. Some of the determiners used are 'there', 'my' and 'which'.

f) A flag indicating if a number starts a headline: It has been observed that to lure the users, clickbait makers start their headlines with a number to instantly catch the user's attention as our brain is made such a way that it attracts to numbers and figures more quickly.

We have excluded some features from our procedure, so as to improve the working of others. One example of this is the exclusion of punctuation patterns. This is because we lemmatised each sentence, and in turn to form the bag of words, we did not take into consideration punctuation patterns. Other such features are:

g) Common Phrases: Several clickbait headlines contain similar phrases like, "You will not believe this!", or "Will blow your mind!". If such phrases are part of a headline, chances are that the link is a clickbait. Downworthy [11] included these kinds of phrases in their model to detect clickbaits.

h) Punctuation: Clickbait headlines usually have a lot of informal punctuations such as exclamation marks '!', question marks '?', a series of full stops '...', or stars '***' to hide something offensive or to get their attention.

i) Internet Slangs: These are words that are commonly used by people to convey emotions. For e.g., LOL means "laugh out loud", and is used to indicate a funny post. Clickbait headlines often make use of these slangs, such as LMAO, AF etc. to immediately catch the eyes of the reader.

4) Model Training

- Applying Naïve Bayes Algorithm on Bag of Words [12]: It is a probabilistic approach, so we first tried it on miniature training set and then on the full training set. We used its output as a feature for the Random Forest and SVC methods.

- Applying Random Forest on Naive Bayes probabilities and non-word features: First, we train it on the miniature training set. Then we fine tune some hyperparameters, and then use Cross Validation on our complete training examples to get an approximate idea of our conduct.

- Training the SVM on Naive Bayes probabilities and non-word features.

- Training the Naïve Forest on all the features.

- Training Naive SVM on all the features

We also rescaled the features in the procedure.

5) Model Validation and Testing

During Model Validation, we evaluated the classifiers made on the Validation Set and ascertained which one gives us better results. We then proceeded to evaluate the model performance on test set.

IV. Discussion

A. Evaluation

Evaluation involves calculating Precision (P), Recall (R), F-measure (F1 score), and Accuracy. These are formulated using the following equations:

$$Precision = \frac{TP}{(TP+FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$F1 = \frac{2*Precision*Recall}{(Precision + Recall)} \tag{5}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{6}$$

For training our classifiers, we first made miniature training sets of 1000 instances only. Applying Naive Bayes, we measured a Precision of 0.902, Accuracy of 0.874 and Recall of 0.840. For Random Forest on Naive Bayes Probability and Non-Word Features, we obtained an Accuracy of 0.836, Precision of 0.842 and Recall of 0.838. On using the full training set, the results improved to Accuracy of 0.875, Precision of 0.897 and Recall of 0.848.

Finally, we evaluated the three classifiers we made, on the validation set to watch their performance and the results obtained were as discussed below.

B. Results

Table 3 shows the obtained scores on applying the various methods. We got the best results from Random Forests. After tuning the parameters, our measures showed better results.

TABLE III. RESULTS OF VARIOUS METHODS

Method	Precision	Recall	Accuracy
Naïve Bayes	0.893	0.825	0.864
Random Forest	0.908	0.869	0.891
Logistic Regression	0.911	0.856	0.887

V. FUTURE STUDY

There is a wide scope for future work involving different approaches for selecting features and exploring the phenomenon behind the wide range of techniques used by us for detection of clickbait.

One way to go forward is inculcating a proof-of-concept application as shown by Varshney [13]. This will help us to focus on clustering methods. We can also apply neural networks, since they may give better results [14][15].

The number of likes, or re-tweets given to the article in social networking websites, can be added as parameters. There is potential to incorporate some more deciding features to our model such as checking for hyperbolic words, internet slangs, and comparing word n-grams [16]. We can also improve the quality of the model by doing sentiment analysis on the headlines.

