An Impose of Dense Neural Network for Detecting Clickbait on Nepali News

Shiva Ram Dam[1,*], Saroj Giri[2], Tara Bahadur Thapa [3] and Sanjeeb Prasad Panday [4]
[1,2]*Department of Information Technology, Gandaki University, Nepal*
[3]*Department of Information System Engineering, GCES, Pokhara University, Nepal*
[4]*Department of Electronics and Computer Engineering, IOE, Tribhuwan University, Nepal*
Corresponding author: shivaram.dam@gandakiuniversity.edu.np

Abstract

Purpose: This research aims to detect clickbaits on Nepali news. Clickbaits are frequently existing in online Nepali digital media. Media house put catchy headlines which, in most of the cases, appears significantly different from the actual content inside it. They embellish the truth to entice readers to click on it.
Methods: A Machine learning model with Dense Neural Network (DNN) is imposed to train and test on the Nepali clickbait dataset. The model takes a featured dataset with cosine similarity and Term Frequency Inverse Document Frequency (TFIDF) to detect clickbaits and non-clickbaits.
Results: Our model achieved a high performance, evidenced by an F1 score of 96.27 on the test data with cross validation, demonstrating its effectiveness in distinguishing between clickbait and non-clickbait content.
Conclusin: Our study presents a successful application fo dense neural networks for clickbait detection in Nepali news, offering a valuable tool for improving news consumption quality. Future works will explore expanding the dataset and incorporating more advanced neural networks.

Keywords: Clickbait, Cosine similarity, DNN, TFIDF

1 Introduction

In the course of advancement of the Internet, web technologies are also advanced day by day (Bhadani & Jothimani, 2017). Transformation of print media to digital media did not remain untouched with this revolution. Online media such as websites, blogs and social media provide a quick delivery of news and information to people who desire to access through digital media (Zuhroh & Rakhmawati, 2019). The ease of low-cost access and low-cost publication makes online news portal take its place over the printed news (Kalombe & Phiri, 2019).

Due to competitive environment in the media houses, clickbaits are used to attract a greater number of readers and increase the rank of the website (Kaushal & Vemuri, 2021). We can find several clickbaits in Nepali news portals, blogs and social sites. Clickbaits are "articles on the Internet with content whose primary purpose is to attract attention and encourage visitors to click on a link to a particular webpage" (Zimdars, 2016). "A clickbait is usually a headline designed to make readers want to click on hyperlinks especially when the links lead to the content of dubious value or interest" (Merriam-Webster, n.d.). But these clickbaits dissatisfy the users if contents, inside it, are beyond their expectation. This annoys users and waste their time. As a result, identifying clickbait headlines ahead of time would better assist readers in deciding which news items to consume. This helps to reduce the chance of going through unwanted information and save time (Jung et al., 2022).

Applying natural language processing (NLP) to Nepali language is quite difficult. Nepali language is very rich grammar and exceptional cases. Finding the base form of Nepali words is bit challenging. This makes complexity in measuring the similarity between news pairs. This paper aims to detect such misleading news headline that is in contrast to its corresponding news body. We propose a dense neural network that learns the textual relationship between the news headline and news body. Our models were trained and tested on Nepali news dataset by Dam et al. (2021) which performed fairly well.

Related works

There has been a lot of research works carried out for detecting clickbaits (Zuhroh & Rakhmawati, 2019). Some researchers conducted their research just on news headline dataset (Manjesh et al., 2017), while some other took pairs of news headline and news body dataset (Yoon et al., 2018). Different methods of machine learning such as Logistic Regression, Naive Bayes, and Random Forest were used by Potthast et al. (2016) with a special emphasis on Twitter network. Chakraborty et al. (2016), used characteristics of sentence form, n-grams, Part of Speech (POS) and special words to identify clickbaits. Framework using Natural Language Processing (NLP) based on semantics and syntactics could be used for identification and classification of news stories as clickbait or non-clickbait (Manjesh et al., 2017).

Cao et al. (2017) used several clickbait features to perform detection of clickbait posts on social media using Random Forest (RF) classifier. Adelson et al. (2018) included TFIDF scores and pre-trained Gobal Vectors (GloVe) for word embedding and to extract baseline features. A Clickbait Convolutional Neural Network (CBCNN) model was utilized as word-embedding structure (Zheng et al., 2018). This structure took account of type-related word sense. Deep learning architectures: DNN, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) was applied to solve the problem of detecting fake news (Thota et al., 2018). Unlike other previous work of binary classification, four stances: Agree, Discuss, Disagree and Unrelated for clickbait detection were introduced. Deep hierarchical models were proposed to detect incongruity in news-headline and news-body (Yoon et al., 2018).

