Clickbait Spoiling

A B. Tech Project(Phase-1) Report Submitted in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Technology

by

Aditya Sinha
(190101004)

under the guidance of

Dr. Sanasam Ranbir Singh

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled “**Clickbait Spoiling**” is a bonafide work of **Aditya Sinha (Roll No. 190101004)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Sanasam Ranbir Singh**

Assistant/Associate

Professor,

May, 2022

Guwahati.

Department of Computer Science & Engineering,
Indian Institute of Technology Guwahati, Assam.

Contents

Chapter 1

Introduction

In this fast-paced and dynamic world, a cut-throat competition exists between different media outlets. With the advent of technology, the general media has shifted from offline to online.In the older times, the consumers didn't tend to change their mode of media consumption but in current scenario, with the vast plethora of media available online, the media outlets have resorted to present frivolous news with catchy headlines to generate more revenue with the *clicks* of users. With a high number of clicks, the outlet will get a higher number of advertisement slots which, in turn, will generate more revenue. These headlines are generally attached with a URL to an article with an unimportant news. These headlines are called **Clickbaits**.

1.1 Definition of Clickbait

Clickbaits are catchy headlines associated with certain links which generate curiosity in the mind of readers by referring to something shocking or someone unnamed. According to Merriam-Webster Dictionary, Clickbait is something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest.

1.2 Cons of Clickbaits

According to G. Loewenstein, Clickbaits generate a *Curiosity Gap* in users with these headlines which affect them to lose their focus while reading different news. This leads the users to avoid important news. Some of these websites redirect us to unprotected links which can be a case of fraudery. These also harm the credibility of journalism sector.

1.3 Clickbait Spoiling

Clickbait spoiling generates a phrase, paragraph or multiple paragraphs which can curb our curiosity by answering the question. In this case, we do not need to click on the clickbaits.

1.4 Problem Statement

Given, different kinds of clickbaits we intend to find appropriate spoilers for them. We have to spoil the clickbait for the ease of users.

Chapter 2

Review of Prior Works

A literature survey was done to understand more about works done on headlines, clickbaits and work done, if any, to spoil them.

2.1 Clickbait

News were made into a kind of spectacle to affect Readers' psychology. Dvorkin wrote a column discussing how clickbait will lead to the eventual demise of journalism. The notion that clickbaits play with the psychology of readers. Bloem and Chen independently discussed about the ill effects of clickbaits. Clikcbaits have been on a slow decline for quite a few time as they have lost their sense of credibility for the mainstream users.

2.2 Clickbait detection

Although there was a boom of clickbaits in the media industry, only limited number of approaches were made to detect clickbaits. Despite the claims of various Social Media Platforms in early 2010s, clickbaits continued to thrive. this led to the birth of some adhoc methods to detect clickbaits but they failed as they made the headlines more incoherent. This was due to the fact that they employed a fixed set of methods and rules to identify clickbaits which wasn't the case all the time. Using frequent terms found in clickbaits and other tweet-specific traits, Potthast made an effort to identify clickbait Tweets on Twitter. The 'Downworthy' browser plugin recognises clickbait headlines by utilising a predefined collection of typical clickbait words and then transforms them to worthless text. Thus, a need of fine tuned model arose. Abhijnan proposed a browser extension, which caution the users about different clickbait headlines. They examined clickbait and non-clickbait headlines and found intriguing distinctions between the two, which coule be used to detetct clickbaits. The browser also allows readers to ban clickbaits and automatically prevents similar ones on subsequent visits.[CPKG16]

2.3 Clickbait spoiling

No work was done in clickbait spoiling prior to Hagen, Fröbe, Jurk and Potthast. They used Information retrieval techniques like Question Answering and Passage Retrieval to generate spoilers for Clickbaits.

