IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1241)
Available online at: https://www.ijariit.com

Online-Fraud Detection, Web-Scraped Dataset Analysis and Algorithm Performance Evaluation

Anam Khan
anam.mail4u@gmail.com
Thakur College of Engineering and Technology, Mumbai, Maharashtra

Dr. Megharani Patil
megharani.patil@thakureducation.org
Thakur College of Engineering and Technology, Mumbai, Maharashtra

ABSTRACT

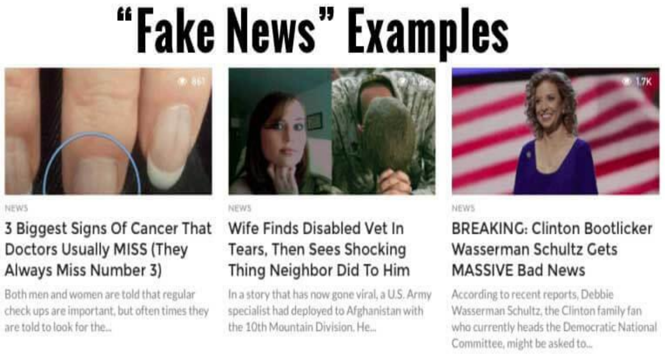
The volume of online information transfer has raised significantly in past few years. It is mainly due to various online portals that are currently one of the primary sources used by people. It is not always the case that the information found online is correct unless gained through some trusted and verified source. This is the reason why the user falls victim in the hands of falsie media. The term "Click-Bait" and "Fake News" are heard several times but people are unaware of the harm it is causing. research in this field is still active in trying to enhance the accuracy of these systems. In this paper, we will use supervised machine learning, research in this field is still active in trying to enhance the accuracy of these systems. In this paper, we will use supervised machine learning. research in this field is still active in trying to enhance the accuracy of these systems. In this paper, we will use supervised machine learning. , research in this field is still active in trying to enhance the accuracy of these systems. In this paper, research in this field is still active in trying to enhance the accuracy of these systems. In this paper, research in this field is still active in trying to enhance the accuracy of these systems. In this paper, the paper incorporates a Novel Neural Network Approach where Headlines are specifically taken into consideration. The model is tested for various types of headlines included in the Dataset. The proposed system is not only tested for Click-Bait Datasets collected from various news sites but also web-scraped data. Considerable results are obtained in terms of Accuracy, F1-score, Precision and Recall.

Keywords— *Click-bait, Fast Text, Natural Language Processing (NLP), Neural Network (NN)*

1. INTRODUCTION

Digital media empowers people and helps in social interaction, gives a platform where they can speak, participate in civic activities and also create different communities for their welfare. It has changed the way of working, boosted productivity, enhanced employers and workers flexibility. But at the same time, starting from Data Security, there have been several issues related to digital media that cannot be neglected, like crime and terrorism, complexity, privacy, social disconnect, work overload, and most destructive Digital Media Manipulation. For a country like India spreading of false information is disastrous. Digital Media Manipulators like "Click-bait" and "Fake News" have the power of creating havoc if caught the wrong attention. If you've spent some time on the web, you've likely seen articles and pictures with headlines just like the examples below. They're just a small sample of what's known as clickbait. Click-Bait may be a sensationalized headline that encourages you to click a link to a piece of writing, image, or video. Instead of presenting objective facts, clickbait headlines often appeal to your emotions and curiosity. Once you click, the web site hosting the link, earns revenue from advertisers, but the particular content is typically of questionable quality and accuracy. Websites use clickbait to attract as many clicks as possible, thus increasing their ad revenue.

While sensational headlines and content are used since the 19th century, they've become widespread within the digital world. Although it's based on an old idea, clickbait still serves the same purpose as its ancestor: to get your attention by whatever means necessary.



3 Biggest Signs Of Cancer That Doctors Usually MISS (They Always Miss Number 3)

Both men and women are told that regular check ups are important, but often times they are told to look for the...

Wife Finds Disabled Vet In Tears, Then Sees Shocking Thing Neighbor Did To Him

In a story that has now gone viral, a U.S. Army specialist had deployed to Afghanistan with the 10th Mountain Division. He...

