Research Master Thesis

Clickbait anatomy: Identifying clickbait with machine learning

Nguyen Ngoc To Ngan

a thesis submitted in partial fulfilment of the requirements

for the degree of

MA Linguistics

(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab

Department of Language and Communication

Faculty of Humanities

Supervisor: Isa Maks

Second Reader: Hennie van der Vliet

Submitted: August 28, 2020

Abstract

This research focuses on the exploration of linguistic patterns in clickbait, aiming at characterizing clickbait from serious formal news using quantitative and qualitative analyses. Two significant findings about the nature of clickbait are discovered: (1) there are noticeable changes in terms of syntactic structures and topics in clickbait headlines, (2) the contents of a clickbait article can provide valuable discourse-level information that can be used to differentiate clickbait from non-clickbait. Based on the results of the analysis, three types of features are selected to be used for machine learning systems: stylometic features with encoded sequential part-of-speech and dependency tags, word embeddings, and document embeddings. The best system which uses Support Vector Machines algorithm and word embedding features achieves precision and recall scores of 0.82, as well as 82% of accuracy.

Declaration

I, Nguyen Ngoc To Ngan, declare that this thesis titled, "Clickbait anatomy: Identifying clickbait with machine learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Acknowledgements

I would like to express my deep gratitude to Dr. Isa Marks, my research supervisor, and Dr. Hennie van de Vliet for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I also would like to thank the research groups of Professor Martin Potthast from University of Leipzig and Dr. Abhijnan Chakraborty from Max Planck Institute for Software Systems who make the data used in this research available.

Finally, I would like to extend my thanks to professors and faculty members of Computational Lexicology & Terminology Lab (CLTL) who provide the best education for my Research Master program in Human Language Technology.

Contents

1 Introduction 12

1.1 Research Problem 12

1.1.1 Research motivation 12

1.1.2 What is a clickbait? 14

1.2 Research Question and Objectives 18

1.3 Outline of the Thesis 18

2 Literature Review 20

2.1 Qualitative analyses on the characteristics of clickbait 20

2.2 Automatic clickbait detecting system 23

2.2.1 Neural Network 23

2.2.2 Machine Learning 29

3 Data 37

3.1 The Webis Clickbait Corpus 2017 37

3.2 Clickbait Headline Corpus 39

4 Linguistic Analysis of the Datasets 40

4.1 Methods of Analysis 41

4.2 Linguistic Analysis 44

4.2.1 Clickbait Headlines vs Non-Clickbait Headlines 44

4.2.1.1 Word/token-level 44

4.2.1.1.1 The average length of words/tokens 44

4.2.1.1.2 Contraction 45

4.2.1.2 Syntax level 47

4.2.1.2.1 Sentence length 47

4.2.1.2.2 Punctuation 48

4.2.1.2.3 The maximum length of dependency path . 49

4.2.1.2.4 Part-of-speech and part-of-speech n-gram . . 53

4.2.1.2.5 Sentence types 59

4.2.1.2.6 Conditionals, superlatives, passives 63

4.2.1.3 Semantic Level 64

4.2.1.3.1 Named entities 64

4.2.1.3.2 Subjects 66

4.2.1.3.3 Pronoun 68

4.2.1.3.4 Sentiment 69

4.2.2 Analysis of Clickbait Contents 70

4.2.2.1 Stylometric characteristics 71

4.2.2.2 The similarity between the contents and headlines . . 72

4.3 Discussion 75

4.3.1 On feature selection 75

4.3.2 On the nature of clickbait 76

5 Automatic Detection Systems 79

5.1 Experiment Settings 79

5.1.1 Feature Engineering 79

5.1.1.1 Sylometric Features 79

5.1.1.2 Embedding Features 81

5.1.2 Machine Learning Architectures 82

5.2 Results 84

5.3 Discussion 86

6 Conclusion 88

6.1 Summary 88

6.2 Suggestions for Further Research Work 89

A (Example of Clickbait and Non-clickbait from Potthast Corpus) 100

B (Report on performance results of each classifier in respect to features) 103

B.1 Stylometric Features 103

B.1.1 Logistic Regression 103

B.1.2 Random forest 104

B.1.3 Support Vector Machine 104

B.2 Word embedding Features 105

B.2.1 Logistic Regression 105

B.2.2 Random forest 105

B.2.3 Support Vector Machine 106

B.3 Combined features 106

B.3.1 Logistic Regression 106

B.3.2 Random forest 107

B.3.3 Support Vector Machine 107

B.4 Document embedding features 108

B.4.1 Logistic Regression 108

B.4.2 Random forest 108

B.4.3 Support Vector Machine 109

List of Figures

1.1 Example of clickbait from Biyani et al. (2016) 13

4.1 The distribution of different average token lengths 45

4.2 The distribution of different numbers of contraction in headlines . . . 45

4.3 Plotting the number of tokens (Chakraborty et al., 2016) 46

4.4 Plotting the number of tokens (Potthast et al., 2018b) 47

4.5 Plotting the frequency of punctuation (Chakraborty et al., 2016) . . . 48

4.6 Plotting the frequency of punctuation (Potthast et al., 2018b) 49

4.7 Dependency tree of the headline "A 22-Year-Old Whose Husband And Baby Were Killed By A Drunk Driver Has Posted A Gut-Wrenching Facebook Plea" 50

4.8 Dependency structure of the two sentences: "The woman who wore a red scarf is killed by a man." and "The woman is killed by a man who wore a red scarf." 51

4.9 Distribution of longest syntactic dependencies between dependencies and roots in clickbait and non-clickbait headlines 52

4.10 The distribution of part-of-speech tags for words in clickbait and non-clickbait headlines (Chakraborty et al., 2016) 53

4.11 Plotting the frequency of each part-of-speech (Potthast et al., 2018b) 54

4.12 Proportion of clickbait and non-clickbait headlines that use modal verbs 55

4.13 Plotting the frequency of each part-of-speech tri-gram (Chakraborty et al., 2016) 56

4.14 Plotting the frequency of each part-of-speech four-gram (Chakraborty et al., 2016) 57

4.15 Plotting the frequency of each part-of-speech tri-gram (Potthast et al., 2018b) 58

4.16 Plotting the frequency of each part-of-speech four-gram (Potthast et al., 2018b) 59
4.17 Plotting the percentage of listicle headlines 60
4.18 Plotting the percentage of questions 62
4.19 Plotting the percentage of conditionals 63
4.20 Plotting the percentage of superlatives 63
4.21 Plotting the percentage of passives 64
4.22 Plotting the distribution of the number of entities in clickbait and non-clickbait headlines 65
4.23 Plotting the distribution of each type of entities in clickbait and non-clickbait headlines 65
4.24 Plotting the most frequent subjects in clickbait and non-clickbait headlines (Chakraborty et al., 2016) 66
4.25 Plotting the most frequent subjects in clickbait and non-clickbait headlines (Potthast et al., 2018b) 67
4.26 Plotting the most frequent pronouns in clickbait and non-clickbait headlines (Chakraborty et al., 2016) 68
4.27 Plotting the most frequent pronouns in clickbait and non-clickbait headlines (Potthast et al., 2018b) 69
4.28 Plotting the distribution of sentiment scores in clickbait and non-clickbait headlines 70
4.29 Plotting the average sentence length in clickbait and non-clickbait contents 71
4.30 Plotting the distribution of the average token length and the sentiment scores in clickbait and non-clickbait contents 72
4.31 Plotting the average similarity scores between headlines and contents 72
4.32 Plotting the proportion of sentences in contents that are similar to headlines 73

5.1 Formula for each performance evaluation metric 83

B.1 Result report for LR classifier and stylometric features 103
B.2 Result report for RF classifier and stylometric features 104

B.3 Result report for SVM classifier and stylometric features 104
B.4 Result report for LR classifier and word embedding Features 105
B.5 Result report for RF classifier and word embedding Features 105
B.6 Result report for SVM classifier and word embedding features 106
B.7 Result report for LR classifier and combined features 106
B.8 Result report for RF classifier and combined features 107
B.9 Result report for SVMne classifier and combined features 107
B.10 Result report for LR classifier and document embedding features . . 108
B.11 Result report for RF classifier and document embedding features . . . 108
B.12 Result report for SVM classifier and document embedding features . 109

List of Tables

2.1 Machine learning algorithms that is used in clickbait literature 31

2.2 Features used in machine learning systems for clickbait detection . . . 33

3.1 Summary of the available corpora for clickbait by Potthast et al. (2018a) 37

3.2 The structure of a JSON file containing the extracted information of a tweet. 38

4.1 Stylometric parameters at different levels of analysis (Lagutina et al., 2019) 42

4.2 The list of information to be extracted for the analysis 44

4.3 Example of the headline and content of a listicle from Potthast et al. (2018b) 61

4.4 Example of the headline and content of a non-clickbait with a listicle headline from Potthast et al. (2018b) 62

4.5 The summary on the importance of each parameter in differentiating clickbait from non-clickbait based on the results of the analysis (++: quite important, +: important, -: not important) 76

5.1 List of features with the methods of engineering 80

5.2 Performance of different classifiers with respect to features. 84

5.3 Training time for each classifier with respect to features. 85

Acronyms

ANN Artificial Neural Network. 23–25

CBOW Continuous Bag of Words. 25, 26, 28

CNN Convolutional Neural Network. 24, 28

DT Decision Trees. 31

GB Gradient Boosting. 31

GloVe Global Vectors for Word Representation. 25, 27, 28

GRU Gated recurrent unit. 25

LogR Logistic Regression. 30

LR Linear Regression. 30, 84, 85, 87–89

LSTM Long Short-term Memory Network. 25, 27, 28

NB Naive Bayes. 30

NLP Natural Language Processing. 23, 25, 41, 42

RF Random Forests. 31, 84, 85, 87–89

RNN Recurrent Neural Network. 24–26, 86

SVM Support Vector Machines. 31, 34, 84, 85, 88, 89

Chapter 1

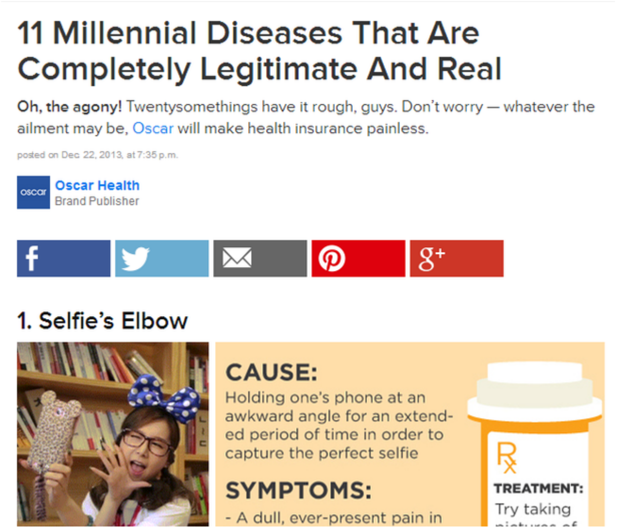
Introduction

1.1 Research Problem

1.1.1 Research motivation

In the modern times, people rely mainly on online media for information and entertainment. As such, news publishers are forced to move from print to digitized format. In addition, the Internet has created a more competitive environment which allows news outlets to freely release their contents and readers to freely access the materials. Therefore, news publishers can no longer depend totally on sales or subscriptions to survive but are required to adapt a new economic model called pay-per-click in which news publishers are paid by advertisers when an advertisement is clicked (Chen et al., 2015a; Beleslin et al., 2017). This new model poses a challenge to news creators to find new strategies to gain readership. Therefore, they developed clickbait as an interim solution to the problem above. However, this model poses some major problems as it is associated with negativity.

According to Chakraborty et al. (2016), the mechanism of clickbait is established essentially by arousing curiosity among readers. Loewenstein (1994) suggests that curiosity involves "intrinsically motivated desire for specific information". The lack of some "specific information" creates what is called "an information gap" in which the mind is triggered to fill producing a feeling of deprivation. Unless one can obtain the missing information, this negative feeling will not be reduced or eliminated, but intensified. Beleslin et al. (2017) and Chen et al. (2015a) argue that this cognitive


11 Millennial Diseases That Are Completely Legitimate And Real

Oh, the agony! Twentysomethings have it rough, guys. Don't worry — whatever the ailment may be, Oscar will make health insurance painless.

posted on Dec. 22, 2013, at 7:35 p.m.