2.3.1 Question Answering

Question Answering is a kind of IR systems where they provide expected answers to the questions. They are unlike most of the IR techniques which work on ranking different documents. Tese systems answer the questions based on resources. The question belong to both-open and close domains whereas the resources can be both structured or unstructured. Stanford Question Answering Dataset(SQuAD) acts as one of the standard datset on which Question Answering Models are trained.Different BERT Models and GRU Models have been used for question answering in different systems. [PB21][ZTS$^+$21] We will discuss two types of Question Answering here - Text based and Data based

1. Text Based : When attempting to answer a question posed by a text, human and machine reasoning differences are at their greatest. Humans can only answer one

question at a time. Identify the appropriate sentence or sentences from the supplied paragraph while it is the machine's challenge, passage can be made with ease. There is now a lot of study being done in this area. not capable of outperforming human performance. datasets like as This is what SQuAD and many more are made to do. Comprehensive challenge and machine improvement performance achieved utilising deep learning algorithms includes, but is not restricted to, attention-based education, unique supervision of education, BERT, GPT models, etc. This can be broken down into three basic, independent subtasks: question analysis, document retrieval, and answer extraction. These three subtasks are generally followed by Question Answering Systems. However, the methods they use to carry out each subtask may vary. The question-answering problem involves using natural language processing to connect users who ask various kinds of inquiries to the QAS. Factoid inquiries, in particular, are those that are primarily about Named Entity, for instance, utilising the words: When, Where, How much/many, Who, and What, which inquire about time and date, location, individual, and group, respectively. The second category includes inquiries regarding the definition of phrase or idea. Another form of inquiry that is challenging to respond to is one that begins with the words "Why" or "How," and almost any attempts are made to do so.

2. Knowledge based: A knowledge graph is a frequent term used to describe the ontology-based information displayed. The information may be represented in the knowledge graph as a triplet of subject, relation, and object, where relation is the link that links subject and object. Finding the right connections between nodes to obtain the solution is the main issue in this situation. Many academics have been drawn to this line of study, which ranges from basic question-answer mapping to breaking the question up into many subsets and processing each using a deep learning-based model. Query languages like SPARQL are also used to query web of data.

2.3.2 Passage Retrieval

Their passage retrieval algorithm needs the question's intended response type or a bridging inference between the question and expected answer type. However, neither the bridging inference techniques nor the answer type ontology were sufficiently explained for us to implement.[KOM+20] Two techniques, Multi-Text and IBM passage retrieval are to be discussed here:-

1. IBM: A density-based passage retrieval system, the MultiText algorithm [2, 3] prefers short passages with plenty of terms and high idf values. The algorithm scores each passage window based on the number of query phrases in the passage and the size of the window. Each window starts and finishes with a query term. Once the passage with the highest score has been determined, our method constructs a new window around the original passage's centre that is the necessary length. The Waterloo MultiText algorithm employs a variation of idf for the term weights due to the structure of their index. Our implementation, however, adheres to the idf standard definition.

2.

They identified different kinds of clickbaits - Phrase, Spoiler and Multi-Part. This work was done as an IR(Information Retrieval) technique. Firstly, Spoiler Classification was done and Clickbaits were divided into the three types based on their spoiler. Then, based on whether they were phrase-based or passage-based, Feature based and transformer based methods were used.

Chapter 3

Objectives and Dataset Analysis

give details of your algorithm

3.1 Objective

Based on the approaches studied in the literature survey, we propose the following objectives

1. Clickbait Spoiler Type Identification

2. Clickbait Spoiler Generation based on types

3. Study of dependence of Spoiler Type on Spoiler Generation

3.1.1 Clickbait Spoiler Type Identification

After reviewing the different kinds of clickbait posts, it was observed that differentiating clickbaits on the lengths of spoilers will be beneficial in generating spoilers for them. If we get the knowledge that whether we are looking for a phrase, passage or multiple parts from the document our task becomes easier on a considerate level. This information retrieval can be done on extending the Question Answering and Passage Retrieval state of the art tasks.

3.2 Clickbait Spoiler Generation based on types

Different spoilers can exist for same type of clickbait posts. Let's say a clickbait which can be spoiled both by a phrase and a coherent passage of text. A prior knowledge of the kind of spoiler will tend the accuracy of the model to be better as they will identify the spoiler of the post on that particular type only.

3.3 Dependency of the spoiler type on spoiler generation

In the prior works, sometimes the model with no classifier outperformed the model which firstly, identified the kind of clickbait through classification. Although, this model performed poorly than the Oracle Model. In the oracle model, pre-determined type of clickbait post is feed which gives best result among all.

