International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 02, 2020
ISSN: 1475-7192

Clickbait Identification in Social Media Text using LSTM based Approach

*[1]Lakshmi Dheeraj, [2]Dr.Nelson

Abstract

Capsule networks can also be used to investigate the working under the input of text data rather than images. Such limitations where taken into account while identifying clickbaits in the usage of texts from social media. Several systems in existing work focus on usage of Long Term Short Memory (LSTM) approach to identify clickbaits. Due to certain disadvantages, the issue was considered as research focus area. The proposed system utilizes three-layered architecture where the first layer takes the input as text. After vectorization, the input is categorized and the final layer produces the output. The layered architecture also takes lesser response time thereby making the system efficient which is showcased in the experimental results obtained after implementation.

Keywords*: Clickbaits, Long Term Short Memory, Social Media Text, Vectorization*

I. Introduction

Social attention are gained by using social networks that stand as an important component in our day-to-day life. Exchange of information also can be increased by the factor of improvising the social networks both in qualitative and quantitative manner. Though several advantages are there, certain problems do exist making the research area to be focused. One of the major issue to be focused is clickbaits. When an user clicks any webpage contents, he/she is able to make money which enables users to design their web pages to attract customers. Idea suggested may be to have headlines in catch manner or the link of the information that makes the customers to click them. Hence identifying clickbait is considered as a categorization problem.

Several existing systems in text categorization makes use of deep learning and machine learning algorithms along with recurrent neural networks named as RNN or Long short term memory named as LSTM. Also certain features are additionally appended in the models based on that the performance is improvised. Patterns as well as sequences are taken into account for LSTM based approach. Confidence parameter is calculated by using the error produced when text input is used. LSTM based approaches suffer from certain limitations including the network issues also and also spatial properties of text not considered here.

[1] UG Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India.
[2] Professor, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India.

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

In order to consider the text spatial properties, our proposed system helps in overcoming the issues related to it. The contributions of the paper include,

- Analyzing the application of the proposed system in the environment where the amount of text is less.
- To identify clickbaits, an automatic method is proposed that prevents the usage of manual entries.
- Compared to deep learning and machine learning algorithms, our proposed system provides better results.

The organization of the paper is as follows: Section 2 discusses about literature related works, Section 3 presents the proposed system. Section 4 discusses the results and Section 5 concludes the paper.

II. Literature Review

Existing works on Clickbaits concentrated more on handcrafted characteristics in order to clearly know the patterns as well as sequences [1]. Several systems have been focusing on clickbaits as their research focus area [2]. The system uses BOW (Bag Of Words) approach where features of sentences as well as text is utilized. Both these data including text as well as features which are given more importance and later for categorization purpose SVM [3] named as Support Vector Machine is used.

Another system considered various tweets from clients in Twitter [4-6] to create a model with features done in handcrafted way. The factors included for categorization includes title, web page link as well as small information required for categorization [7][8]. The proposed system utilizes features extracted and categorizes these features into three including content, text or quote. Features based on similarity also created in the system [9-11].

Later these manual entry methods were not utilized and new models using deep learning methods were used [12]. Few systems based on LSTM methods were used to derive patterns in a sequential methods [13]. The system using CNN(Convolution Neural Networks) were also involved in identifying the clickbaits [14].

The existing systems have several limitations. When the data is entered manually, training data needs to be huge as clickbaits needs to be monitored periodically [15]. In case of social media, the link with clickbaits goes viral within a small timeframe when the features were recovered from the input text [16]. Systems based on CNN and LSTM also suffered from certain disadvantages. Data loss may occur in CNN based methods as pooling happens in that and on the other hand systems based on LSTM fail to focus on spatial properties of input text and focus only on sequential pattern [17].

III. Proposed System

The proposed system to detect clickbaits uses an architecture which is multi-layered. The layers include first block where input is sent, LSTM block and at last another block provides the output. The first block takes the text data as input and so data needs to be vectorized. For this the proposed system makes use of Bag Of Words approach like other systems. The next layer utilizes LSTM technique for the text data sent as input from the previous layer. The last layer mainly helps in ordering of texts from the input data and also helps in the process of routing. Two types of routing stated as static as well dynamic exists, but our approach focus on dynamic routing which mainly aims at reducing the disadvantages of LSTM approach (Fig.1). As a result of these process, the final output is produced from the text data given as input.

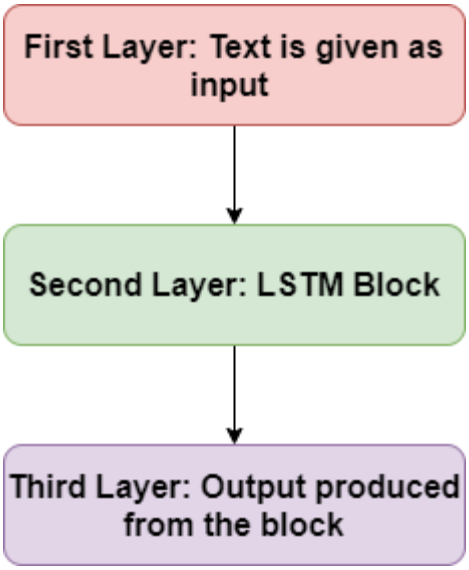

Fig.1. Three layered architecture of proposed system

IV. Results and Discussion

The proposed system when implemented works as inputs are sent as sequence. The usage of LSTM for the proposed system makes it different from the existing system as the time required for classification is also reduced. The various parameters considered for both existing system and proposed system is shown in Table.1

Table.1. Parameters in Existing System Vs Proposed System

Parameters	Existing System	Proposed System
Response Time	Medium	Low
Identification Time	High	Low

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020

Vectorization TIme	High	Low

As the time gets decreased the overall response time of the proposed system also decreases. In order to identify clickbaits, LSTM based approach is effective which is shown in Fig.2.

