A Review Paper on Clickbait Detection

Pankaj Saraswat

SOEIT, Sanskriti University, Mathura, Uttar Pradesh, India

Email Id- pankajsaraswat.cse@sanskriti.edu.in

ABSTRACT: With the development of online advertisements, clickbait spread wider and wider. Clickbait dissatisfies users because the article content does not match their expectation. Thus, clickbait detection has attracted more and more attention recently. Traditional clickbait-detection methods rely on heavy feature engineering and fail to distinguish clickbait from normal headlines precisely because of the limited information in headlines. A convolutional neural network is useful for clickbait detection, since it utilizes pertained Word2Vec to understand the headlines semantically, and employs different kernels to find various characteristics of the headlines. However, different types of articles tend to use different ways to draw users' attention, and a pertained Word2Vec model cannot distinguish these different ways. To address this issue, we propose a clickbait convolutional neural network (CBCNN) to consider not only the overall characteristics but also specific characteristics from different article types. Our experimental results show that our method outperforms traditional clickbait-detection algorithms and the TextCNN model in terms of precision, recall and accuracy.

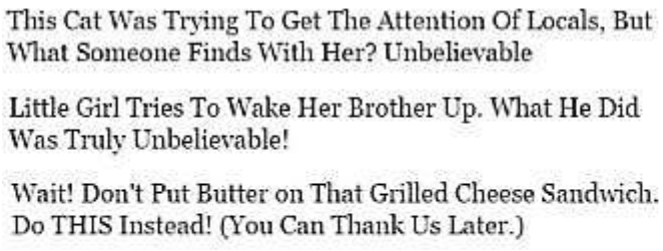
KEYWORDS: Clickbait, Detection, Hierarchical, Hybrid, Network.

1. INTRODUCTION

Clickbait detection is the task of identifying clickbait, a form of false advertisement that uses hyperlink text or a thumbnail link that is designed to attract attention and to entice users to follow that link and read, view, or listen to the linked piece of online content, with a defining characteristic of being deceptive, typically sensationalized or misleading[1].

Clickbait's are internet articles with deceptive names that are designed to get people to visit the intended web page. To monetize the landing page or distribute fake news for sensationalize, clickbait's are used to entice users to click on a certain link. The existence of clickbait's on any news aggregation portal may cause readers to have a negative experience. For the machine learning community, detecting clickbait headlines from news headlines has been a difficult task. In the recent past, many strategies for avoiding clickbait stories have been presented[2]. However, the most recent clickbait detection algorithms are not very reliable. By combining various characteristics, sentence structure, and clustering, this study presents a hybrid classification approach for distinguishing clickbait from non-clickbait publications. The headlines are divided into eleven categories during preliminary classification. Following that, sentence formality and syntactic similarity metrics are used to reclassify the headlines[3]. The headlines are categorized in the final step using clustering based on word vector similarity and the t-Stochastic Neighbourhood Embedding (t-SNE) method. Machine learning models are used to the data set once the headlines have been categorise to test machine-learning techniques. For the real dataset we employed, the experimental findings show that the suggested hybrid model is more resilient, trustworthy, and efficient than any individual classification approach[4].

The popularity of clickbait, which is nothing more than online material that is deceptive, is a prevalent trend in today's internet content nature, with the express purpose of capturing the attention of the audience and enticing people to visit their website. Clickbait is a type of content that is designed to get people to click on it by low quality, low-value material, and the ad agencies who distribute it are highly reliant on ad income. As a result, they develop eye-catching headlines that entice people to click on them, resulting in income[5]. These articles, which often promise a valuable experience or an essential discovery, prey on human psychology and frustrate users since they rarely receive the kind of information they seek. Clickbait on social media has been on the rise in recent years, and even some news publishers have adopted this technique. These developments have caused general concern among many outspoken bloggers, since clickbait threatens to clog up social media channels, and since it violates journalistic codes of ethics.



This Cat Was Trying To Get The Attention Of Locals, But What Someone Finds With Her? Unbelievable

Little Girl Tries To Wake Her Brother Up. What He Did Was Truly Unbelievable!

Wait! Don't Put Butter on That Grilled Cheese Sandwich. Do THIS Instead! (You Can Thank Us Later.)

Figure 1: The Above Figure shows the Examples of Clickbait.

Figure 1 depicts several clickbait instances that could surface during a typical online media browsing session. As can be seen, the titles appear to promise some extremely interesting or instructive material; but, when we click and visit the website, we get just a small amount of useful information. The use of clickbait in social media has increased dramatically in recent years, to the point that certain news organisations are already employing similar tactics.

