Headline Generation for Clickbait Detection

Uddesh Narayan Karda
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
uk8216@rit.edu

***Abstract*—Clickbait are eye-catching links on social media websites.These clickbait try to lure the users into clicking on the link. Clickbait are usually annoying and cause the social media website to have bad reputation which may lead to reduction in the user engagement with the social media website. Moreover clickbait may also contain links to phishing websites that may try to steal user data. Detecting and eliminating these clickbait posts has been an important topic of research. The most important problem that clickbait detection struggles with is the lack of labelled data. This paper proposes the use of stylized headline generator for generating stylized headlines. Stylized Headline Generator uses the original clickbait headlines and their posts to generate headlines with opposite styles.**

***Index Terms*—style transfer: clickbait detection**

I. Introduction

Since the advent of smartphones people have started to gravitate towards reading a lot of news articles on the web. Unfortunately the advent of internet journalism has also brought about the concept of clickbait. These days people only trust internet articles with a lot of engagement. The number of likes and comments on a post has become of importance. A good source of getting high engagement is to clickbait the article with a false headline. The fake headline is made eye catching so as to attract more users to read the article. The problem with clickbait is that often the article attached to the clickbait headline is misleading and creates misinformation. This may create temporary user engagement however it permanently destroys user trust and reduces future user engagement.

Moreover, clickbait are not only used for user engagement purposes but also for spreading fake news and creating misinformation among the masses.A clickbait could also lead to phishing pages, which could harm the user's privacy. Therefore detecting and eliminating clickbait is really important.

Unfortunately the task of clickbait detection is troublesome. One of the reasons making this troublesome is the unavailability of labelled data. There is a lack of labelled data available for the clickbait detection task as collecting data and manually labelling the data is an expensive task. This project is aimed to solve the data generation problem. The data generated process could be used to balance out the clickbait detection dataset and further enhance and improve the clickbait detection task results.

II. Background

Work on text generation and language modelling has been going on since the early 1990's. The earliest work in language modelling involved the use of probability and counting methods[1]. Further feed forward neural networks were used for these purposes.

With the development of Generative Adversarial networks(GAN)[2] and Recurrent neural networks(RNN)[3], using text data to generate summaries has become easier. RNN's provide with the encoder decoder[3] technology that allows for seamless extraction of numerical representation of text data. RNN's have the ability to sequentially remember the data and use the memorized data with all further steps. This allows for a summary to be generated which uses all the hidden states in the RNN.

However, for text generation, understanding and using the semantic meaning of text is an important task. Without understanding the semantic meaning of text, the model will just generate headlines that are just words copied from the original text. Previously there has been work done on this using style transfer[4] concepts. However the work done in this field does not look to preserve contents of the original document. The solution proposed in the following sections aims to use style transfer and content preservation to generate an optimal headline for the given document.

The following sections are as section III solution, section IV experiment setting, section V results, section VI conclusion and section VII acknowledgements.

III. Solution

The goal of the solution is to generate headlines with specific styles and also preserve information from the original documents. The solution structure can be seen in figure 1. In the figure A stands for Auto encoder, z stands for latent content representation of the documents, The solution involves the use of Stylized Headline Generator(SHG)[5]. The SHG makes use of GANs and consists of two parts:

1) A generator Component
2) A discriminator component

A. Generator Component

The generator component is responsible for generating headlines. The generator consists of two sub components :

1) An auto encoder

2) A Generator

1) Auto encoder: An auto encoder is a network that is used to learn efficient content representations for the input data. The learning of the content representation is an unsupervised technique. An auto encoder consists of two parts a encoder and a decoder. The encoder is responsible for encoding the text data and the decoder is responsible for decoding it. In the solution,both the encoder and the decoder are RNN's. An RNN consists of hidden states. The last hidden state before the output, of the encoder, is used as the content representation.

2) Generator: A generator takes input the content representation generated by the auto encoder and the style labels of the input documents i.e labels indicating if the input document is labelled as a clickbait or a non clickbait. The solution consists of two generators. The first one takes input the documents that are labelled as non clickbait and the headlines for those documents and generates clickbait headlines for those documents.

The second generator takes input the documents that are labelled as clickbait and the headlines for those documents and generates non clickbait headlines for those documents.