Classical machine learning approaches were used by Dam et al. (2021) to identify clickbaits on Nepali news using cosine-similarity and TFIDF.

2 Methodology

System model, demonstrated in figure 1, shows different processes that consist of many closely related activities. The first phase involved the construction of dataset. The next phases involved data preprocessing and extraction of features from the dataset. The featured dataset was then split into training and testing dataset. A 10-fold cross-fold validation technique was applied to the training dataset to train the system model (Berrar, 2019). Finally, the testing dataset was used to test and evaluate our model. The obtained result evaluates the performance of the model.

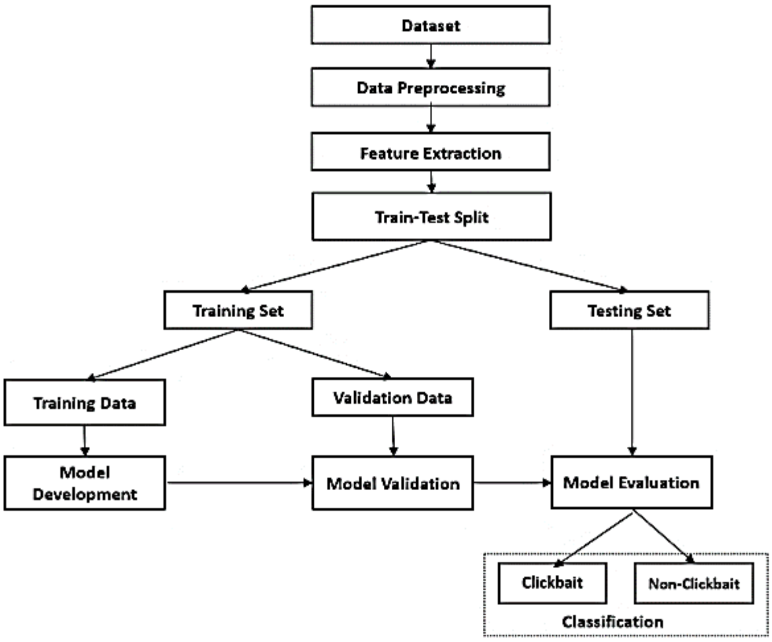


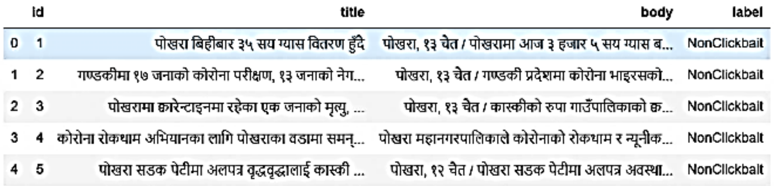
Figure 1: Workflow of clickbait detection process

2.1 Dataset

This research work used the dataset collected manually from various Nepali news portals by Dam et al (2021). The dataset consisted of 10K pairs of news pairs of title and body where 4000 were labeled as “Clickbaits” and 6000 labeled as “Non-clickbaits”.

2.2 System Architecture

To build automated machine learning model on text data, textual data need to be cleaned and converted into machine readable format. For the very purpose, NLP was applied to clean textual data by removing punctuation, English characters and digits, Nepali digits, stopwords and stemming them. In order to perform analysis on text, raw texts were converted into numeric form. TFIDF was used for word to vector representation. These TFIDF vectors of news headline-body pair and cosine similarity were input to our model which predicts the output and classifies as Clickbait or Non-clickbait. Figure 3 shows our high-level system architecture for clickbait detection.



	id	title	body	label
0	1	पोखरा बिहीबार ३५ सय ग्यास वितरण हुँदै	पोखरा, १३ चैत / पोखरामा आज ३ हजार ५ सय ग्यास ब...	NonClickbait
1	2	गण्डकीमा १७ जनाको कोरोना परीक्षण, १३ जनाको नेग...	पोखरा, १३ चैत / गण्डकी प्रदेशमा कोरोना भाइरसको...	NonClickbait
2	3	पोखरामा क्वारेन्टाइनमा रहेका एक जनाको मृत्यु, ...	पोखरा, १३ चैत / कास्कीको रुपा गाउँपालिकाको क्व...	NonClickbait
3	4	कोरोना रोकथाम अभियानका लागि पोखराका वडामा समन्...	पोखरा महानगरपालिकाले कोरोनाको रोकथाम र न्यूनीक...	NonClickbait
4	5	पोखरा सडक पेटीमा अलपत्र वृद्धवृद्धालाई कास्की ...	पोखरा, १२ चैत / पोखरा सडक पेटीमा अलपत्र अवस्था...	NonClickbait