Teaser messages open a so-called "interest gap," increasing the probability that readers would click the destination link to fulfil their curiosity, which is often credited as the reason why clickbait works. “The information-gap theory views curiosity as originating when attention gets focused on a gap in one's knowledge,” according to Lowenstein (p. 87): “the information-gap theory views curiosity as arising when attention becomes focused on a gap in one's knowledge. Curiosity is a sensation of deprivation caused by knowledge gaps. To decrease or remove the sense of deprivation, the interested individual is motivated to get the missing information.”

Publishers desire more clicks on their web sites as web advertising has evolved over the years, in order to generate ad income. In this case, clickbait occurs on the internet in an attempt to capture readers' attention and persuade them to click the link. To achieve their aims, clickbait typically use sexual phrases, deceptive material, unconfirmed news, and exaggerated tones. Clickbait reduces user happiness by raising article click-through rates (CTR), since consumers perceive a disconnect between what they want to know (the headline) and what they actually read (the content). Furthermore, clickbait[6] contributes to the propagation of bogus news on the Internet since many people forward them without reviewing the content. As a result, it is critical to build clickbait detection tools.

Something meant to entice readers to click on a hyperlink, especially when the link leads to information of doubtful worth or appeal, is referred to as clickbait1. Publishers have been using different tactics to create the Curiosity Gap between the information included in uploaded texts and the information that readers actually want to know since the advent of Twitter. This feature encourages readers to visit the publishers' websites by clicking on the links in the tweets. According to, all of Twitter's top 20 most prolific authors used clickbait on a regular basis, with the ratio of clickbait tweets reaching a staggering 26% of all the tweets they produced.

Even without any major hyperactive parameter adjustment, this model achieves a good level of accuracy having two contributions:

- We offer a clickbait corpus available to the public, culled from various social media sources. Currently, no such corpus exists.
- We use and analyse the first deep learning model for clickbait detection, which has good precision, recall, and accuracy. Our algorithm is able to learn generic characteristics rather than platform-specific features by using a corpus generated from several social media sources. We also contribute by bolstering the argument that unsupervised learning pre-training of word vectors is a valuable supplement to deep learning techniques for NLP.

As illustrated in Figure 2, the suggested model may be split into three components. We begin by compiling a data corpus of clickbait and non-clickbait headlines. The textual headlines are then transformed into word embedding, which are then fed into our deep learning models, in this case a CNN.