For generating headlines, the solution uses an adversarial model with the generator being one component and the transfer discriminator bring the second component. The generator is responsible for generating the headlines according to the feedback given by the transfer discriminator.

B. Discriminator component

The discriminator component consists of three sub components.

1) A style discriminator
2) A pair discriminator
3) A transfer discriminator

1) A style discriminator: The style discriminator is responsible for keeping the styles of the original headline and the generated headline different. The style of a headline is the label i.e if the headline is a clickbait or non clickbait. Therefore the style discriminator labels the generated headline with a style label. This label is used by the generator as feedback i.e if the label is same as original headline then change the generation parameters.

2) A pair discriminator: The pair discriminator is responsible for ensuring that the documents and their corresponding headlines i.e both clickbait and non clcikbait headlines are related. Basically the pair discriminator is to ensure that the document and it's two headlines correspond to each other.

3) A transfer discriminator: The goal of the transfer discriminator is to ensure that the generated headline and the original headline are as close as possible i.e that the generated headlines should be able to fool the transfer discriminator. The transfer discriminator should not be able to distinguish between the original and the generated headlines.

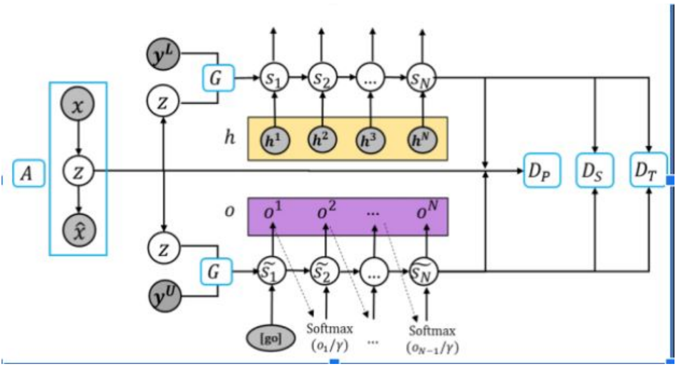

Fig. 1. Solution Structure for headline generation where A is the auto encoder component, G is the Generator component and the D components are discriminators. [5, Fig. 1]

IV. Experiment Settings

A. Parameters

The experiment starts at the encoder-decoder stage. Both the encoder and the decoder are single layer RNN's. The input dimensions for the encoder are 64 and the dimensions for the style representation are 16. The generator is also a single layer RNN with input dimension of 80. The generated headlines have a constraint of maximum length set to 25. There are two sets of experiments, one where the the first 100 words of the input document are considered and the second where the first 200 words of the document are considered. During each experiment a vocabulary file is constructed. This vocabulary file consists of words from all the documents. Only those words are considered for the vocabulary file which occur at least 3 times.

B. Dataset

The dataset was provided by the team working at the action lab at Rochester Institute of Technology under the guidance of Dr. Yu Kong. The dataset consists of one thousand plus videos. These videos were grabbed from social media websites.The video data is converted to textual data for use in processing.

For the conversion process, FFMPEG command line tool was used. Each video file is first converted to audio using FFMPEG. Further each audio file is converted to a Waveform Audio File Format (WAV file). In the next step, Sphinx speech recognition tool[6] was used. This tool is used to convert the WAV files to text files containing the transcribed text. Files with no text were discarded during the conversion process.

For some of the video files, the video had a clickbait headline in all of the frames i.e a few lines of text on each frame of the video. For those videos, one frame was selected and extracted using FFMPEG. Further the headline was extracted from that image using Tesseract[7] optical character recognition(OCR) engine.

The other part of the dataset was the data available for the ongoing webis-clickbait-17 challenge. The data consisted of a mix of clickbait headlines, their corresponding documents and non clickbait headlines and their corresponding documents.

The total dataset consisted of 14566 non clickbait instances and 6354 clickbait instances.

C. Evaluation Parameters

The results compare the current result values with other researched result values. These values have been produced by other systems meant for text generation. A few examples of these systems are the CrossA[8] system meant to transfer sentences across different styles, SVAE[9] which generates text using Variational auto encoders and SeqGAN[10] which uses adversarial networks to generate text.