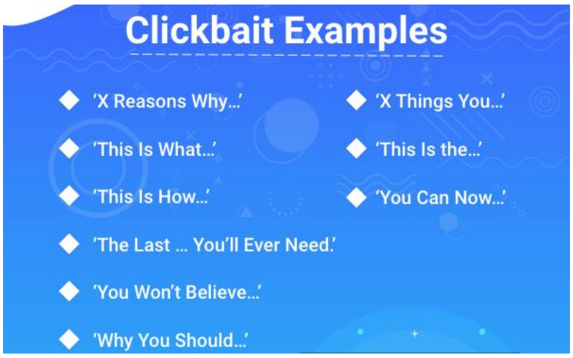
BREAKING: Clinton Bootlicker Wasserman Schultz Gets MASSIVE Bad News

According to recent reports, Debbie Wasserman Schultz, the Clinton family fan who currently heads the Democratic National Committee, might be asked to...

Figure 1: Overview of Clickbait Leading to Fake News

Some common words which form Click-Bait headlines are as follows:

© 2021, www.IJARIIT.com All Rights Reserved



Clickbait Examples

- 'X Reasons Why...'
- 'X Things You...'
- 'This Is What...'
- 'This Is the...'
- 'This Is How...'
- 'You Can Now...'
- 'The Last ... You'll Ever Need.'
- 'You Won't Believe...'
- 'Why You Should...'

Figure 2: Click-Bait Headlines Example

2. Literature Survey

It is not always true that adding more features will increase accuracy. Accuracy might decrease if a feature has a high correlation with another feature [1] or is derived from another feature, the. More data is required to ensure there are enough samples for each combination of values if more features are added. In order to capture the lexical differences between the two classes – clickbait and non-clickbait, Feature Extraction is also been performed on the basis of Sentence Composition, Word structure, Language Analysis, Lexical Nuances. The need to capture the semantic and [2] syntactic dependencies prompted to the use of distributed word embeddings in addition to the lexical features. GloVe and Word2Vec is used to draw a comparative analysis between the two to see which yields the best results. Word2Vec techniques comes into picture as It is used for mapping words in the sentences into multi-dimensional [3] vector space by sliding a window through the sentences and calculating the co-occurrence individually. GloVe being an algorithm for unsupervised learning is used for representing words as vectors. Using GloVe [4] 6B, the tokens were embedded into the 300-dimensional word vectors space. This model works by first pre -computing the co-occurrence matrix, followed by factorizing it to reduce the dimensionality of the matrix by looking at the context in which the given word appears in the corpus. Thus, it is a count-based model that works on the principal of dimensionality reduction and is easier to parallelize. Word N-grams, POS N-grams and Syntactic N-grams [5] were used as features. As the size of the dataset increase N-gram feature space grows linearly. Recurrent Neural Network is a class of artificial neural networks [6] which utilizes sequential information and maintains history through its intermediate layers. The output at each time-step is dependent on that of the previous time-steps of an internal state of , though has got certain drawbacks as Standard Recurrent Neural Networks have difficulty preserving long range dependencies due to the vanishing gradient problem [7]. This is related to interaction between words that are several steps apart. The LSTM using gating mechanism is able to diminish this problem. Also, Convolutional Neural Networks [8] have been utilized for various deep learning tasks. Mainly, a simple CNN having one layer of convolution has been used. MLP being the backbone [9] of Deep Learning is also been used several times for this purpose.

3. Data Collection And Visualization

52,000 headlines from clickbait and non-clickbait sources are taken and the year in which such clickbait headlines are formed is roughly around 2007 - 2020.

- 30,000: 2007-2016 headlines from [10] https://www.kaggle.com/amananandrai/clickbait-dataset
- 22,000: 2019 - 2020 headlines scraped/requested from Twitter, APIs, online publications.

Clickbait sources: Buzzfeed, Upworthy, Viral Nova, Bored Panda, Thatscoop, Viralstories, PoliticalInsider, Examiner, The Odyssey.

Non-clickbait sources: NY Times, The Washington Post, The Guardian, Bloomberg, The Hindu, WikiNews, Reuters.

Table 1: Statistics of Dataset used

Total Headlines	Click-bait Headlines	Non-Click-bait Headlines	Starting Year of Creation	Ending Year of Creation
52,000	27,000	25,000	2007	2020

The Dataset is then divided into 80:20 training and testing data respectively.