Discovering the features using unsupervised ML techniques could yield better results but is impossible to do as of now due to the lack of a bigger and diverse dataset. Also, it would require more time for each of the learning processes to get complete.

We could also try to detect rumours- a similar form of false information whose correctness is not verified at posting time and whose veracity is either ambiguous or not final [17]. The components that can be used for rumour analysis are based on Text are Rumour Detection (RD) and Veracity Assessment (VA) [18].

VI. CONCLUSION

Nowadays reportage is more vulnerable to exaggeration and misinformation. In our research, we tried a different method to classify clickbait headlines. We tried to differentiate between clickbait and non-clickbait elements and also showed how new features could be extracted from different parts of the headlines. However, the work is far from over. New clickbait formats are added each year, and many new methods can be incorporated into our model, to further improve accuracy.

We showed that the clickbaits could be detected with the help of a minimal number of features and compared different models in our work. The conclusions we came to are as follows:

- There is high degree of similarity between the evaluation performance of our proposed model and other existing models.

- Clickbaits can be classified with the help of fewer number of features, which can easily be used to make a real application (such as a browser extension) to move this idea from theory.

REFERENCES

[1] George Loewenstein. 1994. The Psychology of Curiosity: A Review and Reinterpretation." Psychological bulletin, vol. 116 (07 1994), 75–98.

[2] P Meel, DK Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities" in Expert Systems with Applications, Elsevier, p 112986, 2019.

[3] DK Vishwakarma, D Varshney, A Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating

the web search” in Cognitive Systems Research 58, pp 217-229, 2019.

[4] P Meel, F Bano, D Vishwakarma, K Dinesh, “The Problem of Fraudulent Content on the Web: Deep Learning Approaches” in The Problem of Fraudulent Content on the Web: Deep Learning Approaches, 2020.

[5] M. Potthast, S. K¨opsel, B. Stein, and M. Hagen, “Clickbait detection,”in Advances in Information Retrieval. Springer, 2016.

[6] Martin Potthast, Sebastian Köpsel, Benno Stein and Matthias Hagen(2016), Clickbait Detection, Bauhaus-Universität Weimar.

[7] M. U. Rony, N. Hassan, and M. Yousuf, “Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?,” in 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2017.

[8] A.Anand, T.Chakraborty, N.Park, “We used Neural Networks to Detect Clickbaits: You won't believe what happened Next!” in arXiv:1612.01340v2 [cs.CL], 2019.

[9] A.Chakraborty, B.Paranjape, S.Kakarla, N.Ganguly, “Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media” in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[10] Y.Zhou, “Clickbait Detection in Tweets Using Self-attentive Network” in The Zingel Clickbait Detector at the Clickbait Challenge, 2017.

[11] A. Gianotto, “Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean,” downworthy.snipe.net/.

[12] P Meel, M Mishra, D Vishwakarma, K Dinesh, “A Contemporary Survey of Machine Learning Techniques for Fake News Identification”, in A Contemporary Survey of Machine Learning Techniques for Fake News Identification, 2020.

[13] D Varshney, DK Vishwakarma, “Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles” in Journal of Ambient Intelligence and Humanized Computing, pp 1-14, 2020.

[14] B Malhotra, DK Vishwakarma, “Classification of Propagation Path and Tweets for Rumor Detection using Graphical Convolutional Networks and Transformer based Encodings” in 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp 183-190, 2020.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).

[16] DK Vishwakarma, C Jain, “Recent State-of-the-art of Fake News Detection: A Review” in 2020 International Conference for Emerging Technology (INCET), pp 1-6, 2020.

[17] D Varshney, DK Vishwakarma, “A unified approach for detection of Clickbait videos on YouTube using cognitive evidences” in Applied Intelligence, pp 1-22, 2021.

[18] D Varshney, DK Vishwakarma, “A Review on Rumour Prediction and Veracity Assessment in Online Social Network”, in Expert Systems with Applications, p 114208, 2020.