3.4 Dataset Analysis

Corpus of 4000 datasets were analysed to study the problem. There were 14 attributes associated with each record. They are as follows:-

1. uuid : 32-digit Universal Unique Identifier which uniquely labels each clickbait

2. postId : Post ID of the clickbait from the respective platform

3. postText : The Clickbait Text

4. postPlatform : The Platform on which clickbait is posted

5. targetParagraphs : All content of the post URL stored as array of strings seperated by a delimiter

6. targetTitle : The title of the article in the post URL

7. targetDescription : Starting few sentences of the post article

8. targetKeywords : Important relevant words present in the article along with tags(if available)

9. targetMedia : An array of links containing media present in the article

10. targetUrl : URL of the article

11. provenance : Contains information about the source of clickbait, handle which manually spoiled and the subsequent manual spoiler in JSON format

12. spoiler : A phrase, a passage or multiple phrases/passages, present in the article which answers the question without clicking the link

13. spoilerPositions : This is a 2D array. The inner array consists of starting position and ending position of spoiler in the article.

14. tags : An appropriate label is given to the spoiler from these words-'Phrase','Passage','Multi'.

Among these attributes, 5 attributes namely postId, postPlatform, targetDescription, targetKeywords, targetMedia aren't used for training. They just contain information about the records.

Chapter 4

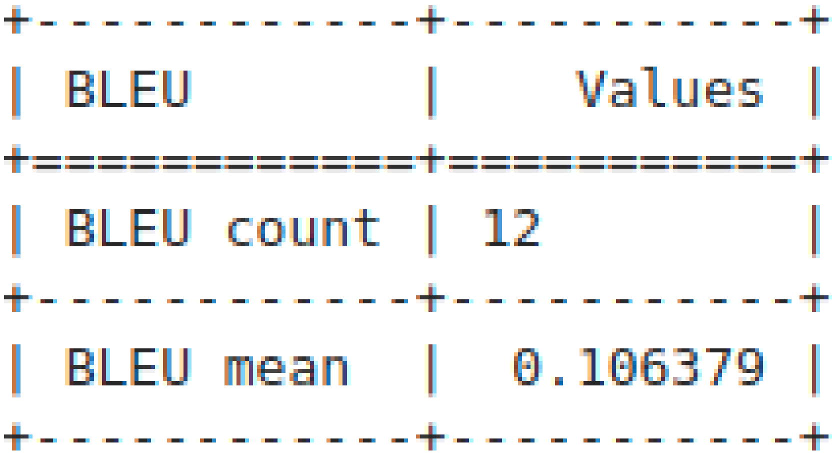
Baseline Models

We will use three Baseline Models to verify our objectives. The baseline models aim to find different stats and evaluate the performance of data on dataset. The baseline models used are Classic Feature Based Models and Transformer Models such as BERT, RoBERTa etc. We used Metrics like Bleu-4, Meteor and BertScore to evaluate the Baseline Models with the Ground Truth of Validation Dataset(consisting of 800 posts)

4.1 Bleu-4

It measures the closeness or proximity between the Ground Truth and Machine Prediction. This can be calculated as $\frac{\sum_{C\in(Candidates)}\sum_{n-gram\in(C)}\text{Count}_{\text{clip}}(n-gram)}{\sum_{C'\in(Candidates)}\sum_{n-gram\in(C')}\text{Count}(n-gram)}$ This equation signifies the modified n-grams precision on text. Bleu-4 is the type of bleu where weight cumulative grams from 1 to 4 is 0.25 each. The merit of BLEU is that it correlates strongly with human judgments by averaging out individual sentence judgement mistakes throughout a test corpus rather than attempting to discern the precise human assessment for each phrase: quantity leads to quality.[Bro17][MLLL16]

we have attached a table with the Bleu-4 metric of Naive Bayes.



```
+------------+-----------+
| BLEU       |    Values |
+============+===========+
| BLEU count | 12        |
+------------+-----------+
| BLEU mean  |  0.106379 |
+------------+-----------+
```