Table 2: Dataset Splitting for training and testing

Number of training data	41737
Number of testing data	10435

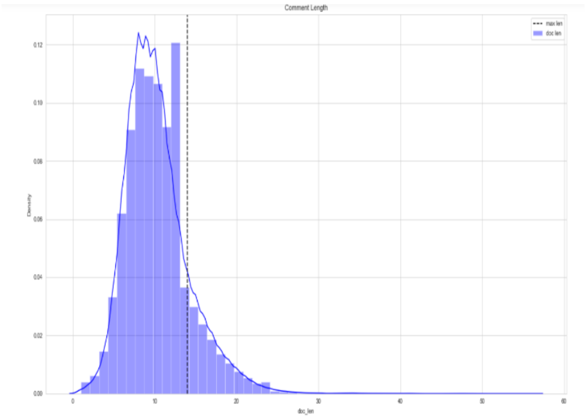


Figure 3: Distribution of headlines for training data

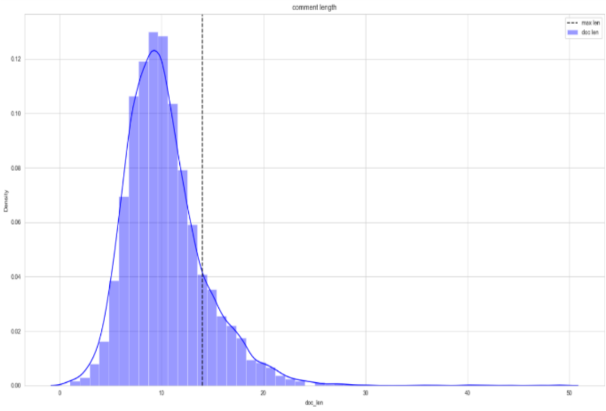


Figure 4: Distribution of headlines for testing data

In order to understand the distribution of headlines in training and testing datasets, an additional variable "document length" is used here which accounts for total words present in a headline.

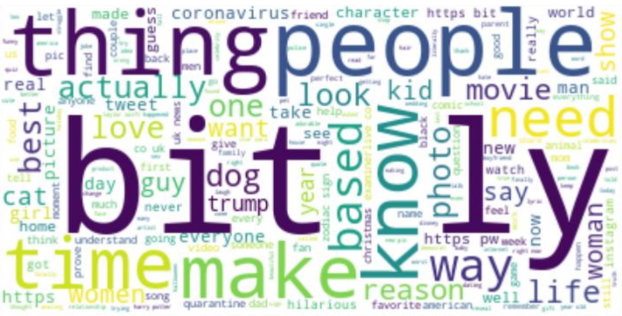


Figure 5: Word Cloud for Clickbait Headlines

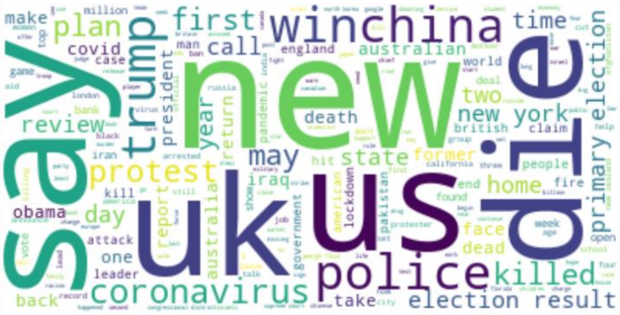


Figure 6: Word Cloud for NON-Clickbait Headlines

The similarities and dissimilarities can be clearly identified with the help of above visualization between these two. Furthermore, frequently occurring words have also been identified that are present in a clickbait headline.

The following figure shows the top words and their count.

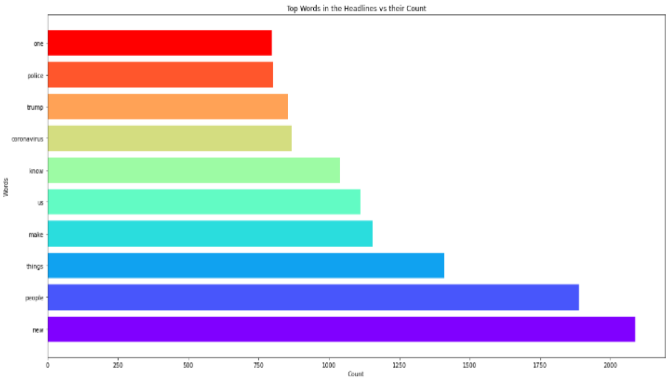


Figure 7: Words and their Count in Headlines

4. PROPOSED SYSTEM

The first phase consists of applying NLP on the entire dataset. The steps involved in the NLP technique are as follows:

