IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

CLICKBAIT DETECTION USING MACHINE LEARNING ALGORITHM

Manasi Churi[1], Prof. Deepali Patil[2]

Computer Engineering
Shree.L.R.Tiwari College of Engineering
Mumbai, India

Abstract

In recent times, people seek information like news stories, blog posts, interviews, infographics, videos etc. from social media platforms due to its accessibility, low cost and fast dissemination. This has led to production of large amount of clickbaits, which are usually referred to as content written specifically to attract as many clicks as possible by creating social posts or sensational headlines. These clickbaits usually lure the readers to click by creating information gap in their minds and arousing their curiosities. The Oxford English Dictionary defines Clickbait as "(On the Internet) content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page". Clickbait is growing fast in volume and it can have many negative societal impacts. The aim of this project is to compare accuracy and efficiency of different machine learning methods used for clickbait detection.

Keywords: **SVM, Logistic regression**

I. INTRODUCTION

Nowadays the usage of social media sites, news sites are increasing rapidly due to its low cost and accessibility. This has led to sites creating creative ways to attract the readers to their website. One of the most widely used way is creating catchy headlines called clickbait. Such headlines are created to grab maximum readers and to earn quick profit by monetizing the pages through it. Clickbait can be defined as a piece of information or headline that deploys various linguistic nuances to create alluring, misleading, exaggerated, eye-catching, provocative with an element of suspense and tempting piece of text [1]. The main aim of clickbait headline is to create the "curiosity gap", withholding important information purposefully to make majority of people anxious to click on the clickbait news and usually these news does not deliver the content the reader is looking for. The example of such clickbaits include "21 Pictures That Are Way Too Real For People Who Grew Up With Strict Parents", "23 Couples Costumes That Will Give You Relationship Goals", "32 Tweets About Alcohol That'll Actually Make You Laugh", "Which Real Housewife Are You Based On Your Birth Month". In this project different classifiers are used to detect whether the news headline is clickbait or non-clickbait. And then performance of these methods are compared to find better solution.

II.LITERATURE REVIEW

Many researchers have been done classification of titles into clickbait and non-clickbait previously using various machine learning and deep learning approaches.

S. Pandey, G. Kaur[1] developed deep learning model that utilize the lexical as well as semantic features of the headlines and the corresponding text to detect clickbait. The author performed the experiment on 80,000 headlines

taken from different sites. Further, BiLSTM using GloVe embeddings and all the numerical features achieved an accuracy of 98.78%.

X. Cao et al. [10] identified clickbait using Random Forest Regression. Feature Engineering with the top ranked 60 features were selected in order to reduce the run time and it also reduces the noise interference. This Random Forest Regression and Feature Engineering achieved an accuracy of 0.819 on a dataset from the year 2017.

S. Manjesh et al. [2] considered features restricted to the headline and multiple machine learning models were used to classify headlines from websites as either clickbait or non-clickbait. Their paper found that deep learning approach fared the best with an average F1-score of 98%, closely followed by the multilayer perceptron neural network with 97%.

Sarjak Chawda et al. [5] used RCNN with GRU and LSTM for clickbait detection to capture long term dependencies. The RCNN+GRU achieved 0.9776 accuracy for clickbait detection.

P. Dimpas et al. [8] used neural network architecture based on Bidirectional Long Short Term Memory (BiLSTM) approach for clickbait detection also the model uses Word2Vec to provide word representation and embedding from the dataset which showed 91.5% accuracy.

A. Chakraborty et al. [7] proposed a SVM based approach for clickbait detection which resulted in an accuracy of 93% and 89% accuracy in blocking clickbaits. In their paper detailed linguistic analysis was done on the 15,000 headlines both in clickbait and non-clickbait categories using Stanford CoreNLP tool.

Praphan Klairith and Sansiri Tanachutiwat [4] has done research on clickbait detection for Thai headlines. Their paper found that deep learning method BiLSTM with word level embedding achieved accuracy rate of 0.98 and f1-score of 0.98.

J. Shin et al. [11] developed a system based on CNN for text classification which integrated feature extraction and content modulation. Their paper found that the performance of R-CNN was better than C-CNN.

A. Kuriakose et al. [6] developed a Webapp for clickbait and fake news detection. For this neural network model is developed which takes news titles and content as input and based on these inputs, the system predicts the percentage of news being clickbaiting.

N. Wongsap et al. [13] used different classifiers such as Support Vector Machine, Decision Tree and Naïve Bayes to compare the effect of special characters such as '!', '?', and '#' on the classification of Thai clickbait headlines. Their paper found that the special characters and decision tree classifier gives 99.90% accuracy.