Oscar Health
Brand Publisher

1. Selfie's Elbow

CAUSE:
Holding one's phone at an awkward angle for an extended period of time in order to capture the perfect selfie

SYMPTOMS:
- A dull, ever-present pain in

TREATMENT:
Try taking

Figure 1.1: Example of clickbait from Biyani et al. (2016)

mechanism is exploited to create clickbait by providing insufficient information in the headlines which requires readers to delve into the remaining content that includes advertisements.

The growth of clickbaits in recent years has become one of the biggest problems on the Internet. First of all, clickbaits draw negative impacts on readers. With the functionality as a web traffic booster, clickbaits have been utilised for economic purposes rather than reporting real events or providing information. Therefore, readers usually do not get what they have expected even though they are promised with a worthwhile experience or an indispensable revelation because the content of a clickbait usually does not align with its headline: while the headlines are generally written with erotic vocabulary and exaggerated tones to achieve their goals, the contents are poorly written, misleading, unverified, shallow and uninformative Zheng et al. (2018). This leads to disappointment, frustration and a sense of betrayal and deception for the readers as they can not truly fill in the "gap" between what they want to know and what they really get from the articles Cao et al. (2017); Agrawal (2016); Beleslin et al. (2017). Chakraborty et al. (2016) also pointed out that clickbaits contribute to the reduction of human attention span as they encourage people to constantly click on new baiting articles without reading the in-depth news stories.

Secondly, it is the online news publishers who suffer the consequences of using clickbaits. Their revenue boost is only temporary, because the dramatic decline in the quality of readers' experience and the negative emotions created by clickbaits is driving customers away from online journalism. These publishers are risking losing not only their credibility and reputation but also loyal reader base who are willing to pay for subscription Rony et al. (2017); Cao et al. (2017); Beleslin et al. (2017). They also violate the general codes of ethics of journalism which clearly state the ultimate goal of journalism is to serve the truth, accuracy, and objectiveness Zhou (2017); Beleslin et al. (2017). However, due to the enormous economic profit, clickbait is still a common practice for a plenty of publishers, regardless of any ethical standards.

Lastly, clickbaits have clogged up the Internet interfering with the search and exchange of information (Biyani et al., 2016; Zhou, 2017). A search engine returns its search results in an ordered list based on click-through rate (CTR) which is a ratio showing how often people click on a hypertext link to an article when they see it. CTR can determine the ranking of a search result in the list: the higher the CTR a search result gets, the higher rank it takes, the easier it can be accessed by readers. Since clickbaits usually have very high CTR, a search engine using the CTR approach cannot differentiate them from genuine high-quality articles and allows clickbaits to show up in very high positions of its search result list Biyani et al. (2016). This is also the reason why clickbaits can spread easily on the Internet, especially on social media as there is no double-checking system for the quality of articles shared by people. With their ubiquity, they have become a powerful tool for advertising, and spreading fake news Chen et al. (2015a); Beleslin et al. (2017); Zheng et al. (2018)

Due to the repercussions caused by the practice of clickbaits, it is necessary to gain a better understanding of clickbaits and develop an automatic system to detect them.

1.1.2 What is a clickbait?

The term "clickbait" is not new. Recorded by the Oxford English Dictionary and Merriam-Webster Dictionary, the first known use of the word "clickbait" dates back to 1999. However, it was not until the 2010s, "clickbait" was recorded in the dictionaries, due to the explosion of the clickbait practice. Etymologically, the word "clickbait" is a compound word formed by "click" referring to "the action or an act of pressing (and releasing) one of the buttons on a mouse or similar device as a means of selecting a particular item or activating a program" and "bait" referring figuratively to "an enticement, allurement, temptation." Below are definitions of clickbait in three well-known dictionaries for the English language: Merriam-Webster Dictionary, Oxford Dictionary, and Cambridge Dictionary:

- "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" (Merriam-Webster[1])
- "Internet content whose main purpose is to encourage users to follow a link to a web page, esp. where that web page is considered to be of low quality or value." (Oxford[2])
- "Articles, photographs, etc. on the internet that are intended to attract attention and encourage people to click on links to particular websites" (Cambridge[3])

The three definitions of clickbait in the mentioned dictionaries share a similar structure which consists of two parts: the first part matches the term clickbait with its form (could be web content, article, headline, photograph, etc.), the second part narrows down its references by describing its function. This typical type of dictionary definition has been employed in a majority of academic literature on clickbait.

The idea on the function of clickbait which is to lure people to open or view a web page is shared and supported across almost all literature (Blom and Hansen, 2015; Chen et al., 2015a; Chakraborty et al., 2016; Potthast et al., 2016; Scacco and Muddiman, 2016; Wang and Wu, 2017; Zheng et al., 2017; Agrawal, 2016; Paulau-Sampio, 2016; Beleslin et al., 2017; Thomas, 2017; Wei and Wan, 2017; Rony et al., 2017; Cao et al., 2017; Anand et al., 2017; Zheng et al., 2018; Zhou, 2017; Bagosy et al., 2018; Zannettou et al., 2019; Li, 2019; Sisodia, 2019; Pujahari and Sisodia, 2019). Meanwhile, the form categorisation of clickbait varies from study to study. There are three main approaches to the categorisation:

[1] https://www.merriam-webster.com/dictionary/clickbait

[2] https://www-oed-com/view/Entry/37263110?redirectedFrom=clickbaiteid

[3] https://dictionary-cambridge-org/dictionary/english/clickbait

1. Clickbait refers to online or web content in general (Potthast et al., 2016; Rony et al., 2017; Chen et al., 2015a; Chen and Rubin, 2017; Indurthi and Oota, 2017; Lockwood, 2016; Munger et al., 2018; Kalveks, 2007). Online or web content can be defined very broadly by Morville and Rosenfeld (2006) as "the stuff that makes up your site" which can include documents, data, applications, e-services, images, audio and video files, personal Web pages, archived e-mail messages, and so on.

2. Clickbait refers to a headline or title or teaser messages of a journal article (Potthast et al., 2018b; Blom and Hansen, 2015; Beleslin et al., 2017; Wang and Wu, 2017; Agrawal, 2016; Wang and Wu, 2017; Thomas, 2017; Rony et al., 2017; Cao et al., 2017; Chakraborty et al., 2016; Anand et al., 2017; Paulau-Sampio, 2016; Wei and Wan, 2017; Zannettou et al., 2019; Li, 2019; Pujahari and Sisodia, 2019; Scacco and Muddiman, 2016; Bagosy et al., 2018; Potthast et al., 2018b),

3. Clickbait does not take any specific form, but has a function of tricking readers into viewing the content of an article by clicking an accompanying link (Elyashar et al., 2017; Wiegmann et al., 2018; Grigorev, 2017; Glenski et al., 2017; Kumar et al., 2018; Cable and Mottershead, 2018; Pandey et al., 2018; Sisodia, 2019).

The problem of regarding clickbait as web content is vagueness as it does not provide a detailed description of clickbait concerning how clickbait is constructed, or what makes people click. Meanwhile regarding a clickbait as a headline of an article is too narrow since it cannot justify the negative connotation of clickbaits. The study on the reaction of readers on three different types of headlines by Scacco and Muddiman (2016) shows that even though clickbait headlines create more uncertainty in comparing to traditional headlines, they do not lead to more negative feelings by themselves, but need to be accompanied with unappealing content. We argue that the negativity does not derive from the headline constructing strategy but from the fact that the content does not provide the information that people look for. If the content is able to provide the missing information to fulfil readers' curiosity, then result of creating information gaps should be readers' satisfaction, not unpleasantness.

Papadopoulou et al. (2017) and Adelson et al. (2017) both regard clickbait in the relation with social media suggesting clickbait is a form of a post on social networking platforms such as Facebook or Twitter. However, this definition limits the existence of clickbait only on social networks and disregards the ability of clickbait existing independently from social media as a news article, or a picture on a website. Moreover, a characteristic of social media is user-generated content in the form of short massages for quick interaction between content creators and audiences via sharing and commenting Burke et al. (2011); Obar and Wildman (2015). This type of interaction does not generate "click" which is necessary for the money making model of clickbait. In this study, we are in favour of the idea from Thomas (2017) that social media just act as a host spreading clickbait on the Internet. The reason for the frequent appearance of clickbait on social networking sites is that each user can have a network of connections with other users and once a clickbait is shared by a user, it can reach the whole network of that user. The social media's power of sharing helps clickbait expose to a large number of readers in a very short amount of time.

So what exactly clickbait is? There have been many attempts to answer this question but yet there is no satisfactory answer provided. We must admit that clickbait is a very broad and elusive concept, and cannot be covered completely in one study. However, not having a unified definition of clickbait make it hard even for humans to determine how to identify clickbait out of non-clickbait, not to mention for machines as they are not able to perceive clickbait as a concept using its functions, but rely on other literal pointers. Therefore, we would like to narrow down the concept of clickbait to be pertinent to the scope of this study. However, the conclusion of what a clickbait is needs to come with empirical evidence. In Chapter 4, we explore the characteristics of clickbait using linguistic analysis on two clickbait corpora: one from Chakraborty et al. (2016) containing only clickbait headlines, and one from Potthast et al. (2018b) containing both clickbait headline and content. After that, the concept of clickbait is discussed in the 4.3 based on the findings of the analysis. For now, we acknowledge the view on the function of clickbait that it is for the purpose of tricking or manipulating. Furthermore, this study only deals with text-based data, visual content such as videos or images are excluded as this study lies within the field of computational linguistics.

1.2 Research Question and Objectives

With the motivation as stated in 1.1.1, this thesis sets out to build an automatic clickbait system using the characteristics of clickbait as features. This general objective can be broken down into three specific ones. The first is to define and characterise clickbaits by analysing and comparing the linguistic patterns of both clickbaits and non-clickbaits. It is important for be able to thoroughly understand the nature of clickbaits in order differentiate them from serious news. To achieve this objective, in Section 4.2, we carry out a linguistic analysis of two clickbait datasets from Chakraborty et al. (2016) and Potthast et al. (2018b). The second, which is the focus of Section 4.3 and 5.1.1, is to select the characteristics of clickbait that are possible to become features for the detection task and transform these characteristics into computational features for machine learning system. The last objective is to apply the features into machine learning systems and evaluate the informativeness of each type of features. This objective is attained in Section 5.1.2 and 5.2.

The three specific objectives of this thesis are achieved in order to give appropriate answers to the three research questions below:

1. What is a clickbait? What are the linguistic characteristics that differentiate clickbait and non-clickbait articles?

2. Among these characteristics of clickbait, which can potentially become features for automatic system and how can they be represented into computational features for machine learning?

3. Which among these features is the most informative for the clickbait detection task using machine learning method?

1.3 Outline of the Thesis

This thesis includes six main chapters. Chapter 1, Introduction, introduces the problem of clickbaits and why it is necessary to understand the characteristics of clickbaits and develop an automatic clickbait system. Chapter 2, Literature Review, gives an overview of the studies have been done on clickbaits in order to gain an overview about what a clickbait is, what the characteristics of a clickbait, what systems were built before for the clickbait detection task, and discusses what has not been done. Chapter 3, Data, describes the details of the two datasets from Chakraborty et al. (2016) and Potthast et al. (2018b) and the reason for utilising these two corpora in this study. Chapter 4, Linguistics Analysis of the Datasets, addresses the methods and the results of the linguistic analysis on the two corpora, and draws the conclusions about the definition and characteristics of clickbaits. Chapter 5, Automatic Detection Systems, presents the experiment set-up for machine learning systems including feature selection and engineering, and system design and optimisation. The performance evaluation for each machine learning algorithm on each feature are also discussed in Chapter 5. Chapter 6, Conclusion, summaries the main research findings, discuss the limitations of research, and provides some suggestions for further study on the subject.