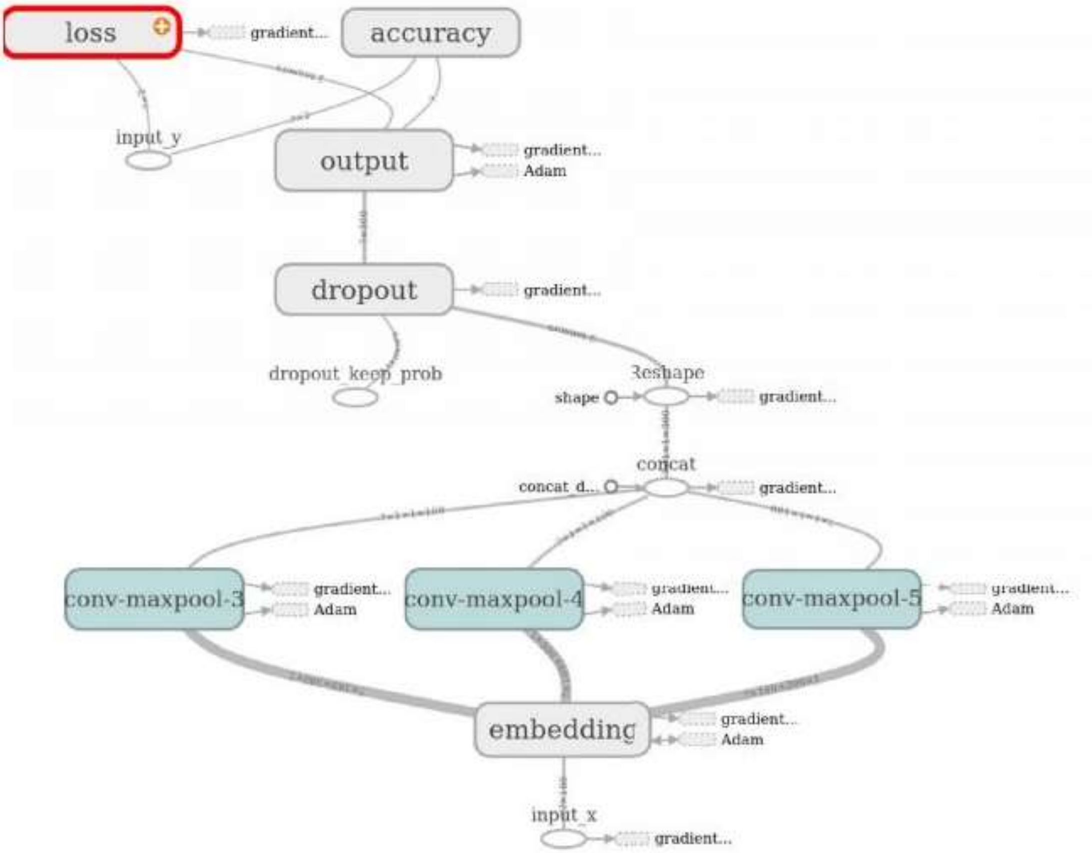


Figure 2: The Above Figure Shows the CNN Model.

1.1 Data Collection:

We develop our own corpus because there is not one available for clickbait. We gathered data from three sources, namely Reedit, Facebook, and Twitter, all of which are prominent social media sites, unlike and who only used data from a single source. This method was used to verify that the characteristics learned by our deep learning model were not reliant on the social media platform. Each social media site has its own set of restrictions, such as Twitter's limit of 140 characters per tweet. As a result, we employed a variety of data sources to train our clickbait detection deep learning model.

1.2 Deep learning Models:

CNN has been used in a variety of deep learning tasks. In this paper, we utilise a basic CNN with only one layer of convolution. Figure 3 depicts a graphical depiction of the whole model that was used. The CNN that we use is based on Kim's CNN architecture. The first layer of the CNN is used to embed words into low-dimensional vectors. We use two types of word embedding:

- Word embedding that are learned from scratch,
- Word embedding that are learned using an unsupervised neural language model that evolves over time. It has been demonstrated that using an unsupervised neural language model to initialise word vectors improves performance.

We present Hierarchical Hybrid Networks, which are inspired by human clickbait detection methods and leverage the link between title and content, whereas most prior algorithms either fail to include both or simply take their similarity into consideration. However, Hierarchical Hybrid Networks, like all previous studies, only function in ideal conditions (data sharing). As a result, we suggest a unique training paradigm called Clickbait Federated Learning as a viable solution to the Data Island problem. For model training, this approach may successfully use data from two parties. Furthermore, Clickbait Federated Learning does not need agreement on network architecture between the two parties. We get Federated Hierarchical Hybrid Networks after training Hierarchical Hybrid Networks with Clickbait Federated Learning, which solves the Data Island problem.

1.3 Hierarchical Hybrid Networks:

Hierarchical Hybrid Networks are made up of four elements, as illustrated in Figure 3, a title feature extractor, a content feature extractor, a link extractor, and a classification network. Hierarchical Hybrid Networks theory is in an ideal scenario (data sharing)[7].

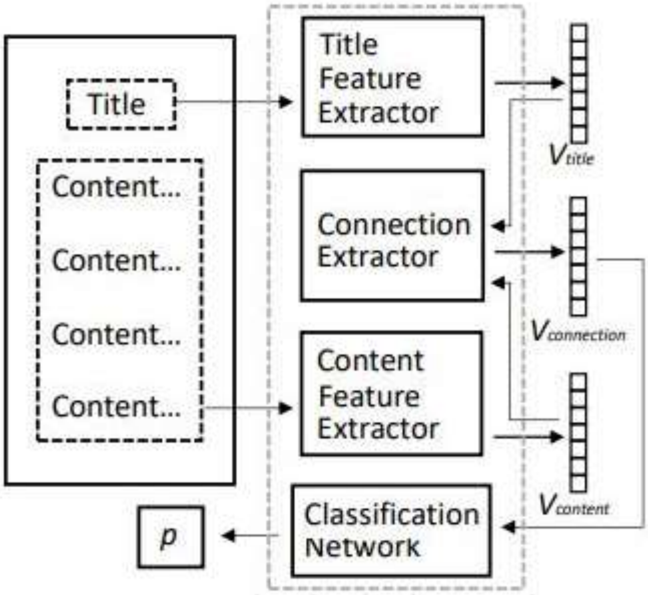


Figure 3: The Above Figure Shows the Hierarchical Hybrid Network [catalyzex].