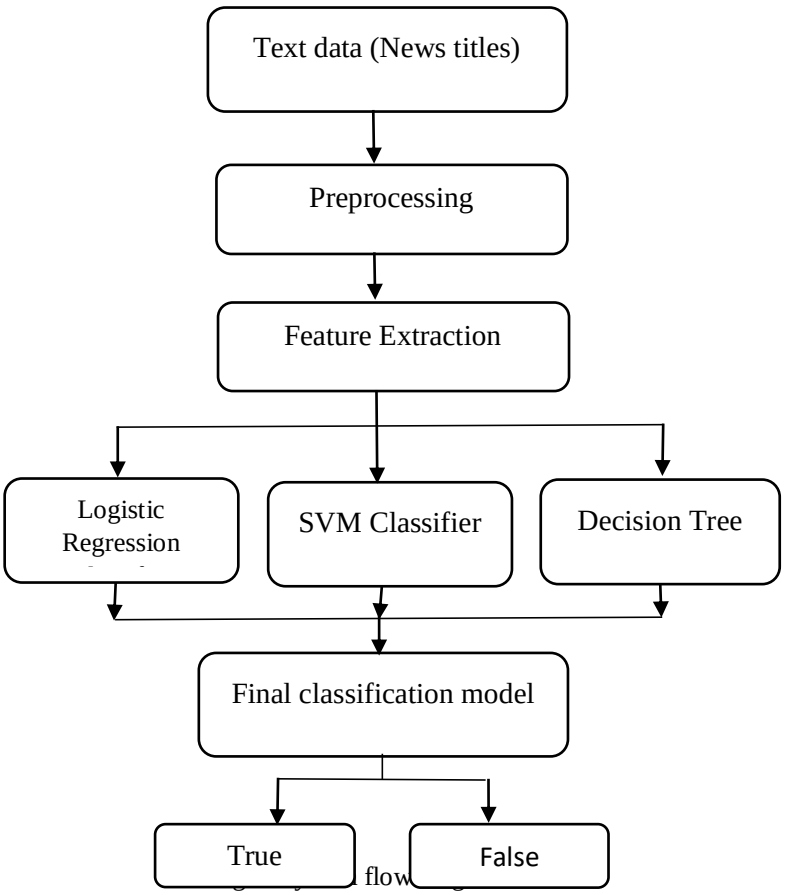
J. Fu et al. [9] proposed clickbait detection using CNN framework. Their empirical experiments showed that the model can better capture the local and global syntactic and semantic relations between words. For this model they used Chinese and English corpus of headlines.

III.PROPOSED SYSTEM:

The proposed system classifies the news titles or headlines as clickbait and non-clickbait. The various machine learning algorithms such as SVM, Logistic Regression, Random Forest are used for classification and to compare the accuracy.

a. **System flow:**

The Figure 1 shows the architecture of proposed system. The aim of the project is to develop one primary component which classify whether a headline is clickbait or non-clickbait. This component will be combined with other features and further machine learning models to work on data to create classification model based on specific attributes.



flow

Data set:

The dataset comprises of a total 32,000 headlines consisting of both clickbait and non-clickbait titles. The news headlines are collected from various popular websites and manually labelled as clickbait or non-clickbait. An equal number of clickbait and non-clickbait headlines were selected randomly in order to prevent a bias.

The following characteristics are chosen to find the lexical differences between the two classes- clickbait and non-clickbait:

Sentence composition: the question words, number of stop words, length of titles

Word structure: punctuation pattern, use of numbers in the title, adverb

SVM: It is a supervised learning algorithm that is used for efficient classification. The Fig. 2 shows the working of SVM i.e. given a labelled dataset SVM defines an optimal hyperplane that separates the dataset into two classes based on the lexical features defined above.

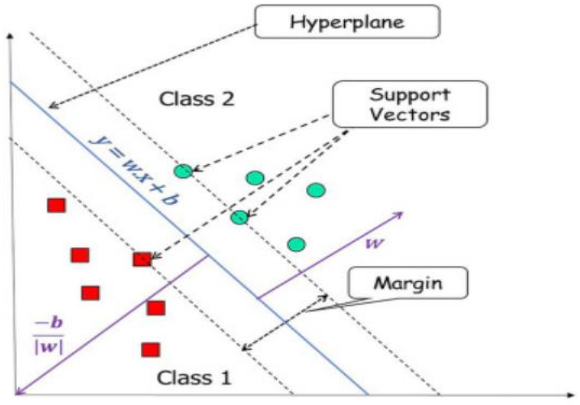


Fig.2 Working of SVM

Logistic Regression: It is a supervised learning algorithm widely used for binary classification problems. It uses set of independent variables to predict the categorical dependent variable. By using the logistic function given in equation 1 it predicts the probability of the given class by taking linear composition of features and weights and applies transformation and it produces the probabilistic values which lie between 0 and 1.