Chapter 2

Literature Review

2.1 Qualitative analyses on the characteristics of clickbait

Clickbait is a central concept of this study, therefore, it is important to have an understanding of what the characteristics of clickbait are in order to be able to differentiate clickbait from non-clickbait. Only a few qualitative studies have been done to profile the characteristics of clickbait to gain a better understanding of clickbait. The earliest study can be taken into account is by Dor (2003) on the characteristics of tabloid news' headlines since it is believed by many that clickbait originates from tabloid journalism (Blom and Hansen, 2015; Chen et al., 2015a; Paulau-Sampio, 2016; Chakraborty et al., 2016; Lockwood, 2016; Beleslin et al., 2017; Rony et al., 2017; Cao et al., 2017). In the study, Dor (2003) characterises tabloid or clickbait headlines by comparing them with appropriate headlines. In journalism, headlines play a crucial role in news communication. The purpose of a headline is not only to summarise the main idea of an article, but also aid comprehension and constrain interpretation of the content by activating the reader's relevant background knowledge or contextual information of the topic, and preparing the reader of what to expect (Ecker et al., 2014). Moreover, headlines serve to draw attention and maximize interest of readers toward the article Dor (2003). He first carries out an overall analysis of new headlines and suggests some properties of an 'appropriate headline' including brevity, unambiguity, attractiveness, freshness, and providing relevant interpretation context for readers. Then, he pointed out that tabloid news' headlines possesses the same properties as quality news headlines, however, one of them is taken to its logical extreme: tabloid headlines are deficient in information, but they carry more contextual effects than information-rich headlines, and these effects can revoke vivid images and emotions in readers' mind but make it harder for readers to appropriately construct the right context of the story.

Lai and Farbrot (2014) field a question of "what makes people click?" by studying the effect of question headlines on readership. They set up two experiments, in each of which they record the number of "clicks" on different types of news headlines. In the first experiment, the types of headline include a declaration ("The hunt for status in the advertising business"), a question without self referencing cues ("Why are advertisers so obsessed with winning prices?") and question headline with self-referencing cues ("Is your boss intoxicated by power?"). In the second one, the types of headline include a non-question ("For sale: Black iPhone4 16GB"), a question headline without self-referencing cues ("Anyone who needs a new iPhone4?"), a question headline with self-referencing cues ("Is this your new iPhone4?") and a rhetorical question headline without self-referencing cues ("Is this your new iPhone4?"). The results from both experiments show that question headlines generate significantly more readership than declarative headlines, while according to the second experiment, question headlines with self referencing cues are particularly effective and generate higher readership than question headlines without self-referencing cues and without rhetorical questions.

Vijgen et al. (2014) studies a special type of clickbait called "listicles" which is an article that compiles a list of things. The author analyses the headlines and content of 720 "listicles" published on BuzzFeed in two weeks of January 2014 and comes to the conclusion that the content of these articles is usually easy to read according to the Gunning fog index which is a readability test for English writing. In addition, the titles, playing the role of teaser messages, are usually constructed with a homogeneous structure which contains a cardinal number—the number of items listed. These titles also contain nouns and adjectives with extreme sentiments conveying authority and sensationalism.

Blom and Hansen (2015) confirm their hypothesis that forward-reference is used as a strategy for clickbaiting by conducting an in-depth analysis of 2000 random headlines from a Danish news website. Forward-reference is a linguistic phenomenon of making reference to forthcoming parts in the discourse, or through the use of unresolved pronouns. The study identifies two common forms of forward-references: deixis and cataphora that occur mostly in commercial, ad-funded, and tabloid news websites. Both deixis and cataphora are words or phrases that refer to the forthcoming part, however, the scope of cataphora is limited within utterance level while the scope of deixis is the whole discourse. It also reports that forward-references in Danish headlines are prototypically expressed using general nouns with implicit discourse deictic reference, adverbs, ellipsis of obligatory arguments, or imperatives with implicit discourse deictic reference, more frequently than interrogatives referring to an answer given in the full text, demonstrative pronouns, personal pronouns, or definite articles.

Paulau-Sampio (2016) researches the clickbaiting strategies of 151 articles published in June 2015 in four online sections of the Spanish newspaper El Pais. In terms of topics, the research reveals a commitment to soft topics or anecdotal themes while serious news topics are absent from these so-called new articles. In terms of headlines, the author also uncovers three main types of writing formulas and the rhetorical linguistic patterns used for capturing the curiosity than as informative elements. The first is to use spoken language discourse markers and interaction including questions, exclamations, parentheses/brackets, inverted commas, rhetorical questions, vocatives, cataphoric elements and the presence of deictics. The second is the use of vocabulary and word games including proverbs, set phrases, idioms, colloquial or informal language, generic or buzz words, and intensifiers. The last concerns the use of simple sentence structures and noun phrases. In terms of content development, the study shows that 39% of articles are structured in a form of a list, 14% has fragmented structure, 42% are organized into paragraphs, following a traditional format with expository-narrative development, and 5% follows the conventional model of question/answer.

In general, the focus of these researches above is only to characterise clickbait by linguistic features of the headlines. What is still missing from the analysis of clickbait is the analysis of the content which might be a good indicator to differentiate clickbait and non-clickbait (Cao et al., 2017; Cable and Mottershead, 2018; Biyani et al., 2016; Rony et al., 2017). Therefore, our analysis in Chapter 4 applies distant and close reading on both clickbait headlines and clickbait contents or text bodies. The reading is to find the relevant answers to the first and the first part of the second research question.

2.2 Automatic clickbait detecting system

Understanding the need for automatically detecting clickbait, researchers in computational linguistics and NLP have made several attempts to develop such systems. We provide a systematic review of what has been done on the task.

2.2.1 Neural Network

In the last decade, neural networks have emerged as powerful computation models, yielding state-of-the-art results for many different tasks in the field of NLP. The development of neural networks was first proposed in 1944 by Warren McCullough and Walter Pitts who were inspired by the work of computational neuroscientists in modeling the human nervous system (Henderson, 2010; Rojas, 2013). An Artificial Neural Network (ANN) can be regarded as "a man-made device that emulates the physical structure and dynamics of the biological brain", according to Moisl et al. (2000).

As widespread experimented, various neural network architectures have been developed. However, neural network implementations still share some commonalities. Similar to the structure of the brain, an ANN consists of numerous interconnected processing units called artificial neurons or nodes (Heaton, 2015; Moisl et al., 2000; Gurney, 1997; Aggarwal, 2018). These neurons are arranged into a series of layers, and each individual neuron of a layer might be connected to several or all nodes on the alongside layers. The connections between the two neurons are represented by a number called a weight which also determines the influence between one unit on another. The general activity of a neuron is to receive information that has real numeric value from different sources, multiplies each incoming value by the corresponding weight, then adds these multiplications and passes this sum to an activation function to produce an output value (Heaton, 2015; Rojas, 2013). However, since an ANN has different types of layers that have different functions, neurons on different layers also perform slightly differently while all neurons on the same layers are exactly the same characters (Heaton, 2015). The first and fundamental layer of an ANN is an input layer designed to receive information from the raw data which the network will attempt to learn about. Each input neuron represents a single dimension of the data and the data is represented with a vector that contains all dimensions (Taylor, 2017). The output layer, staying at the end of the network, receives the information and gives out its responses to the information it receives. In between the input layer and the output layer, there can be one or many hidden layers where information is passed through. This means that the hidden neurons only receive input from other neurons and only output to other neurons. They are not directly connected to the data or produce the final output.

There are two ways for an ANN to learn (Taylor, 2017; Moisl et al., 2000). The first one is called feedforward as the information flows only one way from the input to the output layer. This method of learning requires a certain threshold value for each neuron where all the weighted inputs are summed up and compared to the threshold. If the sum exceeds the threshold value, the output of a neuron is a "1" (or else a "0"), and the neurons which are connected to that one are triggered and pass on the information. A Convolutional Neural Network (CNN) is a typical type of feedforward network. The second way is by a feedback process called backpropagation, as the information can flow both the input layer through hidden layers to output layers and vice versa. This learning method involves comparing the output an ANN produces with the correct output it was supposed to produce and using the difference between them as feedback to fine-tune the weights of the connections between the neurons in the network to minimize the neural network's margin of error. The feedback process can be carried out several times however if the result is not what expected, the system gets punishment which allows it not to repeat what it has already tried before. The training of an ANN only completes when expected results are achieved. Feedback networks can be more powerful and complex than feedforward networks. The most popular feedback network is Recurrent Neural Network (RNN) which includes a diversity of architectures such as Long Short-term Memory Network (LSTM) RNN, bi-directional RNN, or Gated recurrent unit (GRU) RNN.

In the field of NLP, the data needed to be learned by a computing system is language which is represented in the form of a text, a collection of letters. However, the input of an ANN must be a multidimensional vector, each dimension of which is represented by a real numeric value. Therefore, it requires a mapping between the language data and the vector representation. A set of techniques called word embedding is the most popular to model language and learning features in NLP thanks to the power of capturing semantic relations while representing discrete variables as continuous low-dimensional vectors Goldberg (2017); Ganegedara (2018); Biswas et al. (2019); Eisenstein (2018); Deng and Liu (2018); Alekseev and Nikolenko (2017). These techniques are developed based on the distributional hypothesis stating that the meaning of a word can be represented by the set of contexts constructed by its neighboring words in which the word occurs; therefore, words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1935; Sahlgren, 2008). The success of word embeddings is by the generalization ability and maintaining the contextual similarity (Torfi et al., 2020; Naili et al., 2017; Levy and Goldberg, 2014; Torfi et al., 2020). There have been many algorithms to train word embeddings from a corpus, however, the two most popular for the automatically detecting clickbait task are Word2Vec and the Global Vectors for Word Representation (GloVe).

Word2Vec is a predictive model learning word embeddings from unstructured data using a shallow neural network (Mikolov et al., 2013b,a). Two distinct variants of Word2vec are CBOW and Skip-Gram. The CBOW model predicts a current word based on its contexts (Mikolov et al., 2013b,a; Goldberg, 2017). The algorithm will generate a set of contexts in which each context consists of n words and then tries to match each word with the set using conditional probability of the word appearance given the contexts. The output is compared with the expected word to correct its representation based on the back propagation of the error gradient (Goldberg, 2017). Hence, CBOW is a bag-of-words model because the order or sequence of context words is not considered when averaged (Eisenstein, 2018). The Skip-Gram model is the opposite of the CBOW one. It seeks the prediction of the context given a word instead of the prediction of a word given its context like CBOW. The algorithm generates word pairs from the given word and context words, then calculates the probability of the appearance of context words given the target word (Eisenstein, 2018). The final step of Skip-Gram is the comparison between its output and each word of the context to correct its representation based on the back propagation of the error gradient Naili et al. (2017).

The state of the art for the task is achieved by Anand et al. (2017) utilising pre-trained 300 dimension CBOW embeddings concatenated with word embeddings trained with a three-layer using character embeddings as input. The neural network architecture is a bi-directional RNN with one forward and one backward layer as the authors argue that RNN is able to capture contextual information outside individual or fixed sized window of words. This system has a high result of 0.98 F1 score, which is the only system that achieve such success in the classification. Rony et al. (2017) use the Skip-gram model to transform a large corpus of 1.67 million Facebook posts created by 153 US based media organizations into distributed word embeddings The embeddings are later used to map sentences to a vector space over which a softmax function is applied as a classifier. Best performing model achieves 98% accuracy on a labeled dataset. Agrawal (2016) proves the usefulness of using for detecting clickbait using a five-layer with word embeddings learnt from a corpus of article headlines posted on Reddit, Facebook and Twitter, and learnt from a pre-trained Word2Vec model from Google . Zheng et al. (2018) propose a system applying word embedding features learnt from a corpus containing 14,922 headlines taken from four famous Chinese news websites (Tencent, 163, Sohu, Sina), well-known blogs, and Wechat official accounts. After having preprocessed with segmentation, stop-word filtering, and part-of-speech filtering, the headlines are transformed into word embeddings using both Skip-gram and CBOW. These embeddings later are connected to type-related word embeddings and input to a . The authors conclude that their system outperforms 5 other baselines: lexical similarity between headline and content, NB with N-grams, feature-based machine learning systems (Biyani et al., 2016), CBOW word embedding model (Joulin et al., 2016), and word embeddings using (Agrawal, 2016). Thomas (2017) participates in the Clickbait Challenge with a fusion of neural networks called Whitebait achieving a result of 0.564 F1 score. The author uses all text fields available including posttext, target-title, target-paragraphs, target-description, as well as the publication time of the tweet for each training example. The texts are tokenized, normalised then converted to word embeddings using Skip-gram model and input to a LSTM that is randomly initialised. Another neural network with one hidden layer is trained for the publication time limited to one-hour ranges and then converted into one-hot encodings and forwarded to an internal embedding layer. Finally the individually trained networks are fused by concatenating the last dense layer of the individual networks.

