Chapter II

Literature Review

1. Problem of clickbait

Headline title is one of the important components in news that help readers to find good quality reading. For several readers, it helps to filter the article that they are preferred to read. Oftentimes media companies apply a clickbait title, for attracting reader which compromises the representation of the article via headline. The clickbait headline has become one of the issues in journalism, as it has challenged both the media corporate agenda and reader desire. On one side, Media companies need a higher user click-through rate for articles, while users desire a more prominent display of headlines to assist them in locating the most important articles. Computer science research explores this phenomenon in natural language, with one of the most common areas including clickbait detection.

1. We can use

Classification has become one of the ways to combat clickbait. However, there is a possibility to combat clickbait headlines by automatically convert clikbait to non clickbait titles using style transfer. This Could be seen as continuation of clickbait classification. The research of style transfers has been extensively study for various style switching task such as formality (Rao and Tereault. 2018; Briakou et al, 2021), Politeness (Madaan et al, 2020), Toxicity (dos Santos, Melnyk, and Padhi 2018; Tran, Zhang and Soleymani, 2020), Biased (Pryzant et al. 2020)

* Feature extraction

1. Modeling and Classification

There are two methods of text style transfer commonly use that strongly related on the data. If data has a parallel desired output, sequence to sequence become the most commonly use method that incorporate econder-decoder architecture.the text style transfe rcan be seen neural machie translateipn task it can be raeltd to text generation tasks. .A customize version can be also be applied by incormprated using LSTM () , Transformer(). The second method solve the challenge of non parallel corpus, where the text does not have similar pair on different version. Xxx et al (20xx) classifed text styel stranfero into three methods : disentangelemt, prototype edicin, and pseudo curpos construction.

Common task within style transfer are formality (Briakou et al. 2021, Rao and Tetreault. 2018), politeness (Madaan et al. 2020), Toxicity (Tran, Zhang and Soleymani. 2020), Authorship (Xu et al. 2012; Carlson, Riddell, and Rockmore. 2018), Simplicity (Cao et al. 2020; Tan and Goonawardene. 2017, Zhu, Bernhard and Gurevych. 2010, den Bercken, Sips, and Lofi. 2019) and biased (Pryzant et al. 2020). With this type of research, clickbait can be classified as a biased based task which refer as the task for neutralizing the sentence that has clickbait style.

# tak about the division of style style transfer

Jin et al (2021) mentions three approach style transfers for non parallel corpus. First, disentanglement which using latent space for separate both content and attribute. Second, prototype editing which based on taking the template and attribute of input sentence and submit it to neural model for generate new output. Third, pseudo-parallel corpus construction that simulate the use of supervised based sequence-to-sequence data on generated comparison data.

# example disentanglement :

# example of on prototype editing :

Prototype editing refers to style transfer method that incorporate discrete method called delete, retrieve and generate (Li et al.2018). This method try to generate the output by combining both the template of the sentence and the attribute that will determine the style of the output. Xxx et al (xxx) divide the phase/stage of protoyping editing into three : Attribute marker detection, Target attribute retriever, generation from prototypes.

Li et Al (2018) build a model by focus on transforming the attribute markers, which it refer as word and phrase with certain indicated feature. The model utilize four discrete baselines : identifying the attribute marker such as words or phrase, delete the words and phrases, comparing and generating new sentences based on the desired sentence goals. In addition, neural generative model is included in generate the new sentence using RNN decoder that utilize the target parameter. This model is tested on two different style transfer task : converting both positive to negative on Yelp and Aamazon reviews, and converting sentence to become more romantic or humorous. The result shows the model outperform the baseline model with 34% success rate.

The approach described by Li et al. (2018) is based on frequency-ratio methods, in which the frequency of co-occurrence of the attribute of the input sentence with the desired sentence is compared. In comparison, attention-based methods are also used by embed transformers within the retrieve steps (Sudhakar, Upadhyay, and Maheswaran. 2019). Beyond the delete mechanism, improvement is also set up during the generative phase by incorporating a transformer instead of RNN. Sudhakar, Upadhyay, and Maheswaran (2019) method is the improvement ofof Li et al. (2018) which found the flaws on previous method, that delete method using x tend to fail to find representative method of the

# example of Pseudo-Paralel

The idea of pseudo parallel method is adopted from the use of parallel method for text style transfers. Pseudo pralal method try to create a pairs of text which

# Evaluation.

The most common evaluation method for text style transfer has been BLEU (Papineni et al. 2002) which is commonly use for text translation. BLEU has been use on range from text style transfer for borth non deep learning (Xu et al. 2012; Jhamtani et al.2017) and deep learning style transfer algorithm. (Rao and Tetreault. 2019; Li et al. 2018; Jin et al. 2019). In addition, a new evaluation matrix is utilized by focusing on measuring attributes that are correctly classified compared to the total test samples. Hue et al (2017), Shen et al. (2017), Fu et al (2018), and Li et al (2018) use this matrix by classifying the resulting sentences from different works.