1.3.1 Feature Extraction:

We need a feature extractor to extract features from the text for classification because clickbait detection is a text classification problem. Figure 4 shows the Feature extraction[8].

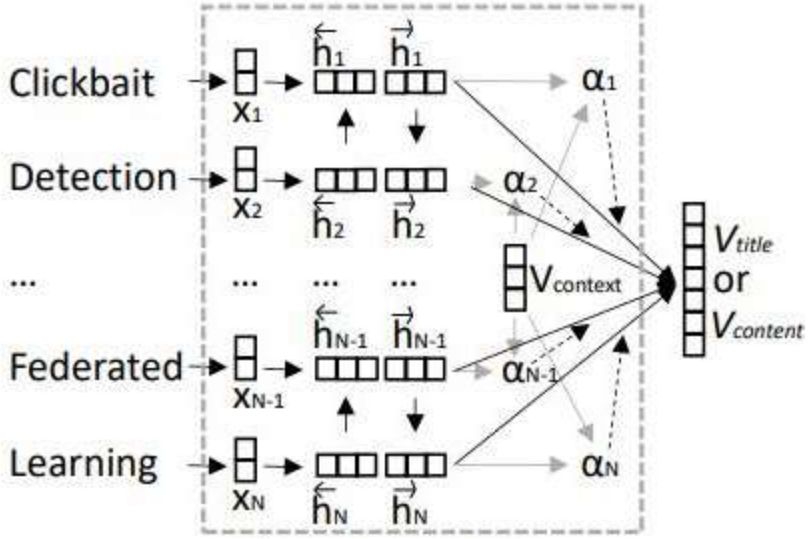

Clickbait
Detection
...
Federated
Learning
x_1
x_2
x_{N-1}
x_N
$\overleftarrow{h}_1$ $\overrightarrow{h}_1$
$\overleftarrow{h}_2$ $\overrightarrow{h}_2$
$\overleftarrow{h}_{N-1}$ $\overrightarrow{h}_{N-1}$
$\overleftarrow{h}_N$ $\overrightarrow{h}_N$
α_1
α_2
α_{N-1}
α_N
$V_{context}$
V_{title} or $V_{content}$

Figure 4: The above figure shows the Feature Extraction [catalyzex].

1.3.2 Connection Extractor:

The problem with clickbait arises when the article's substance fails to deliver on the promise given in the headline. As a result, when humans identify clickbait, they integrate the headline with the content. Human makes the final decision based on the intricate relationship between title and content[9]. We build a link extractor implemented by Convolutional Neural Network, as illustrated in Figure 3, based on human behaviours in clickbait detection.

We concatenate the vectors V_{title} and $V_{content}$ obtained from feature extractors to generate V. To obtain the connection characteristic, we use a convolution technique. The parameters of this filter are W_c and b_c, and the activation function is f. From the i_{th} row through the i+h-1_{st} row of V, a feature c_i is learnt. All features are concatenated in the feature map c. Equation is processed as:

$$V = V_{title} \,||\, V_{content}$$

$$c_i = f(\boldsymbol{W_c} V_{i:i+h-1} + \boldsymbol{b}_c)$$

$$c = [c_1, \dots, c_{n-h+1}]$$

1.3.3 Classification network:

M_4 is a fully connected neural network. W and b is the parameters of the classification network[10]. The process is given by the following equation:

$$p = softmax(\boldsymbol{W} V_{connection} + \boldsymbol{b})$$

2. DISCUSSION

The author has discussed about the clickbait detection, Clickbait's are internet articles with deceptive names that are designed to get people to visit the intended web page. To monetize the landing page or distribute fake news for sensationalize, clickbait's are used to entice users to click on a certain link. The author has also discussed about the Hierarchical Hybrid Networks taught using Clickbait Federated Learning are known as Federated Hierarchical Hybrid Networks. For identifying clickbait, Hierarchical Hybrid Networks use not only the characteristics of the title and content, but the intricate relationship between the title and content. In the ideal case, Clickbait Federated Learning can successfully use non-shared data in a Data Island environment to train a federated model that is similar to a model with the same architecture built using the standard training approach. As a result, it is a promising solution to the Data Island problem, and this approach may be applied to other multi-input classification problems involving non-shared data. Federated Hierarchical Hybrid Networks performed well on clickbait detection tasks, according to our findings. Hierarchical Hybrid Networks are made up of four elements, a title feature extractor, a content feature extractor, a link extractor, and a classification network. Hierarchical Hybrid Networks theory is in an ideal scenario. Hierarchical Hybrid Networks, which are inspired by human clickbait detection methods and leverage the link between title and content, whereas most prior algorithms either fail to include both or simply take their similarity into consideration. However, Hierarchical Hybrid Networks, like all previous studies, only function in ideal conditions (data sharing). As a result, we suggest a unique training paradigm called Clickbait Federated Learning as a viable solution to the Data Island problem. For model training, this approach may successfully use data from two parties. Furthermore, Clickbait Federated Learning does not need agreement on network architecture between the two parties. The main features of Hierarchical Hybrid Networks are feature extraction, classification abstraction, connection extraction.