GloVe is another word embedding approach developed by Pennington et al. (2014) that is claimed to outperform other models on word analogy, word similarity, and named entity recognition tasks. It is a new global log-bilinear regression model for the unsupervised learning of word representations that employs both global matrix factorization and local context window methods. The reason for the combination of the two techniques is that each of them suffer from distinct disadvantages (Pennington et al., 2014). Matrix factorization methods rely solely on co-occurrence statistics to construct contexts and similarity. Therefore, the most frequent words which usually are stop-words such as “the” or “a” have a disproportionate effect on the measure of similarity between two words in the corpus despite conveying fairly little about their semantic relatedness. In addition, these methods are unable to maintain linear relation between vectors of similar words in the vector space. Meanwhile, shallow window-based methods cannot utilize the statistics of the corpus since they are trained on separate local context windows instead of on global co-occurrence counts. In order to overcome the drawbacks from the two approaches, the GloVe algorithm constructs a global matrix of word-word co-occurrence counts from the whole corpus. This means each entry of the matrix stores the number of times a word occurs with a context word. To observe the similarity between two words, the algorithm calculates the ratio of the co-occurrence probabilities of each word with the same context words to distinguish the relevant contexts from irrelevant ones and discriminate between the two relevant context. The representation vector of a word in a context is learned by reducing the difference between the dot product of the vectors of the target word and the context word and the logarithm of their number of co-occurrences.

There are two systems that use GloVe embeddings to construct the feature representation of their datasets for the task. The first one is Pineapplefish by Glenski et al. (2017) which relies on linguistically-infused LSTM and CNN. Each model consists of two sub-networks: a text sequence sub-network initialized with a word embeddings layer using pre-trained 200-dimensional GloVe embeddings trained on tweets and the text of each article, and a vector representation sub-network that learn from linguistic cues and/or image vector representations. The output of the two sub-networks are later concatenated to form data representations allowing the models to learn the strength of clickbait content from the text present not only in the tweets themselves but also the linked articles, as well as images present in the tweets. The second system by Zhou (2017) is the best-performing system of the Clickbait Challenge 2017 achieving a result of 0.683 F1 score. It employs a token-level, self-attentive mechanism on the hidden states of bi-directional Gated Recurrent Units to measure the weights of tweet tokens in predicting the annotation distribution.The input to the network is headlines represented as a sequence of word embeddings that have been pre-trained on Wikipedia using GloVe.

Besides word embeddings, many researchers have experimented with different methods of data representation in search of the improvement for their systems. Gairola et al. (2017) join the Clickbait Challenge with Tuna, a bi-directional LSTM with an attention layer system using text embeddings and image embeddings as input. For the text representation, they combine a pre-trained 300 dimensional CBOW word embeddings with character level embeddings calculated through . For the image representation, they apply a pre-trained object detection network to the image data and take the embeddings from the convolutional layer.The model achieves an F1 score of 0.6537, beating the baseline. Kumar et al. (2018) develop an updated version of Tuna (Gairola et al., 2017). In their version, they utilise Doc2Vec model for the teaser text and the target description representations. In addition, they add other two Siamese neural networks on top of the LSTM, one of which calculates the similarity score between the teaser text and the description field of the linked article, and the other one calculates the similarity score between the teaser image and the target description. The study is conducted over a test corpus of 19538 social media posts, attaining an F1 score of 0.6537.

Even though applying the most advanced technology, clickbait detecting systems using neural networks do not seem to achieve remarkable results for the task, except for the system of Anand et al. (2017). It cannot be denied that neural network models have many advantages such as the ability to implicitly detect complex nonlinear relationships between independent and dependent variables, and the versatility in optimisation. However, these systems still suffer from critical shortcoming (Tu, 1996). Neural network methods are usually referred to as a "black box" as they have limited ability to explicitly identify possible causal relationships. This means that neural network systems do not provide any deeper insight into the nature of clickbait since it is not possible to point out how they are able to differentiate clickbait and non-clickbait.

In this study, we decide not to apply this technology since our goal is to have comprehension about clickbait, and with their disadvantage of being a "black box", neural networks cannot facilitate that goal.

2.2.2 Machine Learning

Parallel to the development of neural networks, machine learning is also widely applied in many different computing problems (Rogers and Girolami, 2016). Machine learning, also known as predictive analytics or statistical learning, is an interdisciplinary research field concerned with statistics, artificial intelligence, and computer science. It gives machines, specifically computers, the ability to learn without being explicitly programmed (Samuel, 1959). Traditionally, in order to carry out a task, computers need to be given instructions or commands for every step. With machine learning, the machine can do a task by itself though a learning process (Richert, 2013). This process is described by (Mitchell et al., 1997) as "a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". In a way, the learning process of computers resembles the learning process of humans. For example, when a child learns about a new object such as a chair, the parents will tell him/her what a chair is by associating the word "chair" with the object in real life. The learning starts when that child can observe the features of a chair which can be height, width or function that differentiate a chair with other objects such as a table. These features should be significant or regular among a collection of chairs. As the child has seen a chair and told that the object he/she has seen is a chair, the child gains some experience with chairs, and later, he/she can rely on that experience to identify a chair. Machines or computers also learn by using experience or "the past information" (Alpaydin, 2020; Faul, 2019; Mohri et al., 2018) which can be gained from the discovery of the significance or regularities of provided samples data (Bishop, 2006; Kelleher et al., 2015).

There are many different methods of machine learning, however, supervised methods are the most popular for the task of clickbait detection which is normally treated as a classification problem. The fundamental concept behind supervised machine learning is that each sample is associated (or labeled) with a target variable (a class or labels). The goal is for the computer to learn the mapping between the target variables and the samples that results in prediction of the target variables for unseen data (Kelleher et al., 2015; Mohri et al., 2018). Table 2.1 shows the most common machine learning algorithms that are used in previous literature.

Algorithms	Description	Literature
Naive Bayes	NB algorithm is based on the Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.	Potthast et al. (2016); Zheng et al. (2017)
Linear Regression	LR models the relationship between a dependent variable and one or more explanatory variables using a linear function. It draws a linear interpolation between data points and constructs a hyperplane such that the distance is minimized between the points and the hyperplane	Indurthi and Oota (2017); Wiegmann et al. (2018)
Logistic Regression	LogR models the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic/sigmoid function.	Potthast et al. (2016); Cao et al. (2017); Papadopoulou et al. (2017); Zheng et al. (2017)

Support Vector Machines	SVM algorithm find an "optimal" hyperplane which separates the training data into two classes. The separating hyperplane is selected by maximizing its margin which is the shorted distance between the hyperplane and any of the samples in which one can move the separating hyperplane without any misclassification.	Chakraborty et al. (2016); Grigorev (2017); Wei and Wan (2017); Zheng et al. (2017)
Decision Trees	A decision tree consists of a set of partially ordered set of tests, each sequence of tests defining one branch in the tree is terminated by a leaf. Each leaf is associated with a class. Decision tree algorithm make predictions based on the sequences of tests on the descriptive feature values of a sample. The result of each test indicating what should happen next: either the algorithm goes through another test, or a decision about the class label of the sample if a leaf has been reached.	Chakraborty et al. (2016); Biyani et al. (2016); Elyashar et al. (2017)
Random Forests	RF algorithm creates a "forest" which is an ensemble of decision trees. Each new data sample is passed through each tree in the forest and each tree give out a class prediction. The final result is the one that is most predicted.	Chakraborty et al. (2016); Cao et al. (2017); Elyashar et al. (2017); Zheng et al. (2017)
Gradient Boosting	GB is an ensemble of shallow and weak successive trees with each tree learning and improving on the previous one.	Biyani et al. (2016); Elyashar et al. (2017); Zheng et al. (2017)

Table 2.1: Machine learning algorithms that is used in clickbait literature

Whichever methods are applied, the success of a machine learning system lies in the feature extraction process. Therefore there have been many experiments with features for the task. The first publications exploring potential features that could be used for the task of automatic clickbait detection is by Chen et al. (2015a). They postulate three sets of features. The first one is linguistic patterns in the content examined at two different levels. At the lexical and semantic levels of analysis, features include stylometric features such as parts of speech, word length and subjective terms; demonstrative pronouns; adverbs; interrogatives; and imperatives. At the syntactic and pragmatic levels, forward-referencing the affective and suspenseful language, action words, overuse of numerals, and reverse narratives are suggested to be informative.This set of features is established based on the quantitative studies on the characteristics of headlines, yellow journalism or tabloids in general including Knobloch et al. (2004), Blom and Hansen (2015), Chen et al. (2015b), Lex et al. (2010), Lary et al. (2010). The other two sets of features are non-linguistic: the congruence between textual content and image; and user reading and commenting behavior.

Even though the proposal of Chen et al. (2015a) is not supported by quantitative analysis on a large scale data set, nor validated by applying a corresponding automatic clickbait system, it has set a foundation for other studies on clickbait using automatic detecting systems and statistical features. Most common features used in later literature are quite similar to what Chen et al. (2015a) suggested. These features can be grouped into two categories: linguistic and non-linguistic features. Table 2.2 shows the list of most common features used in different machine learning systems for clickbait detection.

Linguistic features	Stylometric features	number of tokens
		number of words
		average word length
		number of stopwords
		stopwords-to-words ratio
		easy words-to-words ratio
		has abbreviation
		number of dots
		starts with number
		punctuation patterns
		question marks
		number of noun phrases
		number of adverb phrases
		number of adjective phrases
		number of verb phrases
		question words

		number of characters
		character n-gram
	Semantic/lexical features	sentiments
		sentiment lexicon
		suspenseful language
		slangs
		similarity between post and target title
		word embeddings
		bag-of-words
		word n-gram
		word list
	Syntactic/grammar features	forward reference
		reverse narrative
		determiners
		possessive case
		part of speech
		syntactic N-grams
		adverbs
		interrogatives
		imperatives
		demonstrative pronouns
		superlatives
		past participle
		present participle
	Discourse features	topics
		readability
Non-linguistic features	Tweet features	hashtag
		user tagging
		emojis
	Image features	use image
	User behavior features	commenting
		sharing patterns

Table 2.2: Features used in machine learning systems for clickbait detection

Chakraborty et al. (2016) is the first to carry out a statistical analysis of the significant discrepancies between clickbait and non-clickbait posts over a number of features for an automatic clickbait detection system, including sentence structure features, word pattern features, clickbait language features, as well as N-gram features. The best performance is achieved with SVM at an accuracy rate of 93%, a precision of 0.95 and recall of 0.9.

Potthast et al. (2016) compile the first clickbait corpus including 767 clickbait from the top 20 most prolific publishers on Twitter from which the authors extract 215 features include basic text features of teaser message (headline) and the main content such as bag-of-word, polarity, the number of stop words; dictionary features using word lists, and common phrases; and Twitter meta information features such as attached images and videos, the number of retweets or hashtags, and timestamp of each post.

Biyani et al. (2016) develop new features to detect clickbait based on textual similarity between the headline and the content. In combination with content features such as sentiment scores, and n-grams as well as informality and forward reference features.

Cao et al. (2017) propose a machine learning system called Pike trained on a set of 175 linguistic features from the teaser message, two features related to the linked article, as well as three features that capture semantic relations between the teaser message and the linked article. The authors introduce some new features: the text similarity between post text and target title calculated by the number of overlapping words between the two files normalized by the length of the files; and the part-of-speech patterns which are number+noun phrase+verb, and number+noun phrase+that.The best system to perform is Random Forest Regression with selected 60 most important features.

Indurthi and Oota (2017) represent the teaser message of a tweet using hand-crafted features including the number of words, the number of stop words, the average length of the word, the presence of question form, the presence of numbers at the start of headline, the presence of continuous form of verbs, and the presence of superlative forms of adjectives; and along with pre-trained distributed word representations using 300 dimensions of the GloVe embeddings. These features are used to train a Linear Regression model which achieves an F1 score of 0.65.

Grigorev (2017) applies the standard Bag-of-Words approach to encode individual words, 2- and 3-grams into a spare vector space after The preprocessing which includes removing stopwords, removing duplicate sentences, removing HTML tags, replacing number with [n], and extracting word stems.

Elyashar et al. (2017) create a tweet representation that consists of three feature types: (1) image-related features that map the image in the post and the headline with the text, keywords, and images in the article, (2) linguistic features extracted from the headline and the linked article and (3) user behavior patterns. The best performance of 0.819 precision, and 0.966 recall is obtained by XGBoost with all features.

Papadopoulou et al. (2017) combine text features adapted from the previous studies including morphological features (character, capitalised or not,..), stylistic (slangs, number of hashtags,..), grammatical features (pos, number of verbs, nouns. . .), sentiment, and bag-of-word. Papadopoulou et al. (2017) also refer to Paulau-Sampio (2016) and calculate readability of the text body.