- Stop-Word Addition and removal
- Tokenization.
- Lower Casing
- Removal of Extra White Spaces and Punctuation

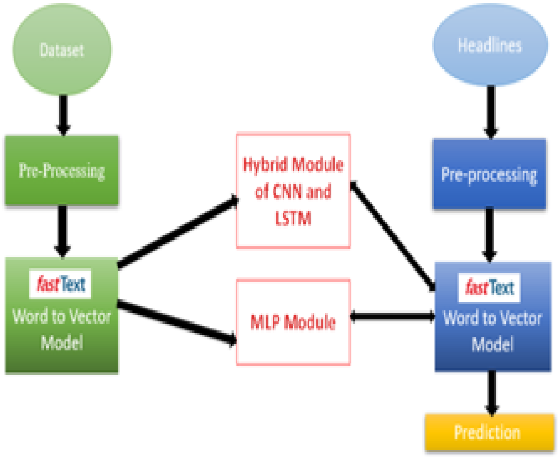


Figure 8: Architecture Diagram for the proposed model

The second phase uses FastText word embeddings that are obtained from official FastText Website https://fasttext.cc/docs/en/english-vectors.html for the dataset that will be fed to the neural network.

The third phase includes two robust models that are as follows:

A) Hybrid Bi-Directional LSTM-CNN model:

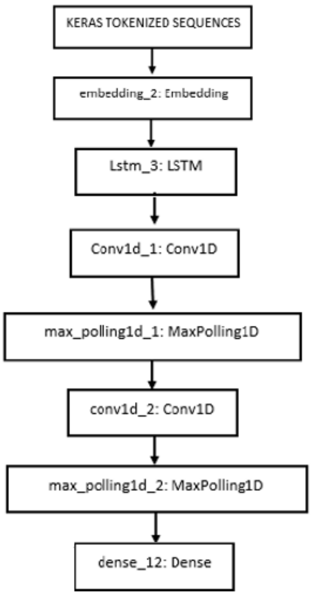

KERAS TOKENIZED SEQUENCES
embedding_2: Embedding
Lstm_3: LSTM
Conv1d_1: Conv1D
max_polling1d_1: MaxPolling1D
conv1d_2: Conv1D
max_polling1d_2: MaxPolling1D
dense_12: Dense

Figure 9: Bi- Directional LSTM-CNN Hybrid Model

The binary cross entropy is used here as loss function and Adam optimizer is used along with "sigmoid" activation function.

B) Multi- Layer Perceptron Model:

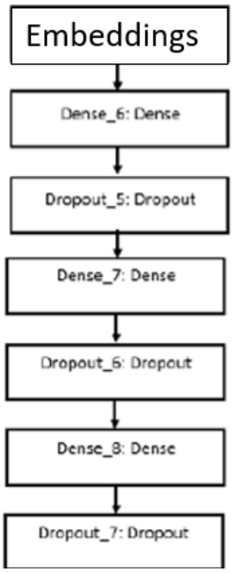


Figure 10: Architecture of MLP Model to achieve Best Accuracy

5. MODEL EVALUATION AND RESULT

The last and most crucial phase comprises of model evaluation and discovering really considerable results.

Table 3: For Bi-Directional LSTM-CNN

Total parameters	10,404,781
Trainable parameters	218,881

Table 4: For MLP

Total parameters	10,189,625
Trainable parameters	3,725

It can be clearly seen that for both models, count of Trainable parameters is much lower as compared to total parameters. It clearly states that these are optimal weights found by the model to reduce the cost function of the model.