The evaluation parameters used for the experiments are:

1) Readability
2) Similarity

1) Readability: Readability is the measure of how readable a text is. It is measured by the Flesch-Kincaid readability score. For Flesch-Kincaid readability score, the higher the score the easier it is to read the text. This score can also be used to calculate the Flesch-Kincaid readability grade level for better understanding of the results.

2) Similarity: Similarity here refers to the similarity of the generated headline to the original document for which the headline was generated. For clickbait headlines, the similarity must be low and for non clickbait headlines the similarity must be high. BLEU(4-gram) scores were also calculated for this purpose.

V. Results

The results for the experiments were as follows:

1) Experiment considering first 100 words of the document: Following are sample results from the experiments:

1) Original Non Clickbait Headlines:
   a) uk s response to modern slavery leaving victims destitute while abusers go free
   b) the forgotten trump roast relive his brutal 2004 thrashing at the new york friars club
   c) tokyo s subway is shut down amid fears over an imminent north korean missile attack on japan
2) Generated Clickbait Headlines:
   a) the best places to retire without a car
   b) the most common passwords in 2016 are truly terrible
   c) these are the most polluted cities in the world
3) Original Clickbait Headlines:
   a) we ll always have newsweek
   b) 18 uplifting documentaries guaranteed to put a smile on your face
   c) 12 things you realize at the end of a relationship
4) Generated Non Clickbait Headlines:
   a) conor mcgregor s final 100 years of the bull posts on the entire 2017 game
   b) the guardian view on 2017 workers the best efforts to advance the best films
   c) chrisette michele announced the president trump s inauguration with inauguration

The vocabulary file generated consisted of around 26500 words.

Readability Scores

Headline type	Clickbait	Non Clickbait
SHG	8.45	10.16
Current Solution	88.7	73.5
SeqGAN	14.25	14.54
SVAE	8.04	10.38
CrossA	8.98	10.48

Note - The authors of the original research[5] have calculated the minimum age of text readability according to Flesch-Kincaid readability score. In current solution results a Flesch-Kincaid readability score of 80.00 to 90.00 means that an average 12 year old(6th grader) can read the text easily. A score of 70.00 to 80.00 indicates that an average 13 year old(7th grader) can read the text easily.

BLEU Scores

Headline type	Clickbait	Non Clickbait
SHG	0.453	0.446
Current Solution	0.724	0.684
CrossA	0.407	0.432

Note - Higher BLEU score means the headline is more similar to the reference document/text.

Similarity Scores

Headline type	Clickbait	Non Clickbait
SHG	0.37	0.40
Current Solution	0.204	0.213
CrossA	0.20	0.22

2) Experiment considering first 200 words of the document: The vocabulary file generated consisted of around 37500 words. Note - The original research[5] does not contain experiments with 200 words of the document being considered.

1) Original Non Clickbait Headlines:
   a) uk s response to modern slavery leaving victims destitute while abusers go free
   b) the forgotten trump roast relive his brutal 2004 thrashing at the new york friars club
   c) tokyo s subway is shut down amid fears over an imminent north korean missile attack on japan
2) Generated Clickbait Headlines:
   a) the best profile picture to get the tolerance crowd riled up
   b) this is the best beard style for every face shape
   c) the best entertainment stories from the day
3) Original Clickbait Headlines:
   a) we ll always have newsweek
   b) 18 uplifting documentaries guaranteed to put a smile on your face
   c) 12 things you realize at the end of a relationship
4) Generated Non Clickbait Headlines:
   a) conorjohn glenn s neil skip a celebration of his year s eve

b) the best entertainment news from the super bowl
c) the new psa league is useful for ios and size

Readability Scores

Headline type	Current Solution
Clickbait	94.3
Non Clickbait	75.9

Note - In current solution results a Flesch-Kincaid readability score of 90.00 to 100.00 means that an average 11 year old(5th grader) can read the text easily. A score of 70.00 to 80.00 indicates that an average 13 year old(7th grader) can read the text easily.

BLEU Scores

Headline type	Current Solution
Clickbait	0.754
Non Clickbait	0.720

Note - Higher BLEU score means the headline is more similar to the reference document/text.