Wei and Wan (2017) use two sets of features: body-independent including 10 features suggested to be most informative by Chakraborty et al. (2016); and body-dependent features including informality, and sentiment of headlines and text bodies, informality gap, sentiment gap, and similarity between headlines and text bodies. The authors also extract textual entailment by search sentences in the body for dependency pairs occurred in the headline. The system achieves the best result of 0.753 F1 score.

Zheng et al. (2017) proposed a model that takes into account the user-behavior features to detect clickbait. Their model takes the outputs of traditional models as inputs, and recalculates the clickbait probability according to user behaviors. They conclude that user-behaviour features can significantly improve the performance of machine learning algorithms.

Wiegmann et al. (2018) attempt to improve the baseline for the Clickbait Challenge 2017 using a heuristic feature selection approach called leave-many-out feature selection. This process starts with the full set of features and then randomly removes one at a time until a subset with a minimum number of features is reached. For each removal, a leave-one-out error is recorded, and after so many times the process repeats, an average error of a feature is calculated and considered a score of its usefulness. The authors claim that choosing the top-scoring features can optimise the prediction.

Wang and Wu (2017) explore the existence of clickbait in the data science articles of WeChat official accounts - a social media. They train a regression model on top 200 keywords with highest frequency and the relation with page views. They find that the clickbait phenomenon in the data science articles of WeChat official accounts is not prominent.

In all, these studies provide some fundamental information about the landscape of machine learning application in clickbait detection task. For our study, since we use the Clickbait Challenge 2017 dataset (Potthast et al., 2018b) as the training data, we also use the baseline result from the challenge which belongs to Potthast et al. (2016) as our benchmark to evaluate our system. Moreover, we will follow the study of Chakraborty et al. (2016) as a guideline for our analysis, and a reference for our findings.

Chapter 3

Data

The key to this study is data. Quite a few datasets for clickbait have been built, however each of them has their own advantages and disadvantages. Table 3.1 provide a detailed description of the most popular corpora used for the task. Among these corpora, we choose two corpora as our data for the clickbait analysis and building machine learning system.

3.1 The Webis Clickbait Corpus 2017

The first subject of this study is the Webis Clickbait Corpus 2017 [1] by Potthast et al. (2018b) which is used in the Clickbait Challenge 2017 as train and test dataset. The corpus is collected from Twitter which is a social media platform where news pub-

[1] Available to public on https://zenodo.org/record/3346491

Publication	**Type of data**	**Size**
Agrawal (2016)	headlines	2388
Potthast et al. (2016)	headlines & contents	2992
Biyani et al. (2016)	headlines	4073
Chakraborty et al. (2016)	headlines	15000
Rony et al. (2017)	headlines	32000
Potthast et al. (2018b)	headlines & contents	38517

Table 3.1: Summary of the available corpora for clickbait by Potthast et al. (2018a)

lishers can distribute their content in the form of a tweet which is a short text that can be followed by a link to a news article and/or a picture. Therefore, the collected data includes not only the tweet texts but also the headline, text body, caption, description and keywords of the accompanied articles, and media attachments (images). All of the information above is provided in a JSON file structured as in table 3.2.

Objects	**Information**
postText	the content of the tweet
postMedia	the image that was posted alongside with the tweet (also made available by the organizers)
targetTitle	the title of the actual article
targetDescription and targetKeywords	the description and keywords from the meta tags of the article
targetCaptions	all captions in the article
targetParagraphs	the actual content of the article

Table 3.2: The structure of a JSON file containing the extracted information of a tweet.

The corpus is made up of 38.517 entries sampled from the top 27 most retweeted news publishers ranked by NewsWhip social media analytic service in the period from December 1, 2016, through April 30, 2017. The data acquired from each tweet was annotated regarding their "clickbaitiness" by five human evaluators on a four-point-graded scale: not click baiting (0.0), slightly click baiting (0.33), considerably click baiting (0.66), and heavily click baiting (1.0) and a two-point-graded scale: clickbait and no-clickbait. Then the mean judgments of the annotations was calculated and used to determine the final class of each data entry. The annotation was implemented with the crowdsourcing platform Amazon Mechanical Turk (AMT).

To annotate the clickbait corpus, Potthast et al. (2018b) provided a function description to aid the understanding of the clickbait concept to annotators, who were asked to determine the "clickbaitness" - how likely they are to be a clickbait - of a tweet and the article linked with it. This notion concerns with four properties: the emergence of clickbait from the optimization of teaser messages to maximize click-through, the intention of publishers using clickbait to lure people, the effect of clickbait on the readers which is manipulation, and reader perception regarding clickbait as a bait to click a link. The method of using these properties to identify clickbait depends greatly on the individual opinion of evaluators, which poses a question of the validity of the annotation. Meanwhile, Potthast et al. (2018b) report a fair agreement level with Fleiss'K of 0.36.

The reason for choosing this corpus is twofold. First of all, it is the largest clickbait corpus that is publicly available at the moment. Second, both the headline and the text body of a clickbait or non-clickbait article are included in the corpus, which is critical point for this study since our goal is to analyse the characteristics of clickbait from both headlines and contents to determine what a clickbait is.

3.2 Clickbait Headline Corpus

Beside the Webis Clickbait Corpus 2017, we use the second corpus which is the clickbait headline corpus from Chakraborty et al. (2016). The corpus Chakraborty et al. (2016) consists of the headlines of 16,000 articles published in September 2015 from 10 news publishers, five of which are well-known to practice clickbait. The authors collected the headlines and later asked six volunteers to label the headlines of these articles as either clickbait or non-clickbait. Each article is labeled by at least three volunteers. They were able to obtain an inter-annotator agreement with a Fleiss'K of 0.79. More than 7000 headlines were marked as clickbait. This corpus is chosen because of its size, as well as the quality of being publicly available. Even though the corpus of Rony et al. (2017) is much lager than the corpus of Chakraborty et al. (2016), it is not available for public use.

This corpus is used in section 4.2.1 as a point of reference for the linguistic analysis of clickbait headlines. It is also included in the experiment of automatic identification, since we would like to build more generic systems that learn from different datasets with different characteristics. We understand that if we train our system with two datasets that have broad differences, the system might not be able to pick up the significant, however, we are willing to trade off variance for bias, since we observe a quite low agreement level in the Webis Clickbait Corpus 2017, meaning that there can be many noises in this dataset.

Chapter 4

Linguistic Analysis of the Datasets

This chapter presents in details the linguistic analysis of clickbaits and non-clickbaits in the Webis Clickbait Corpus 2017, in reference to the analysis of Chakraborty et al. (2016)'s corpus. As mentioned in Section 2.1, Chakraborty et al. (2016) was the first to carry out a statistical analysis on a corpus and reports on the significance of linguistic characteristics of clickbait headlines in comparing to non-clickbait headlines, while there has not been any kind of analysis on the corpus of Potthast et al. (2018b) from the authors themselves or any other participants of the Webis Clickbait Corpus 2017 who use the corpus for their system. Therefore, the analysis is necessary for yielding insight into the differences between clickbait and non-clickbait in the corpus. In addition, some participants of the Clickbait Challenge 2017 including Elyashar et al. (2017); Papadopoulou et al. (2017); Indurthi and Oota (2017) developed their machine learning system based on the result of Chakraborty et al. (2016) using the corpus of Potthast et al. (2018b), but the testing results of these system are even below the baseline set by Potthast et al. (2016). In comparison, Wei and Wan (2017); Pujahari and Sisodia (2019); Pandey et al. (2018) are able to achieve fairly good results reapplying features from Chakraborty et al. (2016) analysis on Chakraborty corpus. This means that the findings of Chakraborty et al. (2016) is not relevant for 2017 corpus. As the result, the analysis of clickbait characteristics in the Webis Clickbait Corpus can able us to make some informed choice for features that are able to represent the data in both corpora.

For the ease of following the analysis, we refer to the corpus from Chakraborty et al. (2016) as Chakraborty corpus, and the corpus from Potthast et al. (2018b) as

Potthast corpus

4.1 Methods of Analysis

In this study, in order to characterise clickbait and differentiate it with formal non-clickbait, a mix of two different types of analysis, quantitative and qualitative, are applied.

The quantitative analysis is carried out using stylometry method concerning stylistic characteristics of texts that can be statistically quantified. Even though most popularly used in authorship attribution researches (Gómez-Adorno et al., 2018), stylometric analysis is applied in various NLP tasks including authorship profiling, style change detection, sentiment analysis and classification of written texts thanks to the ability to capture text complexity and multidimensionality (Lagutina et al., 2019).

For qualitative evaluation, a close-reading is performed to analyse and compare different examples from the two corpora. This method may not provide general descriptions of the corpora or the phenomenon. However, with the focus on and within each individual example, it can supply some explanation, clarification, justification or question for quantitative observation, while quantitative analysis highlights what may deserve further investigation (Jänicke et al., 2015).

Parameters of stylometry can be at many different levels of linguistic analysis. Table 4.1 systematically lists these parameters by each level. Chakraborty et al. (2016) have applied stylometric analysis on the clickbait corpus on word/token, syntax and semantic level. Our analysis follows Chakraborty et al. (2016) analysis scrutinizing the two corpora at the same level; however, there are some modifications to the parameters as we adapt the metrics of stylometry from Lagutina et al. (2019). Some parameters are omitted, for example, parameters at phonetic level since the subject of the study is textual.

In order to conduct the analysis, we build a processing system using Python which is an interpreted, high-level, general-purpose programming language, and Python libraries for text analysis including SpaCy, and TextBlob. SpaCy [1] comes

[1] https://spacy.io/

Level of analysis	**Parameters**
Character level	*character n-gram, the frequency of characters, lower- and upper-case letters, figures, and space*
Word/token level	*bag-of-word, word frequency, word length, average word length, contraction, word n-grams*
Text level	*topic, genres, sentiment*
Phonetic level	*intonation, melody, the number of syllables, vowels, and consonants, rhythm patterns*
Syntax level	*sentence length, average sentence length, punctuation mark frequency, functional word ratio, part-of-speech, part-of-speech n-gram, syntactic trees, sentence types, the complexity of sentence construction, ellipsis*
Semantic level	*expressions, synonyms, antonyms, named entities, topic words, slang, sentiment of words*

Table 4.1: Stylometric parameters at different levels of analysis (Lagutina et al., 2019)

with pretrained statistical models for work tokenization, lemmatization, syntax parsing, part-of-speech tagging, and named entity recognition as well as pretrained vectors for word embeddings utilising the state-of-the-art convolutional neural network models. It's commercial open-source software, released under the MIT license. TextBlob [2] is another library for processing textual data providing a simple API for diving into common NLP tasks. It is developed on top of NLTK which is another architecture for text analysis, however, TextBlob is much simpler.

SpaCy provides the main technology for the analysing system. The reason is that SpaCy can offer the fastest tools for all NLP fundamental tasks with great accuracy that is within 1% of the best available, according to Choi et al. (2015). We integrate TextBlob into our architecture as a support.

Before we carry out the analysis, all data is pre-processed with tokenisation, dependency parsing, part-of-speech tagging and named-entity recognition. From this preliminary processed data, we extract information for the analysis which is presented in Table 4.2

[2] https://textblob.readthedocs.io/en/dev/

Extracted information	**Technical method**
Tokens, number of tokens, average token length	*Use SpaCy tokeniser to break down each text into tokens, then count the number of tokens in a sentence and the average token length*
Punctuation, contraction	*Get any tokens that is punctuation or contraction. Count the number of contraction in each sentence and count the occurrence of each type of punctuation*
Part-of-speech (pos), POS tri-grams and POS four-grams	*Use SpaCy POS-tagger and get the tri-grams and four-grams*
Dependencies, the longest dependency path	*Use SpaCy dependency parser, calculate the path from each dependency to the ROOT of the sentence*
Subjects	*From dependency structure, extract the subject of each sentence.*
Named entities and named entity labels	*Use SpaCy named entity recogniser to extract the named entities and their categories (labels) in the texts*
Sentiment scores	*Use TextBlob to calculate the sentiment score of each text*
Is a question	*Check if a text is a question by checking if the part-of-speech and the dependency of the first token indicate wh-question or auxiliaries*
Start with number	*Check if a text starts with a number that is not a year or an age based on the part-of-speech and the dependency of the first token.*
Use passive structure	*Check if there are any passive structures used in a text by checking the present of any dependencies marked with "pass"*

Use superlative	*Check the present of superlative structures in a text by checking if "JJS" or "RBS" is in the part-of-speech list*
Use conditional	*Check the present of conditional structures in a text by checking if "if" or "unless" is in the token list*

Table 4.2: The list of information to be extracted for the analysis

For the analysis of the content, we apply some similar methods of extracting information from headlines, especially for calculate the number of tokens, the average length of tokens, and the sentiment scores. For sentence features, we use the sentencizer from SpaCy which detects sentence boundary.