Bab 3

1. Text Style Transfer
2. Clickbait

Clickbait is the content of the internet that is based on sensational sentences while reducing the quality and accuracy with the goals of generating revenue. Clickbait relies on the curiosity of the reader but not fulfilling the understanding of the user. (Manjesh et Al, 2017). Clickbait is the phenomenon that resulted from the understanding that click-through has become the currency of the internet, which is the idea that internet creators strive for more clicks (Molina et Al, 2021).

Beyond that, Clickbait is also phenomenon that result from the incentivisation of the popularity as the metric within content creation. Clickbait history can be traced from the rise of new american news media and it has since in journalism. Critics argue that clickbait create mass of additive distraction for reader (Lischka and Marcel, 2021).

1. Preprocessing
2. Tokenization

Tokenization is the activity of breaking the documents or the sentence into pieces called “token”. This process is the first step of document processing in natural language processing. Token-based on the occurrence of the words can also be used as the vector input that represents the document. In machine learning, tokenization is useful for numerical conversion of words. Tokenization can be used for separating words, sentences, characters and also subwords. This day, neural language model approach has become one of the common tokenization options (Mielke et Al, 2021)

1. BLEU

Bilingual Evaluation Understudy (Papineni, 2002) is the evaluation metric that is commonly used for text translation which compares the reference sentence to the generating result. BLUE act as the alternative to human evaluation. The idea is to compare the n-gram of the sentence from both result and reference which is independent of the order and based on the co-occurrence of the word. In addition, it counts the total of the words that existed in the reference sentence by the total output sentence. BLUE also incorporates a penalty strategy if the sentence outputs are shorter than the references called brevity penalty.

1. N-grams Language Models

N-grams language models are the model of the probability models on the current word based on the previous word which concentrates on the N-1 previous words as the representation of the entire history of the base of the word on one word. The Markov model is used to predict the current word without seeing the whole history of all previous words. the n-gram used maximum likelihood estimation to find the estimated probabilities by normalizing counting the corpus to result between 0 to 1 (Jurafsky, 2020). [need formula]. The relative Frequency Cooccurrence Matrix is used in this research. the n-gram estimate is based on the observed frequency of sentence divided by the frequency of the prefix. Relative frequency is based on maximum likelihood estimation.

1. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a probability metric that counts the frequency of the words in a document within a collection of documents. The Term Frequency depends on the frequency of a certain word selected compared within the relative of the document Luhn, H. P. (1957). Inverse document frequency shows the commonality of the word compared to the corpus which a higher result of the TF-IDF value indicates the importance of the term (Sparck Jones, 1972). [need formula]

1. Encoder-Decoder Model

Encoder-decoder is the sequence-to-sequence model that comprehends the context of the data and generates predefined output. Encoder-decoder model has three keys: encoder, context vector, and decoder. The encoder works to generate a context vector that is the input sequence's context representation. A context vector is a vector that represents the essence of the input sequence that will be used for generating the output in the decoder. Decoder results in a predetermined length output sequence resulting from hidden states (Jurafsky and Martin, 2020).

1. Neural Network

Neural networks are a subtype of machine learning in which the structure is inspired by the human brain, and closely reflects the way biological neurons communicate. An artificial neural network is composed of multiple layers that are stacked on top of one another [@lecun2015deep]. The input layer is the first, the concealed layer is the second, and the output layer is the third.

Within a neural network, the weighted sum of the inputs is computed with bias, and non-linearity is applied toward the ending of the computation. To minimize the error, the neural network is trained by propagating the error into the previous layer using the chain rule as backward propagation This is accomplished by counting the gradient based on the loss function computed by comparing the predicted output to the reference output.

Regularization is used to optimize the neural network by using dropout (Hinton et al. 2012). Apart from regularization, contemporary neural network architects embed a computational graph that supports gradient computing and parallelism enabled by GPU (Jurafsky, 2020).

1. Recurrent Neural Network

A recurrent neural network is a type of neural network that is involved with sequential data. Recurrent neural networks encode memory by taking information not only from the input but also from within themselves. Recurrent neural networks are commonly used to detect sequence-based data such as handwriting, genome sequence, and time-series data. RNN incorporates cycles when passing the data between the nodes. As a sequential network, recurrent neural network input depends on the previous result of the network as the next input of the work. Recurrent neural networks utilize backpropagation through time as the mechanism for calculating the gradient which coins the error within each of the steps.

Vanishing gradient and exploding gradient [@glorot2010understanding] are the two main problems that recurrent neural networks face. One of the ways to solve it is LSTM (Long Short-Term Memory) was introduced as a special type of RNN that has the ability to store information for a long period of time. A deep recurrent neural network (DRNN) is a combination of LSTM and RNN it has been shown that DRNNs are able to learn long-term dependencies in time series data. Another type of Recurrent neural network incorporating bidirection (RNN-Bidirectional) is used to learn the dependencies in both forward and backward directions.

Abstrct

Luhn, H. P. (1957)

(Mielke et Al, 2021) Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP

(Manjesh et Al, 2017) Clickbait Pattern Detection and Classification of News Headlines using Natural Language Processing

(Lischka and Marcel, 2021). Clickbait news and algorithmic curation: A game theory framework of the relation between journalism, users, and platform