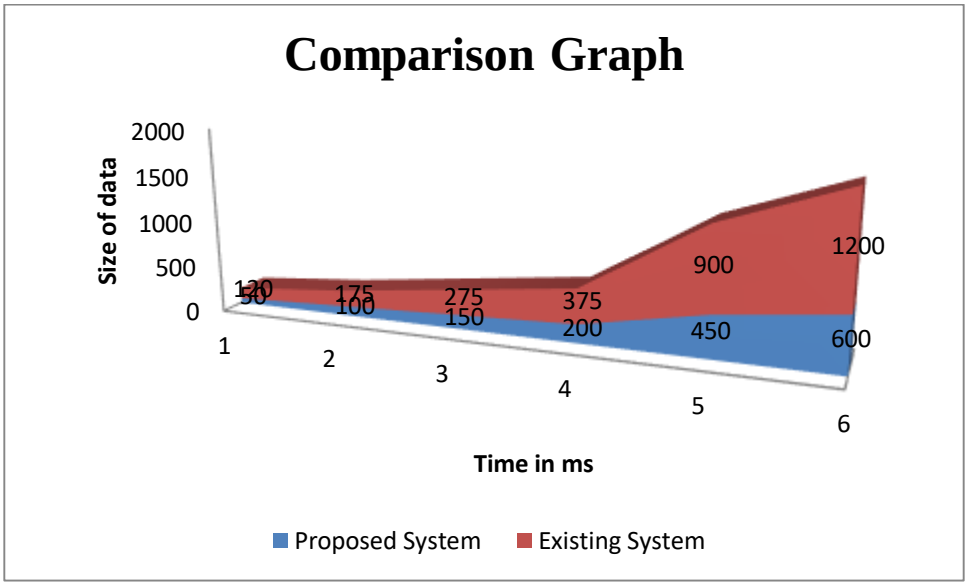

Fig.2 Comparison graph with existing system

V. Conclusion

The usage of capsule network along with its importance is discussed in the proposed system. The main advantage of the system is to identify the clickbaits using Long Term Short Memory in the second layer. The system acts as a multi-layered architecture with three layers where the input text is processed, classified using LSTM and finally the output data is received after vectorization by Support Vector Machine (SVM). The experimental results shows the efficacy of the system with the system being compared with the existing works in the literature.

References

[1] A. Srivastava, Anuradha and D. J. Gupta, ”Social Network Analysis: Hardly easy,” 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), Faridabad, 2014, pp. 128-135. doi: 10.1109/ICROIT.2014.6798311.

[2] A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, ”Stop Clickbait: Detecting and preventing clickbaits in online news media,” 2016 IEEE/ACM International Conference on Advances in Social

Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 9-16. doi: 10.1109/ASONAM.2016.7752207

[3] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16), Gerhard Brewka (Ed.). AAAI Press 3818-3824.

[4] Kim, Jaeyoung & Jang, Sion & Choi, Sungchul & Park, Eunjeong. (2018). Text Classification using Capsules.

[5] D. Li and J. Qian, "Text sentiment analysis based on long shortterm memory," 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), Wuhan, 2016, pp. 471-475. doi: 10.1109/CCI.2016.7778967.

[6] Yann LeCun and Yoshua Bengio. 1998. Convolutional networks for images, speech, and time series. In The handbook of brain theory and neural networks, Michael A. Arbib (Ed.). MIT Press, Cambridge, MA, USA 255-258.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Sabour, S., Frosst, N., & Hinton, G.E. (2017). Dynamic Routing Between Capsules. NIPS.

[9] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. 2011. Transforming auto-encoders. In Proceedings of the 21th international conference on Artificial neural networks - Volume Part I (ICANN'11), Timo Honkela, Duch Wodzisaw, Mark Girolami, and Samuel Kaski (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 44-51.

[10] Samujjwal Ghosh and Maunendra Sankar Desarkar. 2018. Class Specific TF-IDF Boosting for Short-text Classification: Application to Short-texts Generated During Disasters. In Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1629-1637. DOI: https://doi.org/10.1145/3184558.3191621.

[11] SChristopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[12] Pennington, Jeffrey, Richard Socher and Christopher D. Manning. Glove: Global Vectors for Word Representation. EMNLP (2014).

[13] Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 1997), 1735-1780. DOI=http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[14] A. Tjandra, S. Sakti, R. Manurung, M. Adriani and S. Nakamura, "Gated Recurrent Neural Tensor Network," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 448-455. doi: 10.1109/IJCNN.2016.7727233.

[15] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781. [16] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc...

[17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785-794. DOI: https://doi.org/10.1145/2939672.2939785

Received: 02 Dec 2019 | Revised: 24 Jan 2020 | Accepted: 16 Feb 2020