In order to calculate the similarity between headlines and contents, we depend on Gensim [3] which is a Python library specifically for unsupervised semantic modelling from plain text. This library allows building a vector space from the vocabulary of a corpus. Words are transformed into vectors using Latent semantic analysis (LSA) [4] which is a technique of representing documents based on the frequency of the terms in a set of documents and the set itself. The similarity of two documents is determined by the cosine similarity of their representation vectors.

4.2 Linguistic Analysis

4.2.1 Clickbait Headlines vs Non-Clickbait Headlines

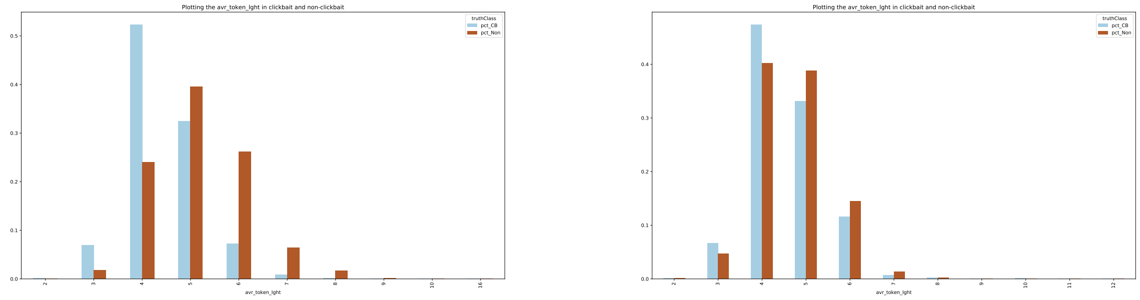
In the analysis of clickbait and non-clickbait headlines, we are able to make some comparisons between the two corpora on word/token level, syntax level and semantic level.

4.2.1.1 Word/token-level

4.2.1.1.1 The average length of words/tokens On word/token-level of analysis, Chakraborty et al. (2016) first examine the length of words in clickbait and

[3] https://radimrehurek.com/gensim/index.html

[4] https://en.wikipedia.org/wiki/Latent$_{s}emantic_{a}nalysisLatent_{s}emantic_{i}ndexing$



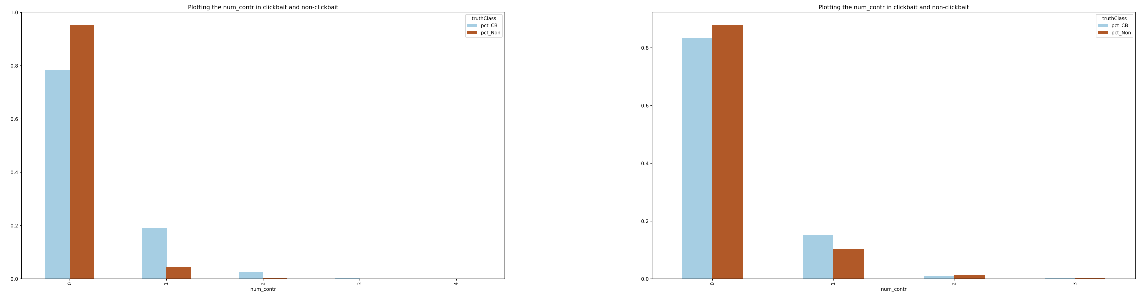
(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.1: The distribution of different average token lengths

non-clickbait headlines. The conclusion is that the average token length in clickbait headlines is shorter than in non-clickbait ones. Figure 4.1a shows the distribution of average token length of each headline in Chakraborty corpus. It shows a majority of clickbait headlines have an average token length around 4 or 5 characters, while tokens in non-clickbait headlines usually have an average length around 4 to 6 characters.

Meanwhile, Figure 4.1b reports the distribution of different average token lengths in Potthast corpus. As can be seen from the graph, both clickbait and non-clickbait headlines in the corpus have the average token length of 4 to 6. There are still more clickbait headlines that have an average token length less than 4 and there are more non-clickbait headlines that have an average token length more than 6, but only with a negligible percentage.



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.2: The distribution of different numbers of contraction in headlines

4.2.1.1.2 Contraction Chakraborty et al. (2016) explain that the difference between the average length of words/tokens in clickbait and non-clickbait headlines is due to the frequent use of shorter function words and word shortenings in clickbait headlines, while it is uncommon practice in non-clickbait headlines. Figure 4.2a compares the distribution of different numbers of contraction use in headline in Chakraborty corpus. More than 20% of clickbait headline use contractions and, while only less than 10% of non-clickbait headlines do, meaning that contraction is used about twice as more often in clickbait than non-clickbait headlines in this corpus.

The distribution of contractions in Potthast corpus is demonstrated in Figure 4.2b. Here, the percentage of non-clickbait headlines that use contractions are quite similar to the percentage of clickbait one. Even though there are more clickbait headlines that use one contraction, there are more non-clickbait headlines that use two contractions, and about the same percentage of clickbait and non-clickbait headlines use three contractions.

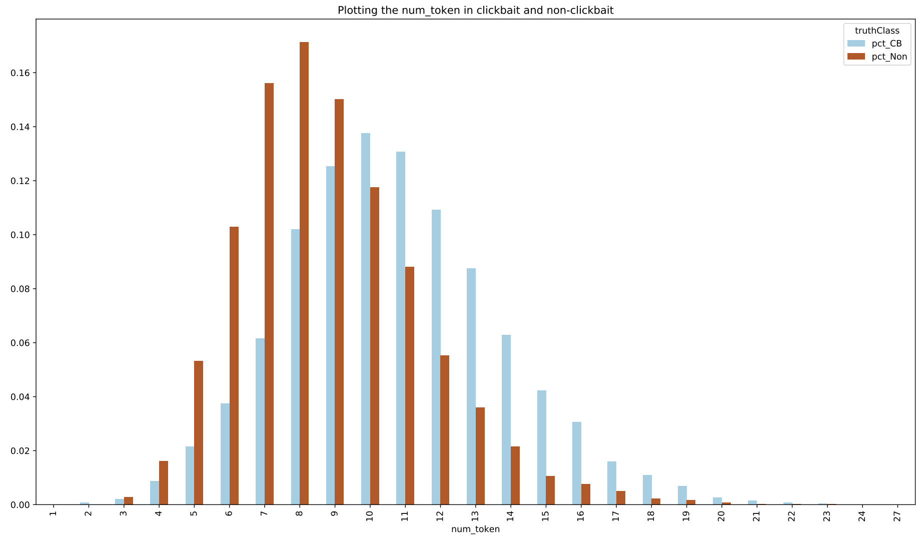


Figure 4.3: Plotting the number of tokens (Chakraborty et al., 2016)

In all, we barely observe any considerable differences between the distribution of token lengths and the frequency of contraction use between clickbaits and non-clickbaits in Potthast corpus. As the result, the findings on word/token level characteristics of clickbait in Chakraborty corpus cannot be confirmed to be universal. This means the word/token level characteristics can be uninformative feature for the automatic detection systems.

4.2.1.2 Syntax level

4.2.1.2.1 Sentence length Regarding the length of the headlines (or the number of tokens), Chakraborty et al. (2016) come to the conclusion that clickbait headlines are usually longer than the non-clickbait headlines.

Figure 4.3 shows the plotting of the number of tokens in the clickbait and non-clickbait headlines in Chakraborty corpus.The number of tokens in clickbait headlines is frequently in the range of 8 to 13, while non-clickbait 6 to 11. The average length of clickbait headlines is 10 while non-clickbait 8. However, there are more clickbait headlines that have more than 10 tokens and there are more non-clickbait headlines that have less than 8 tokens.

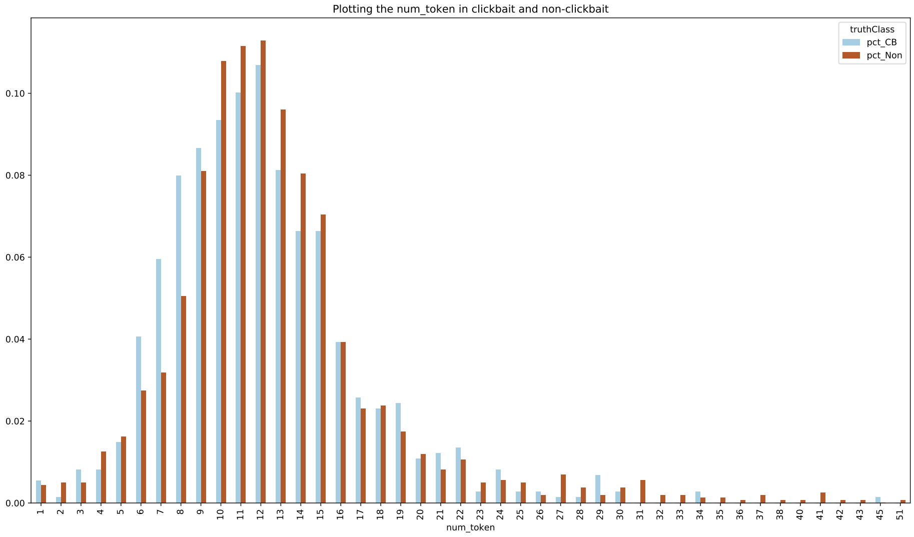


Figure 4.4: Plotting the number of tokens (Potthast et al., 2018b)

However, the conclusion of Chakraborty is not true for Potthast corpus. Figure 4.4 shows the number of tokens in clickbait and non-clickbait headlines in Potthast corpus. The graph shows a significant change in the length of headlines. The number of tokens in clickbait headlines is in the range of 6 to 15 while the number of tokens in non-clickbait headlines is in the range of 8 to 16. Both clickbait and non-clickbait headlines have an average length of 12. There are more clickbait headlines that is shorter than average, and there are more non-clickbait headlines that is longer than average. This means that clickbait headlines are getting shorter while the formal non-clickbait headlines are getting longer. In addition, as can be seen in 4.4, there are quite a few headlines that exceed the maximum length in Chakraborty corpus.

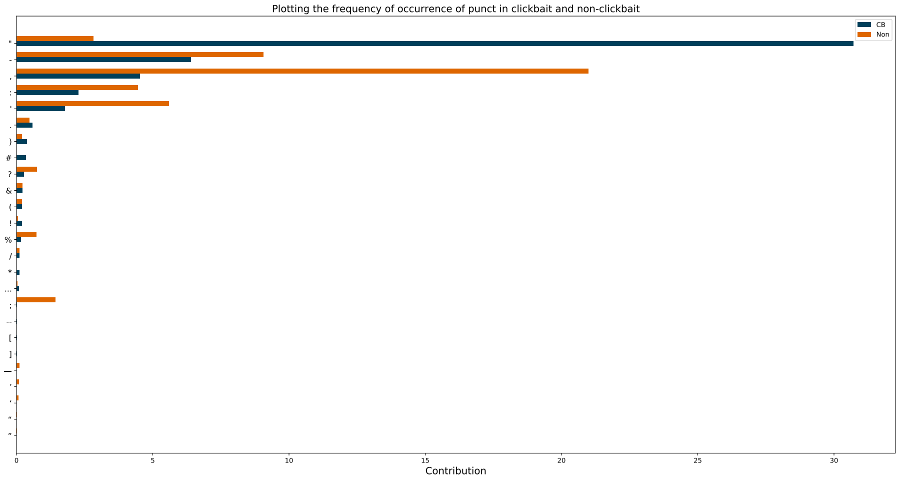


Figure 4.5: Plotting the frequency of punctuation (Chakraborty et al., 2016)

4.2.1.2.2 Punctuation Another aspect that Chakraborty et al. (2016) look into is punctuation.They discover some informal punctuation patterns such as *!?*, *...*, *!!!* or ***** that only appear in clickbait headlines. As we carry out a close reading on the corpus, we find that these patterns either appear at a very low frequency in clickbait headlines or actually appear in the non-clickbait headlines. For example, the pattern ***** only appears in one clickbait headline *"How Do You Spell Br*nd*n Fr***r's Name"*, and the pattern *...* also appears in non-clickbait headlines like *"ITV fined A$3000 for cruelty to rat on "I'm A Celebrity...Get Me Out Of Here!"*. As the result, we decide to plot the occurrence of every punctuation mark that appear in the corpus. The results are presented in Figure 4.5

Closer inspection of the table shows a considerable use of quotation mark in clickbait headlines, while comma and semi comma is more popular in non-cickbait ones. Question mark is actually used more in non-clickbait headlines in Chakraborty corpus, despite previous findings that clickbait headlines are frequently in the form of a question (Scacco and Muddiman, 2016).