© 2021, www.IJARIIT.com All Rights Reserved

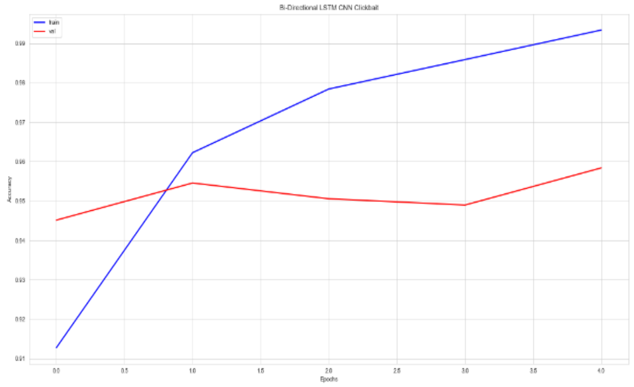


Figure 11: Accuracy Graph with FastText Word Embeddings on Bi-Directional LSTM-CNN model

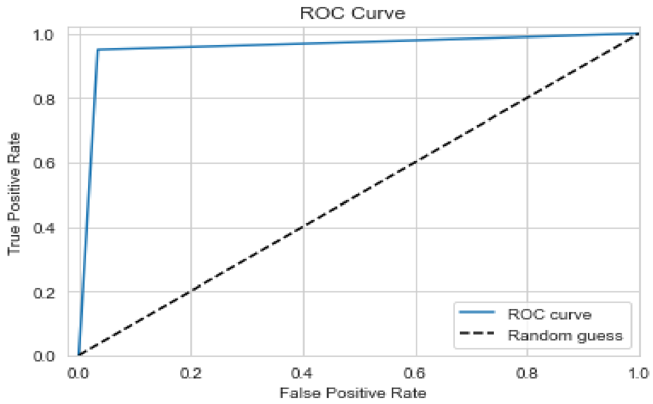


Figure 12: ROC Curve with FastText Word Embeddings on Bi-Directional LSTM-CNN model

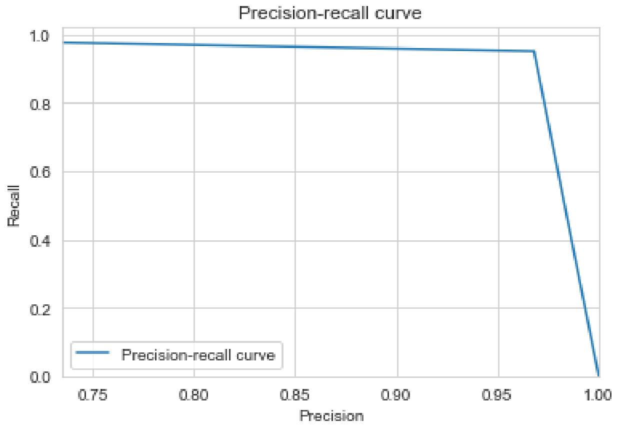


Figure 13: Precision-Recall curve with FastText Word Embeddings on Bi-Directional LSTM-CNN model

Table 5: Analysis of Bi-Directional LSTM-CNN model using performance metric

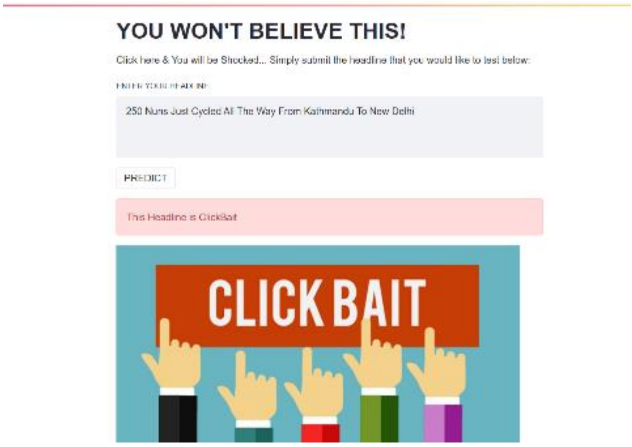
	Accuracy	Precision	Recall	F1-Score
Click-bait	96.00	95.00	97.00	96.00
Non-Click-bait		97.00	95.00	96.00

Table 6: Analysis of MLP model using performance metric

	Accuracy	Precision	Recall	F1-Score
Click-bait	74.00	73.00	75.00	74.00
Non-Click-bait		75.00	73.00	74.00

6. USER INTERFACE

The user interface for the model is created using Streamlit Framework.



YOU WON'T BELIEVE THIS!