Figure 2: Clikcbait-NonClickbait Dataset and its attributes

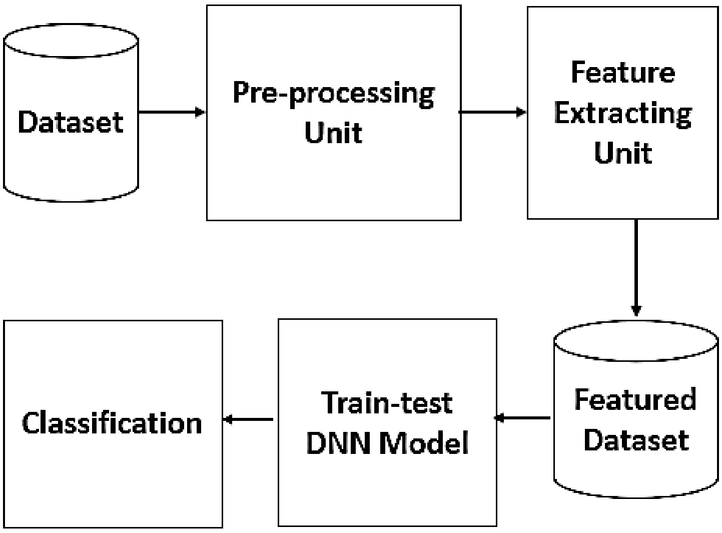


Figure 3: High-level system architecture

2.3 Data Preprocessing

Data preprocessing is an essential step in building a machine learning model. It is an initial task that is performed to clean the data for better accuracy and better performance. The results are seen depending on how well the data have been preprocessed. Data preprocessing was applied to both news headlines and their news body (Yoon et al., 2018). Preprocessing of the dataset involved the phases as shown in figure 4. Various punctuations like comma, full stops, apostrophes, hyphen, single quotes, double quotes, question mark, exclamation sign, asterisk, parenthesis, braces and many other special characters were removed and replaced with a blank (or white) space. Since English characters and words have very less significance in this research, they were removed out. All Nepali digits were replaced with their corresponding word form. The words were split and then stemmed by removing prefixes and suffixes to get their root form (Bal & Shrestha, 2004). Finally, stopwords listed in the Natural Language Toolkit (NLTK) were removed out from the text.

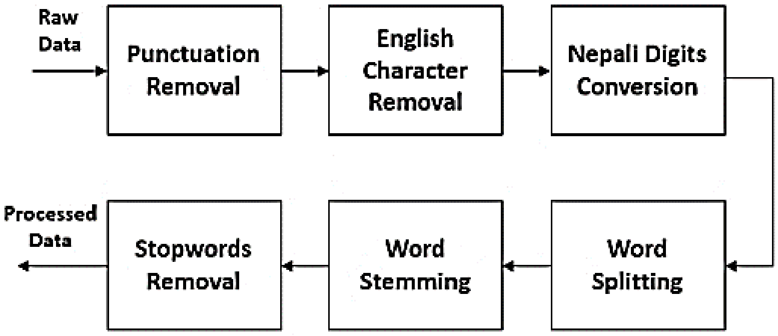


Figure 4: Phases in data preprocessing

2.4 Word to Vector Representation

Words needs to be converted into numeric form so that it can be implemented for Machine Learning (ML). Here, TFIDF vectorizer were utilized to express the words in the form of vectors. This TFIDF transforms text to a meaningful numerical representation. TFIDF is a metric that describes the value of a word to a document compared to the whole vocabulary (Chowdhury, 2010). Mathematically:

$$\text{TFIDF}(w) = \text{TF}(w) \times \text{IDF}(w) \tag{1}$$

Where, TF is Term Frequency and IDF is Inverse Document Frequency.
Cosine similarity is another metric used to create featured dataset.
Cosine similarity $\cos(\theta)$ is expressed as:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{2}$$

where, A is vector of TFIDF of news-title and B is vector of TFIDF of news-body.
It computes cosine angle between two vectors (Schutze et al., 2008). Two exactly same documents have a value of 1 and two entirely different documents have a value of θ. Other in-between values show intermediate similarity (Singhal, 2001).