Fig. 4.1 BLEU-4 Metric

4.2 Meteor Score

METEOR is a machine translation evaluation measure that is based on an extended idea of unigram matching between machine translation and human-produced reference translations. METEOR is evaluated by calculating the connection between metric scores and human translation quality judgements. BLEU falls short in various areas, such as the fact that the BLEU score is more precision-based than recall-based. In other words, it is based on determining if all terms in the created candidate appear in the reference generated by a manual evaluator. However, it does not check to see if all of the terms in the reference are covered. In this example, a meteor enters the frame. Precision(P)=$\frac{m}{w_t}$
Recall(R)=$\frac{m}{w_r}$

$$F = \frac{10PR}{R + 9P}$$

4.3 BERTScore

BERTScore is a semanctic method of evaluating performance. We employ a reference sentence and a candidate sentence. To represent the tokens, use contextual embeddings, and compute matching using cosine similarity, with the option of adding weight from inverse document frequency scores. BERTSCORE, is a relatively new statistic for comparing generated text to already established practices BERTSCORE was specifically created to be basic, independent of task, and simple to use. Our investigation demonstrates how BERTSCORE overcomes some of the drawbacks of conventional metrics, particularly for difficult adversarial scenarios.[ZKW$^+$19]

4.4 Baseline Model 1 : Machine Learning Model for Spoiler Generation

In the first Baseline Model, feature based- models such as Naive Bayes, SVM and Logistic Regression were used. Appropriate Features were extracted from both ths clickbait text and the attached articles. These features were used for Clickbait Post:-

1. Term Frequency(TF)
2. Term Frequency, Inverse-Document Frequency(TF-IDF)
3. Part-Of-Speech(POS) Tag consisting of Unigrams and Bigrams.

The following features were used for Linked document.:-

1. Term Frequency, Inverse-Document Frequency(TF-IDF)
2. Part-Of-Speech(POS) Tag consisting of Unigrams and Bigrams.

4.5 TF-IDF

TF-IDF, which stands for term frequency-inverse document frequency, is a metric that can be used to quantify the significance or relevance of string representations (words, phrases, lemmas, etc.) in a document among a group of documents. It is used in the fields of information retrieval (IR) and machine learning (also known as a corpus).

4.6 POS

The process of assigning a portion of speech tag or supplemental philological class is known as part of speech (POS) tagging. every single word in a sentence should be given the signal.

A chi-square feature selection phase picked all post-based features and 70 percent of document-based features for the three feature-based techniques. Post-based characteristics are weighted four times more than document-based ones.

4.7 Baseline Model 2: Transformer Based Model for Spoiler Classification

In both the transformer models, the input for the attention models were the concatenation of the clickbait post and linked document.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Thus, firstly we learnt about clickbaits, the motivation of spoiling the clickbaits, did a literature survey on the problem of clickbait spoiling and learnt that it is a novel problem which hasn't much. In the literature survey, we learnt about Question Answering and Passage Retrieval systems. We learnt about prior works done in Clickbaits and its detection. We learnt about different kinds of spoilers and spoiler generation techniques. After the literature survey, we proposed three objectives to work upon in this problem of clickbait spoiling. We explained the reasoning behind each objective. Then, dataset analysis was done and important attributes were identified which will help to solve our objective. After proposing the objectives, the code was run on three kinds of baseline models. Two of them were transformer models and one was feature-based machine learning model. Performance of these models were evaluated on different parameters like Bleu-4, Bertscore which differed syntactically and semantically etc. Transformation based models showed better accuracy as expected.

5.2 Futurework

We need to propose a model which improves the accuracy over baseline models. Ranking Models can be beneficial in improving these things. So, we will explore in the field of ranking models. By ”ranking,” we imply arranging documents by relevance in order to identify relevant material in relation to a query. We will also explore in the Summarisation models as they will also deliver the same objective. The task of summarisation is to create a shorter version of a document while retaining its significant content. This is the proposed futurework for this problem. [GFP$^+$20]

References

[Bro17] Jason Brownlee. A gentle introduction to calculating the bleu score for text in python, Nov 2017.

[CPKG16] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16, 2016.

[GFP+20] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing Management*, 57(6):102067, 2020.

[KOM+20] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[MLLL16] Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. Interactive attention for neural machine translation. 10 2016.

[PB21] Hariom A. Pandya and Brijesh S. Bhatt. Question answering survey: Directions, challenges, datasets, evaluation matrices, 2021.

[ZKW⁺19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2019.

[ZTS⁺21] Munazza Zaib, Dai Hoang Tran, Subhash Sagar, Adnan Mahmood, Wei E. Zhang, and Quan Z. Sheng. Bert-coqac: Bert-based conversational question answering in context, 2021.