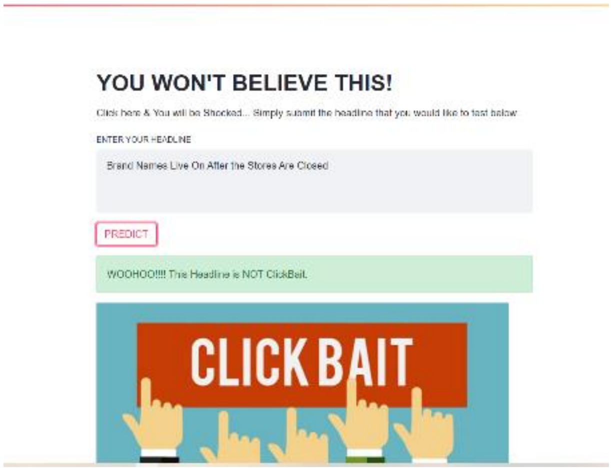
Click here & You will be Shocked... Simply submit the headline that you would like to test below:

ENTER YOUR HEADLINE

250 Nuns Just Cycled All The Way From Kathmandu To New Delhi

This Headline is ClickBait

Figure 14: Click-Bait Headline Detected on Streamlit UI



YOU WON'T BELIEVE THIS!

Click here & You will be Shocked... Simply submit the headline that you would like to test below:

ENTER YOUR HEADLINE

Brand Names Live On After the Stores Are Closed

WOOHOO!!!! The Headline is NOT ClickBait.

Figure 15: Non-ClickBait Headline Detected using Streamlit UI

7. CONCLUSION

In order to demote and block the usage of Click-Bait headlines by agencies in order to lure the audience towards fake and cheap content for earning revenue, novel neural network architectures are been proposed, and it has been found out that how the data is being fed to the neural network plays a major role in increasing the accuracy and reliability of the system. Also, the Bi-Directional LSTM-CNN model performs better than almost all the currently existing systems as it has 96% accuracy and MLP model has 74% accuracy on 52000 rows of data, with FastText word embeddings is performing better than any classification algorithm. It overcomes all the shortcomings of Word2Vec and GloVe embedding techniques. The Future Scope holds in softwares to not only detect but block the Click-Baited headlines at the same time.

8. ACKNOWLEDGMENT

I gratefully acknowledge the support, guidance and encouragement of my Dissertation Guide and Associate Professor Dr. Megharani Patil ma'am for this novel work.

9. REFERENCES

[1] Daoud M. Daoud, M. Samir Abou El-Seoud , "An Effective Approach for Clickbait Detection Based on Supervised Machine Learning Technique", International Journal of Online and Biomedical Engineering (iJOE),Volume 15, Issue 3, 2019.

[2] Saumya Pandey, Gagandeep Kaur, “Curious to Click It?-Identifying Clickbait using Deep Learning and Evolutionary Algorithm”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[3] Hai-Tao Zheng , Jin-Yuan Chen, Xin Yao , Arun Kumar Sangaiah, Yong Jiang, Cong-Zhi Zhao,“ Clickbait Convolutional Neural Network”, 2018.

[4] Peter Adelson , Sho Arora , and Jeff Hara, “Clickbait; Didn’t Read: Clickbait Detection using Parallel Neural Networks”, 2018.

[5] Philogene Kyle Dimpas, Royce Vincent Po, Mary Jane Sabellano “Filipino and English Clickbait Detection Using a Long Short-Term Memory Recurrent Neural Network”, 2017.

[6] Kai Shu, Suhang Wang, Thai Le, Dongwon Lee,Huan Li, “Deep Headline Generation for Clickbait Detection”, IEEE International Conference on Data Mining, 2018.

[7] Praphan Klairith, Sansiri Tanachutiwat , “Thai Clickbait Detection Algorithms using Natural Language Processing with Machine Learning Techniques”, International Conference on Engineering, Applied Sciences, and Technology (ICEAST), 2018.

[8] Amol Agrawal, “Clickbait Detection using Deep Learning”, 2nd International Conference on Next Generation Computing Technologies -NGCT, 2016.

[9] Lanyu Shang, Daniel Zhang, Michael Wang, Shuyue Lai, Dong Wang, “Towards Reliable Online Clickbait Video Detection: A Content-Agnostic Approach”, 2017 *IEEE*/*ACM* International *Conference* on Advances in Social, 2019.

[10] https://www.kaggle.com/amananandrai/clickbait-dataset.