The frequency of punctuation marks in Potthast corpus is shown in Figure 4.6. Here, the differences between the use of punctuation marks in clickbait and non-

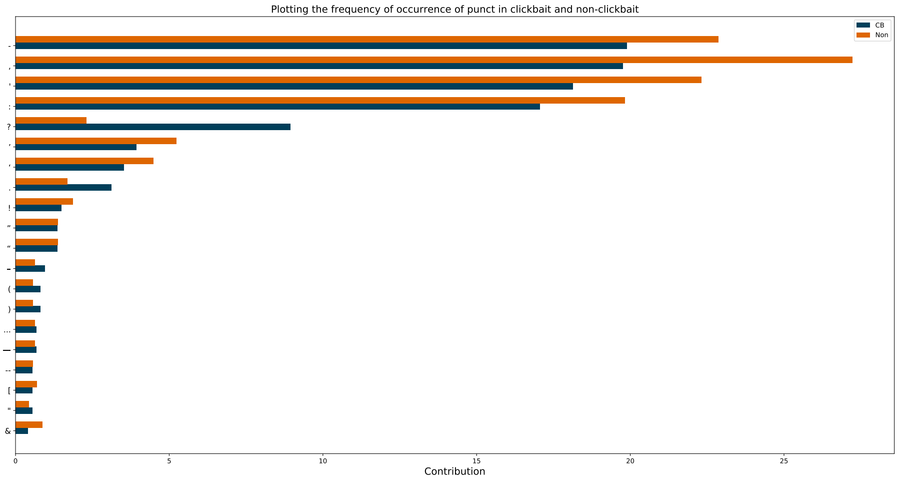


Figure 4.6: Plotting the frequency of punctuation (Potthast et al., 2018b)

clickbait headlines is less significant. However, it is apparent from the chart that there is a greater use of question marks in clickbait headlines than in non-clickbait headlines. Punctuation use is less informative in differentiating clickbait and non-clickbait in Potthast corpus than in Chakraborty corpus.

4.2.1.2.3 The maximum length of dependency path Chakraborty et al. (2016) calculate the length of dependency path which is the number of words separating the governing and the dependent words. An example from their study is *“A 22-year-old whose husband and baby were killed by a drunk driver has posted a gut-wrenching Facebook plea”*. The maximum length of the syntactic dependency which is 11 is calculated by counting the number of tokens in the adjective clause separating the subject *"22-year-old"* and the verb *"posted"*. They come to a conclusion that on average, clickbaits have longer dependencies than non-clickbaits; the main reason being the existence of more complex phrasal sentences as compared to non-clickbait headlines.

However, after examining the dependency structure, we decided not to follow the measure method of Chakraborty et al. (2016). Considering the dependency tree of the example from Chakraborty et al. (2016), we can see that "22-year-old", "has", "plea" are all direct dependency of the word "posted", even though the number of words between each dependency and its governor are different as shown in 4.7.

Therefore, we decide to calculate the dependency paths of a sentence based on

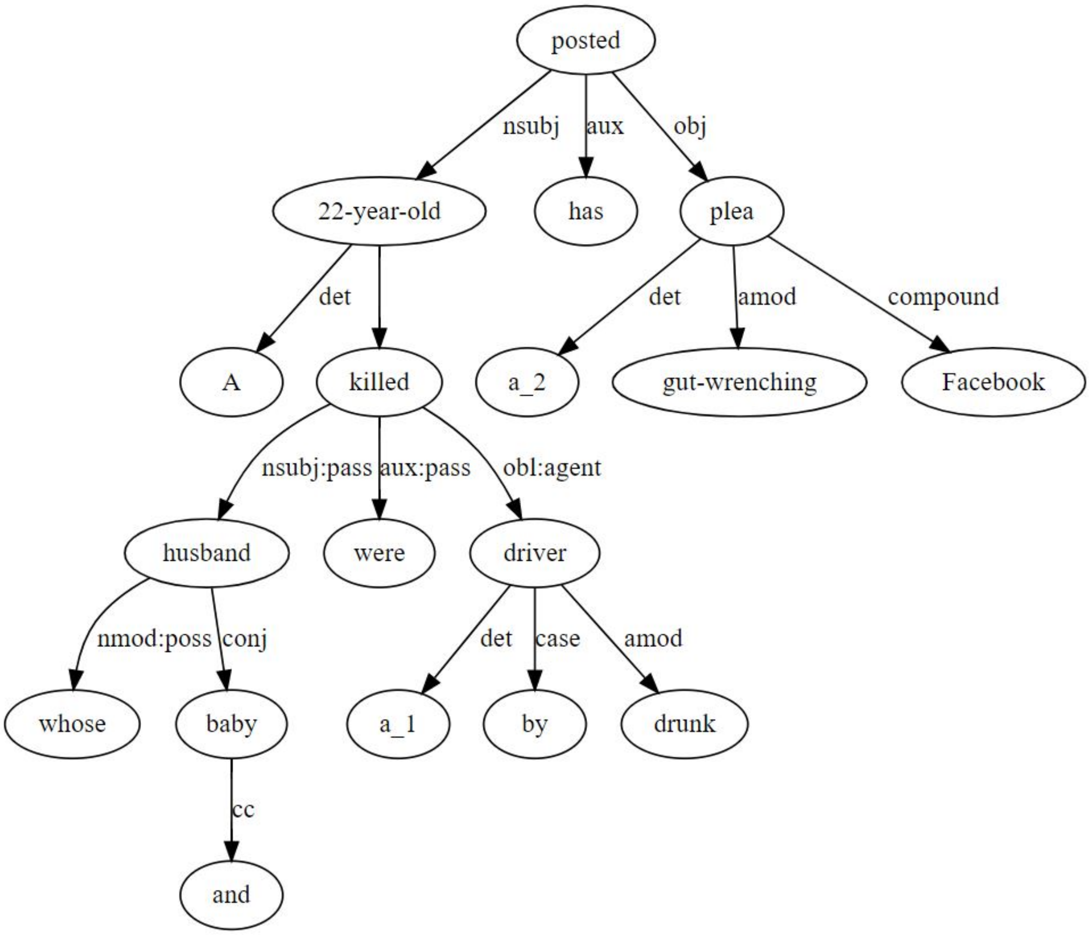


Figure 4.7: Dependency tree of the headline "A 22-Year-Old Whose Husband And Baby Were Killed By A Drunk Driver Has Posted A Gut-Wrenching Facebook Plea"

its syntax tree structure. A tree starts with the root since the root is the only independent element in a sentence. Then the tree is expanded down with each dependency as a leaf node and the dependency relation as the edge that connect the governor and the dependency. We calculate the dependency path of each node at the lowest level of the tree to the root. The longest dependency path for the headline *"A 22-Year-Old Whose Husband And Baby Were Killed By A Drunk Driver Has Posted A Gut-Wrenching Facebook Plea"* calculated using our method is 6, between *"and"* and *"posted"*, instead of 11.

We believe that our method is better at reflecting the complexity of the sentence structure than Chakraborty et al. (2016)'s as we are able to capture the depth of the dependency tree. Considering the two sentences in 4.8, if we apply Chakraborty et al. (2016)'s method, then the first sentence *"The woman who wore a red scarf is killed by a man."* has the maximum dependency length of 6 while the second one *"The woman is killed by a man who wore a red scarf."* maximum dependency length of 8. These numbers indicate that the second sentence has a more complex structure than the first one does despite the fact that they have the exact same main clause *"The woman is killed by a man"*, only with the relative clause *"who wore a red scarf"* in different positions. Nevertheless, our method gives the same result to the two sentences with the maximum length of 4 showing the similarity of the structure of the two sentences.

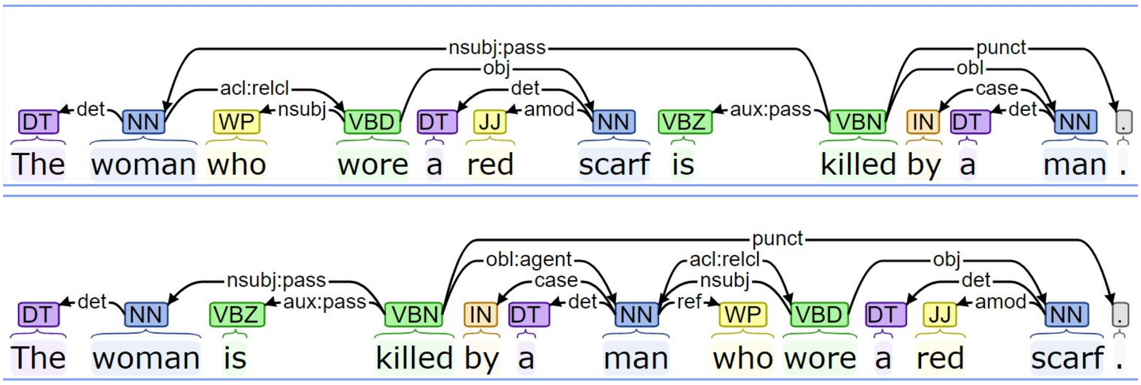
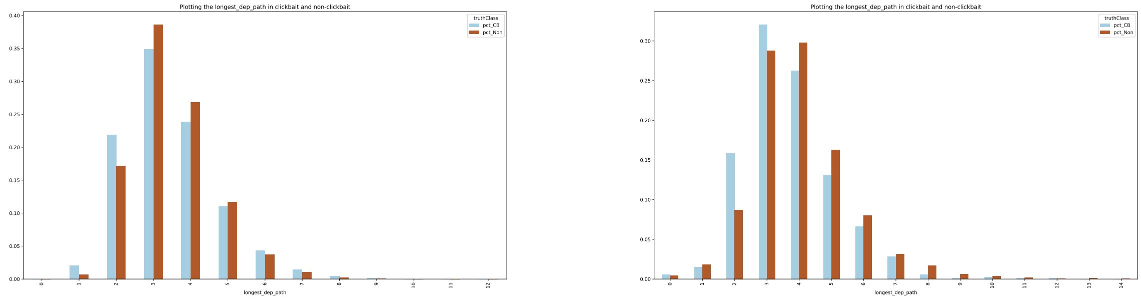

nsubj:pass
obj
punct
obl
acl:relcl
det
case
det
nsubj
amod
aux:pass
det
DT NN WP VBD DT JJ NN VBZ VBN IN DT NN .
The woman who wore a red scarf is killed by a man .
punct
obl:agent
acl:relcl
obj
nsubj:pass
case
nsubj
det
det
aux:pass
det
ref
amod
DT NN VBZ VBN IN DT NN WP VBD DT JJ NN .
The woman is killed by a man who wore a red scarf .

Figure 4.8: Dependency structure of the two sentences: "The woman who wore a red scarf is killed by a man." and "The woman is killed by a man who wore a red scarf."

As the consequence of the changing in calculation method, our analysis produces different results for calculating the maximum length of dependency path of each headline. Figure 4.9a shows the distribution of the maximum length of dependency path in both clickbait and non-clickbait headlines in Chakraborty corpus. A clickbait headline can have either a quite shallow dependency trees with the length of 1 or 2, or a quite deep one with the length of 6 or more. The length of non-clickbait dependency trees are normally only around 3 to 5.



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.9: Distribution of longest syntactic dependencies between dependencies and roots in clickbait and non-clickbait headlines

The reason for shallow dependency tree in clickbait headlines is that some clickbait headlines are only fragments. For example,the clickbait headline *"The 22 best Adele memes"* is only a noun phrase with "memes" as the root and the only governor while all other words are its direct dependencies. This means the dependency tree of this headline only has two levels, making the maximum dependency path 1. Meanwhile, a non-clickbait example "Lebanese Soldiers Killed in Ambush" has the same number of words, but as it is a full sentence, its dependency structure is more complex with the maximum length of dependency path of 2.