2.5 Algorithms

The system model imposes Dense Neural Network (DNN), illustrated in figure 5, which is interconnected with fully linked neurons in a network layer. Each neuron in every layer received an input from every neuron in preceding layer and provided learning features from previous layer. The input layer received TFIDF vector of news title, TFIDF vector of news body and cosine similarity of each pair. The output layer produced 0 or 1 (i.e. clickbaits or non-clickbaits). The first three layers were provided with Rectified Linear Unit (ReLU) activation function and the output layer with Softmax. The model used Adaptive Moment Estimation (Adam) as optimizer and Binary Cross-entropy (BC) as loss function. Table 1 shows the detail of DNN model.

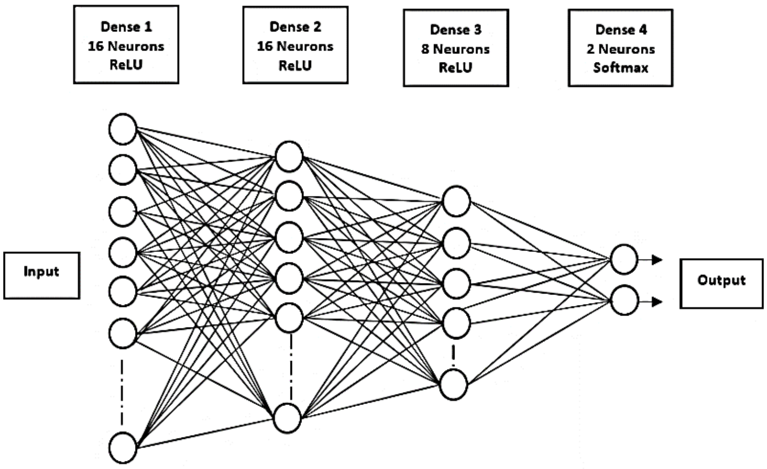


Figure 5: Layers and activation function in our dense neural network

3 Results

Figures 6 shows the accuracy plot for cross-validation with DNN model. Figure 7 show the confusion matrix of DNN model with testing dataset. Tables 2 and 3 show the accuracy, precision, recall and F1-score at 70:30 and 80:20 dataset split into training and testing dataset for SVM, RF and DNN model.

Table 1: Detailed components and parameters in our dense neural network

Hyperparameters	Value
No. of layers 4	4
Input layer activation function	ReLU
Hidden layer activation function	ReLU
Output layer activation function	Softmax
Optimizer	Adam
Loss function	BC
No. of epochs	25
Batch size	50

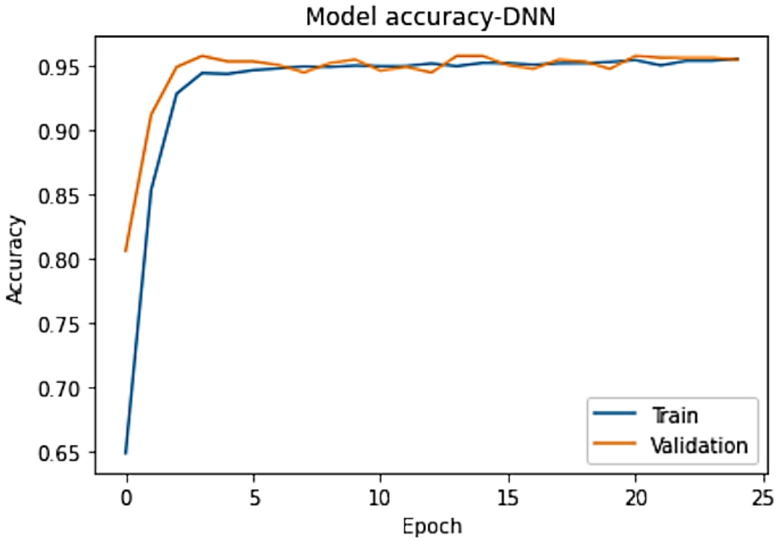


Figure 6: Accuracy plot with DNN

Table 2: Performance of models on test dataset at 70:30 data-split

Model	Accuracy	Precision	Recall	F1-Score
DNN	95.30	95.33	94.93	95.12
SVM	95.03	95.16	94.56	94.83
RF	94.93	94.92	94.59	94.73

Table 3: Performance of models on test dataset at 80:20 data-split

Model	Accuracy	Precision	Recall	F1-Score
DNN	96.40	96.46	96.11	96.27
SVM	95.15	95.22	94.76	94.97
RF	94.35	94.35	94.35	94.33