3. CONCLUSION

The author has concluded about the clickbait detection and the objective of this study is to discover messages in a social stream that are meant to exploit cognitive biases to improve the chance of readers clicking an associated link. The practical success of clickbait, as well as the following rush of clickbait on social media, might turn it into another type of spam, cluttering up social networks and causing annoyance to users. The use of clickbait by news organisations is particularly concerning. Traditional clickbait-detection methods rely on heavy feature engineering and fail to distinguish clickbait from normal headlines precisely because of the limited information in headlines. A convolutional neural network is useful for clickbait detection, since it utilizes pertained Word2Vec to understand the headlines semantically, and employs different kernels to find various characteristics of the headlines. However, different types of articles tend to use different ways to draw users' attention, and a pertained Word2Vec model cannot distinguish these different ways. To address this issue, we propose a clickbait convolutional neural network (CBCNN) to consider not only the overall characteristics but also specific characteristics from different article types. The author has also concluded about the Hierarchical Hybrid Networks. Hierarchical Hybrid Networks use not only the characteristics of the title and content, but the intricate relationship between the title and content. In the ideal case, Clickbait Federated Learning can successfully use non-shared data in a Data Island environment to train a federated model that is similar to a model with the same architecture built using the standard training approach. Hierarchical Hybrid Networks theory is in an ideal scenario. Hierarchical Hybrid Networks, which are inspired by human clickbait detection methods and leverage the link between title and content, whereas most prior algorithms either fail to include both or simply take their similarity into consideration.

REFERENCES

[1] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9626, no. 1, pp. 810–817, 2016, doi: 10.1007/978-3-319-30671-1_72.

[2] A. Agrawal, "Clickbait detection using deep learning," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 268–272, 2017, doi: 10.1109/NGCT.2016.7877426.

[3] Y. Zhou, "Clickbait Detection in Tweets Using Self-attentive Network," 2017.

[4] H. T. Zheng, J. Y. Chen, X. Yao, A. K. Sangaiah, Y. Jiang, and C. Z. Zhao, "Clickbait convolutional neural network," *Symmetry (Basel).*, vol. 10, no. 5, pp. 1–12, 2018, doi: 10.3390/sym10050138.

[5] D. López-Sánchez, J. R. Herrero, A. G. Arrieta, and J. M. Corchado, "Hybridizing metric learning and case-based reasoning for adaptable clickbait detection," *Appl. Intell.*, 2018, doi: 10.1007/s10489-017-1109-7.

[6] A. Anand, T. Chakraborty, and N. Park, "We used neural networks to detect clickbaits: You won't believe what happened next!," 2017, doi: 10.1007/978-3-319-56608-5_46.

[7] J. Wang, H. Zhang, Z. Wang, and D. W. Gao, "Finite-Time Synchronization of Coupled Hierarchical Hybrid Neural Networks with Time-Varying Delays," *IEEE Trans. Cybern.*, 2017, doi: 10.1109/TCYB.2017.2688395.

[8] G. Hu *et al.*, "3D Graphene-Foam-Reduced-Graphene-Oxide Hybrid Nested Hierarchical Networks for High-Performance Li-S Batteries," *Adv. Mater.*, 2016, doi: 10.1002/adma.201504765.

[9] W. Hou, L. Guo, X. Wang, and X. Wei, "Joint port-cost and power-consumption savings in hybrid hierarchical optical networks," *Opt. Switch. Netw.*, 2011, doi: 10.1016/j.osn.2011.03.003.

[10] S. Arshad, B. Shahzaad, M. A. Azam, J. Loo, S. H. Ahmed, and S. Aslam, "Hierarchical and Flat-Based Hybrid Naming Scheme in Content-Centric Networks of Things," *IEEE Internet Things J.*, 2018, doi: 10.1109/JIOT.2018.2792016.