Similarity Scores

Headline type	Current Solution
Clickbait	0.212
Non Clickbait	0.225

As it can be seen from the above results, statistically the experiment considering the first 200 words from each document performs better than the experiment considering only 100 words from the document. Moreover analysis of the results shows that the concept of a style discriminator works very well. Generated clickbait headlines are clearly formatted as clickbaits. Generated non clickbait headlines are clearly formatted as normal sentences with an object, a subject and and the subject complement. However the results also indicate that generating non clickbaits is problematic as seldom the generated non clickbait headlines look like clickbait. Moreover the generated non clickbait headlines often have wrong subjects or objects. This indicates a problem with the pair discriminator as the pair discriminator is responsible for keep the relation between the original headlines and their documents.

A reason for non clickbait headlines being generated with clickbait style could be the presence of noise in the dataset. As the dataset contains videos that have been converted to text, the data may contain noise which could affect the generation process. Another observation through multiple experiments was that during the vocabulary file construction if the minimum number of appearances of a word was reduced to 1 then this had no positive effect on the results. Reducing the minimum number of appearances of a word did not increase the quality of the results however it did increase the runtime. Moreover the model also tends to plateau at certain points. This was solved by reducing the back propagation occurrence threshold.

Another approach to improve the model was to reduce similar sentences from the document. By doing this, the total number of words in the document would go down, however the semantic meeting of the text would not change. For this purpose, textrank[11] algorithm was used. Textrank algorithm is based on the pagerank[12] algorithm. It uses cosine similarity of statements to eliminate similar statements. Textrank algorithm was used on one of the documents and the results were as follows:

Original document length : 3214 characters
1st ranked summary length : 1358 characters
2nd ranked summary length : 749 characters (best result)
3rd ranked summary length : 213 characters
4th ranked summary length : 750 characters
5th ranked summary length : 162 characters

As seen in the results, textrank algorithm has great usage in reducing text document lengths. However upon inspection it was observed that the best result was the document length with 749 words. The best result means that the document is comparatively shorter and also keeps the semantic meaning intact. If the textrank algorithm is used on all the documents, the amount of preprocessing would be much larger than currently being done. Therefore the textrank algorithm was not deemed fit for use.

VI. Conclusion

Data generation has been and currently is a challenging problem. However the experiments performed demonstrate that the solution i.e Stylised Headline Generator(SHG)[5] is capable of producing readable text that can be used for clickbait detection purposes. From the experiments performed it can also be concluded that generating non clickbait headlines for documents is very challenging even with SHG. Therefore the solution SHG generates high quality headlines for the documents that can be used for clickbait detection purposes.

Acknowledgment

The author would like to thank Kai Shu and his team for permission to work on their work. The author would also like to thank everyone at the Rochester Institute of technology action lab for being helpful and providing the dataset.

VII. Future Work

The results of the experiments show that there are words being repeated multiple times in the same headline. This shows that the model is unstable. To solve this problem the concept of attention[13] can be applied here. Moreover the current problem is the document size. The solution only considers a maximum of 200 words. A solution to this problem is the use of models and algorithms like textrank[11] algorithm to summarize documents. Also the solution could be further expanded to generate longer headlines.

References

[1] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ser. ACL '96. USA: Association for Computational Linguistics, 1996, p. 310–318. [Online]. Available: https://doi.org/10.3115/981863.981904

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, p. 139–144, Oct. 2020. [Online]. Available: https://doi.org/10.1145/3422622

[3] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[4] J. Mueller, D. Gifford, and T. Jaakkola, "Sequence to better sequence: Continuous revision of combinatorial structures," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2536–2544. [Online]. Available: http://proceedings.mlr.press/v70/mueller17a.html

[5] K. Shu, S. Wang, T. Le, D. Lee, and H. Liu, "Deep headline generation for clickbait detection," 08 2018.

[6] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[7] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 629–633.

[8] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," 2017.

[9] S. Semeniuta, A. Severyn, and E. Barth, "A hybrid convolutional variational autoencoder for text generation," 2017.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[11] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[12] I. Rogers, "The google pagerank algorithm and how it works," 2002.

[13] F. Meng, Z. Lu, H. Li, and Q. Liu, "Interactive attention for neural machine translation," 2016.