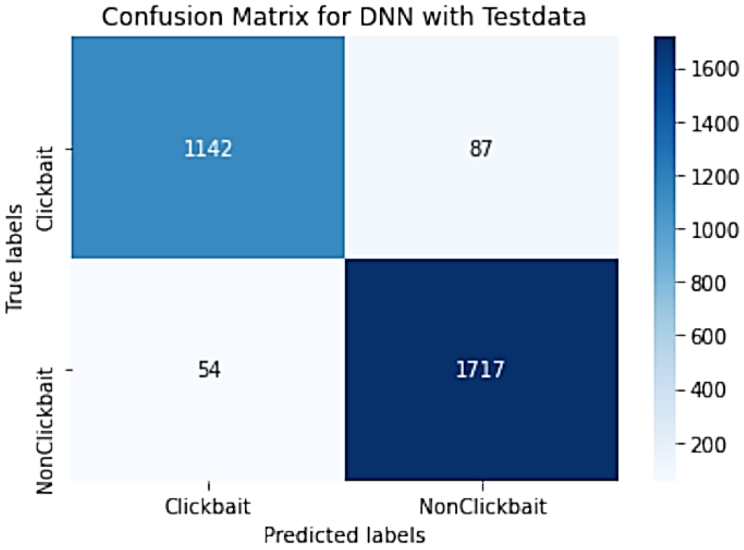


Figure 7: Confusion matrix for Test dataset with DNN

4 Discussion

Two experiments were conducted in this research. The first experiment was to compare the proposed model with the results obtained with other classical methods such as SVM and RF Classifier by Dam et al (2021). A 10-fold cross validation technique was used to validate our dataset (Berrar, 2019). The second experiment was to find the best performance of our model as compared to baseline model by Thota et al. (2018).

4.1 Comparison with classical model

The results show that DNN model achieved an F1-score of 96.27 on testing dataset at 80:20 train-test split. Figure 8 shows that DNN performed better than the classical model (Dam et al., 2021) in terms of the metrics: accuracy, precision, recall and F1-score.

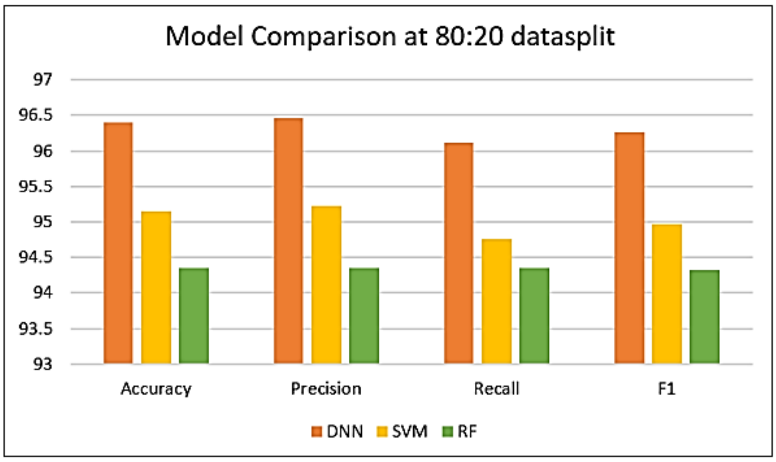


Figure 8: Model comparison between SVM, RF and DNN at 80:20 data-split

4.2 Comparison with baseline model

Table 4 shows the comparison with baseline model of Thota et al. (2018). Their dataset was split into train and test at 67:33 ratio. Train dataset was further split into train and validation at 80:20. A 3-fold cross-validation was applied for all experiments by Thota et al. (2018). Their methodology used DNN fed with TFIDF and cosine similarity into the model. Compared to result obtained by Thota et al. (2018), our model achieved better accuracy with a lead of 2.19 %. However, as shown in figure 7, few misclassifications were observed. This was because of TFIDF vectors not being able to recognize ambiguous words that have several meanings.

Table 4: Baseline performance of proposed model

Models	Accuracy
Thota-TFIDF-DNN	94.21
DNN	96.40

5 Conclusion

Nepalese media are gradually using clickbaits and this is an increasing phenomenon. News reader are getting trouble with this trend. They are not getting the actual content as in the headline, and hence wasting their time. A machine learning model with DNN has been imposed to classify clickbaits and non-clickbaits. This performed more better than the SVM and RF classifiers (Dam et al., 2021). The DNN model detects clickbaits with 96.4% accuracy in Nepali news. DNN performed slightly better than them.
This research can be extended to develop and improve prediction model by considering morphotactic and pragmatic analysis for Nepali language. A general way is to use parts-of speech tagging.