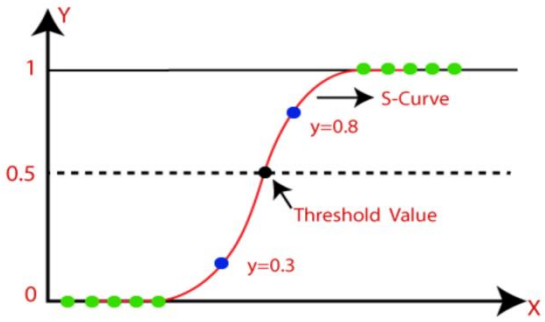


Fig.3 Working of Logistic Regression

Eq. 1 shows the logistic function which maps any real numerical value 'x' to a value between 0 and 1.

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$……Eq. 1

Decision Tree: It is a supervised algorithm preferably used to solve classification problems. It is a tree-structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In this algorithm the decisions or the test are performed on the basis of features of the given dataset.

IV.RESULT:

Result can be discussed on two levels namely pre-processing and feature extraction and accuracy of different algorithms which are applied on it. NLTK library is used for feature extraction. Sentence composition and word structure is considered for both clickbait and non-clickbait headlines. According to that, number of stopwords, length of headline, use of numbers are found comparatively more in clickbait headlines.

The Table I highlights the statistics of the evaluation metrics for the baseline models in which Logistic Regression performs well as compare to other models with a precision and recall of 0.97. This baseline approach only considers the headline and the lexical features.

Table I Result analysis

Model	Precision	Recall	F1-score
SVM	**0.93**	**0.93**	**0.93**
Logistic Regression	**0.97**	**0.97**	**0.97**
Decision Tree	**0.87**	**0.88**	**0.88**

V.CONCLUSION:

In this paper, the classifier is able to classify the headlines whether it belongs to clickbait or non-clickbait category. For classification of headlines linguistic differences such as sentence compositions, word structure, language analysis and lexical nuances are considered. Different machine learning algorithms are used to compare results, and according to the results logistic regression gives better accuracy.

REFERENCES

[1] 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.

[2] Suraj Manjesh, Tushar Kanakgiri, Vishak P, Vivek P, Vivek Chettiar, Shobha G, "Clickbait Pattern Detection and Classification of News Headlines using Natural Langugae Processing", 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) 2017

[3] Ayse Geckil, Ahmet Anil Mungen, Esra Gundogan, Mehmat Kaya, "A Clickbait Detection Method on News Sites", in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018.

[4]Praphan Klairith and Sansiri Tanachutiwat, "Thai Clickbait Detection Algorithms using Natural Language Processing with Machine Learning Techniques", in International Conference on Engineering, Applied Sciences, and Techmology (ICEAST), 2018.

[5] Sarjak Chawda, Aditi Patil, Abhishek Singh, Prof.Aditi Save, "A Novel Approach for Clickbait Detection", IEEE Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI), 2019.

[6] Ammu Kuriakose, Dinnu Sebastian, Esther Mahima Mathew, Hannu Mathew, Er. Gokulnath G, "ALIKAH – A Clickbait and Fake News Detection System using Natural Language Processing", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI), 2019.

[7]Abhijan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, Niloy Ganguly, "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media", in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[8]Philogene Kyle Dimpas, Royce Vincent Po, Mary Jane Sabellano, "Filipino and English Clickbait Detection Using a Long Short Term Memeory Recurrent Neural Network", in International Conference on Asian Language Processing (IALP), 2017.

[9]Junfeng Fu, Liang Liang, Xin Zhou, Jinkun Zheng, " A Convolutional Neural Network for Clickbait Detection", in 4th International Conference on Information Science and Control Engineering (ICISCE), 2017.

[10]Xinyue Cao, Thai Le, Jason (Jiasheng) Zhang, Donwong Lee, "Machine Learning Based Detection of Clickbait Posts in Social Media", 2017.

[11]J. Shin, Y. Kim, S. Yoon and K. Jung, "Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification", IEEE International Conference on Big Data and Smart Computing (BigComp) 2018

[12] Oxford Dictionaries. Clickbait [Online].

https://en.oxforddictionaries.com/definition/clickbait

[13]N. Wongsap, L. Lou, S. Jumun T. Prapphan, S. Kongyoung, N. Kaothanthong, "Thai Clickbait Headline News Clasification and its Characteristic".