Acknowledgement

The authors are grateful to all those who had supported and guided throughout the course of the research.

Authors contribution:

Shiva Ram Dam conceptualized the study, created the dataset, designed the model and experiments. Saroj Giri developed the methodology and was involved in result analysis. This research was supervised by Sanjeeb Prasad Panday and Tara Bahadur Thapa.

References

Adelson, P., Arora, S., & Hara, J. (2018). Clickbait; didn't read: Clickbait detection using parallel neural networks. Retrieved January 9, 2024, from https://cs229.stanford.edu/proj2017/final-reports/5231575.pdf

Bal, B. K., & Shrestha, P. (2004). A morphological analyzer and a stemmer for Nepali. PAN Localization, Working Papers, 2007, 324-331.

Berrar, D. (2019). Cross-Validation. In Elsevier eBooks (pp. 542–545). https://doi.org/10.1016/b978-0-12-809633-8.20349-x

Bhadani, A., & Jothimani, D. (2017). Big Data: Challenges, Opportunities and Realities. arXiv. Retrieved from https://doi.org/10.48550/arXiv.1705.04928

Cao, X., Le, T., & Zhang, J. (2017). Machine Learning Based Detection of Clickbait Posts in Social Media. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1710.01977

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop Clickbait: Detecting and preventing clickbaits in online news media. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). https://doi.org/10.1109/asonam.2016.7752207

Chowdhury, G. G. (2010). Introduction to modern information retrieval. Facet Publishing.

Dam, S., Panday, S., & Thapa, T. (2021). Detecting Clickbaits on Nepali News using SVM and RF. Retrieved February 29, 2024, from http://conference.ioe.edu.np/publications/ioegc9/ioegc-9-018-90032.pdf

Jung, A., Stieglitz, S., Kissmer, T., Mirbabaie, M., & Kroll, T. (2022). Click me. . .! The influence of clickbait on user engagement in social media and the role of digital nudging. PloS One, 17(6), e0266743. https://doi.org/10.1371/journal.pone.0266743

Kalombe, C., & Phiri, J. (2019). Impact of online media on print media in developing countries. Open Journal of Business and Management, 07(04), 1983–1998. https://doi.org/10.4236/ojbm.2019.74136

Kaushal, V., & Vemuri, K. (2021). Clickbait—Trust and credibility of digital news. IEEE Transactions on Technology and Society, 2(3), 146–154. https://doi.org/10.1109/tts.2021.3073464

Manjesh, S., Kanakagiri, T., Vaishak, P., Chettiar, V., & Shobha, G. (2017). Clickbait pattern detection and classification of news headlines using natural language processing. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (pp. 1-5). Bengaluru, India. doi: 10.1109/CSITSS.2017.8447715.

Merriam-Webster. (n.d.). Clickbait. In Merriam-Webster.com dictionary. Retrieved January 9, 2024, from https://www.merriam-webster.com/dictionary/clickbait

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait Detection. In Lecture notes in computer science (pp. 810–817). https://doi.org/10.1007/978-3-319-30671-1_72

Schutze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39). Cambridge, UK: Cambridge University Press. Retrieved February 29, 2024, from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. SMU Data Science Review, 1(3), Article 10. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/10

Yoon, S., Park, K., Shin, J., Lim, H., Won, S., Cha, M., & Jung, K. (2018). Detecting Incongruity Between News Headline and Body Text via a Deep Hierarchical Encoder. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.07066

Zheng, H. T., Chen, J. Y., Yao, X., Sangaiah, A. K., Jiang, Y., & Zhao, C. Z. (2018). Clickbait Convolutional Neural Network. Symmetry, 10(5), 138. https://doi.org/10.3390/sym10050138

Zimdars, M. (2016). False, misleading, clickbait-y, and satirical ”news” sources. Google Docs. Retrieved from https://docs.google.com/document/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/preview

Zuhroh, N. A., & Rakhmawati, N. A. (2019). Clickbait detection: A literature review of the methods used. Register, 6(1), 1. https://doi.org/10.26594/register.v6i1.1561

Correct citation: Dam, S. R., Giri, S., Thapa, T. B., & Panday, S. P. (2024). An Impose of Dense Neural Network for Detecting Clickbait on Nepali News. *Jagriti-An Official Journal of Gandaki University, 1*(1), 58-65.