Figure 4.9b shows the distribution of maximum dependency path length in Potthast corpus. It shares the same result that there is a larger proportion of clickbait headlines with shallow dependency tree, even though the clickbait headlines in Potthast corpus has a slightly longer dependency path length in the range of 2 to 4. Non-clickbait headlines tend to have much deeper dependency trees with the maximum length of dependency path of more than 3. The proportion of clickbait and non-clickbait headlines that have the maximum length of dependency path at 3 or 4 is quite the same and it contributes to a larger portion of the corpus. Therefore, using the maximum length of dependency path as a feature for the machine learning system could creates noises.

The calculation of dependency path using our new methods shows that the finding of Chakraborty that clickbaits tend to have a more complex structure than non-clickbaits is not valid for both corpora. This feature can show the difference between clickbaits and non-clickbaits, however, the difference can be quite moderate.

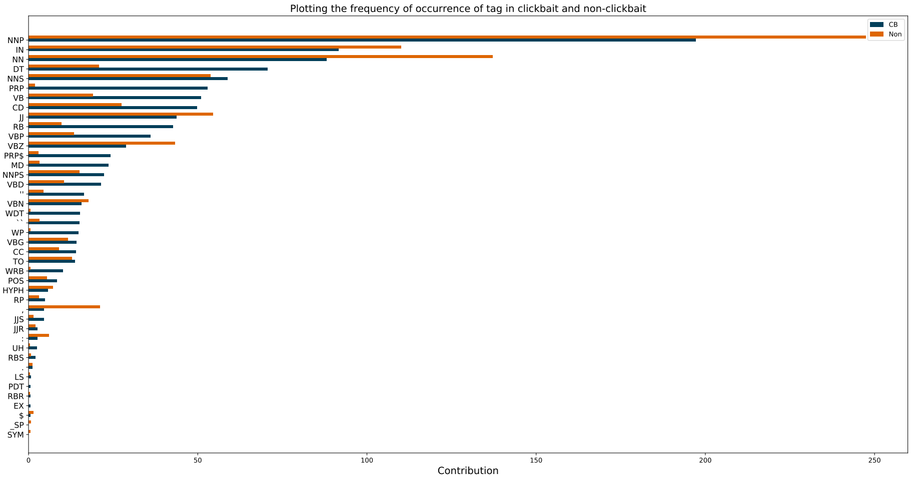


Figure 4.10: The distribution of part-of-speech tags for words in clickbait and non-clickbait headlines (Chakraborty et al., 2016)

4.2.1.2.4 Part-of-speech and part-of-speech n-gram Chakraborty et al. (2016), when examining the distribution of each part of speech in their corpus as shown in Figure 4.10, make an observation that clickbait headlines have higher proportion of adverbs and determiners (RB, DT, and WDT) than non-clickbait personal and possessive pronouns (PRP, and PRP$) compared to non-clickbaits. They also claim that there is a much larger proportion of proper nouns indicating more content words and entities in conventional non-clickbait than in clickbait headlines. In addition, clickbaits use more verbs as they focus on forming well-formed sentences. Verbs in past participle and 3rd person singular form (VBN and VBZ) tend to be used more in non-clickbait headlines, whereas clickbaits use mostly past tense and non-3rd person singular forms (VBD and VBP).

In addition to the findings of Chakraborty et al. (2016), our analysis shows a significant use of quotation marks corresponding to the large proportion of proper nouns, and wh-pronouns (WP) and wh-adverbs (WRB) signifying interrogative structures in clickbait headlines.

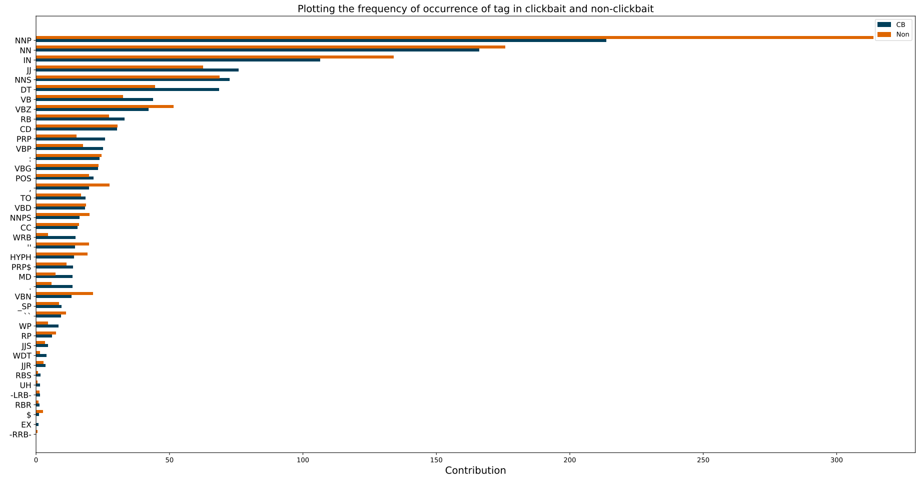
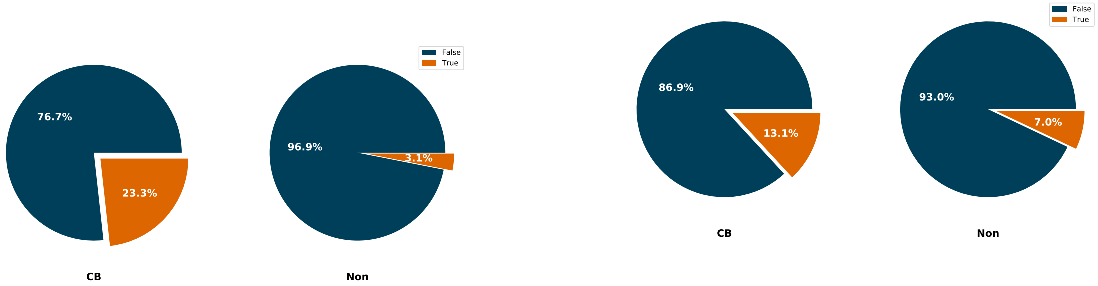


Figure 4.11: Plotting the frequency of each part-of-speech (Potthast et al., 2018b)

The results from analysing Chakraborty corpus is also true for Potthast corpus, however, the differences between indexes is much less considerable. For example, in Chakraborty corpus, 23,3% of clickbait and only 3,1% of non-clickbaits headlines contain modal verbs as shown in Figure 4.12a, while in Potthast corpus 13.1% and 7% relatively as shown in Figure 4.12b. The difference between the proportion of clickbait headlines that use modal verbs and the proportion of non-clickbait headlines that use modal verbs decreases from 20,2% in Chakraborty corpus to only 6,1% in Potthast corpus. It is not only there are less clickbait headlines using modal verbs, but also there are more non-clickbait headlines using modal verbs in Potthast corpus, comparing to in Chakraborty corpus. This is also relevant to other types of part-of-speech, which leads to the fact that it is harder to make a distinction between clickbait and non-clickbait headlines in Potthast corpus that in Chakraborty corpus using the distribution of part-of-speech tags.

Since the distribution of each individual part-of-speech tag cannot really tell much about the difference in the syntactic structures of clickbait and non-clickbait headlines, we look in to the distribution of part-of-speech n-gram which is a contiguous sequence of part-of-speech of words that appear consecutively, with the hope of capturing the syntax of clickbait and non-clickbait headlines by the part-of-speech associations. In their research, Chakraborty et al. (2016) consider word n-grams as a characteristic to distinct clickbait and non-clickbait. They conclude that there



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.12: Proportion of clickbait and non-clickbait headlines that use modal verbs

is a pattern of phrases repeated in clickbait headlines while non-clickbait headlines are unique as they report on facts and events with example phrases. Our argument is that using word n-grams cannot truly generalise the characteristics of either clickbaits or non-clickbaits because phrases can be varied with the change of only one words. For example, Chakraborty et al. (2016) mention that the phrase *"can you guess your"* occurs very frequently in clickbait headlines. We found several variance of this phrase with the same structure *modal verb-pronoun-verb-prossesive pronoun* (*MD-PRP-VB-PRP$*) but only different vocabulary such as *"should you buy your"*, *"should you instagram your"* or *"can you pass your"*. Even though other phrases do not occur as frequent as "can you guess your", their shared syntactic structure is typical for clickbait headlines as it does not appear in non-clickbait headlines. Therefore, identifying clickbait or non-clickbait headlines using phrases or word n-grams is not as exhaustive as using structures or part-of-speech n-grams.

In addition, non-clickbait headlines actually also follow some patterns but these patterns are structural and cannot be detected with word n-grams due to the diversity in vocabulary. As can be seen in Figure 4.13 and 4.14, there is a noteworthy dissimilarity between the distributions of 30 most frequent part-of-speech tri-grams and four-grams in clickbait and non-clickbait headlines.There are several part-of-speech tri-grams and four-grams that appear significantly more frequent in clickbait than in non-clickbait and vice versa. The tri-gram *noun-adposition-proper noun* (*NOUN-ADP-PROPN*) and four-gram *noun-adposition-proper noun-proper noun* (*NOUN-ADP-PROPN-PROPN*) occurs considerably more frequent in non-clickbait than in clickbait headlines. Some phrases in non-clickbait headlines that have the

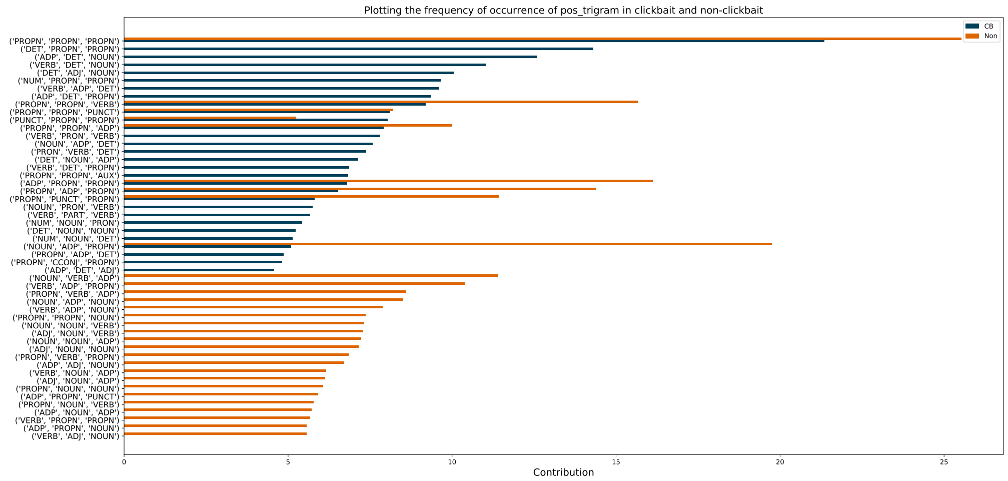


Figure 4.13: Plotting the frequency of each part-of-speech tri-gram (Chakraborty et al., 2016)

structure above are *"box at United Airlines Flight"*, *"elections in El Salvador"*, *"expulsion of New York"*, *"troops to Afghan War"*, etc. Some other structures such as *number-proper noun-proper noun-proper noun* (*NUM-PROPN-PROPN-PROPN*), *verb-adposition-determiner-noun* (*VERB-ADP-DET-NOUN*), or *determiner-proper noun-proper noun* (*DET-PROPN-PROPN*) occur most frequently in clickbait headlines but quite infrequent in non-clickbait ones. Some structures can appear in both clickbait and non-clickbait headlines, although, in Chakraborty corpus, there are not many. Special cases are the tri-gram *proper noun-proper noun-proper noun* (*PROPN-PROPN-PROPN*) and *proper noun-proper noun-proper noun-proper noun* (*PROPN-PROPN-PROPN-PROPN*) which can represent named entities such as *New England Patriots*, or *Floyd Mayweather Jr.*, but also can be noises from the corpus because as headlines are collected from the real news articles, they still keep the published format in which all words are capitalised. This format tricks our analysis tool to recognise normal nouns as proper nouns. For example, the phrase *Boston Terror Suspect Usaamah Rahim* only contains three words that are proper nouns, but the analysing program assigns the tag *"PROPN"* to all of them.

Figure 4.15 and 4.16 provide the distribution of tri-grams and four-grams in Potthast corpus. As can be seen from the two charts, 50% of most frequent tri-grams and four-grams are the same in both clickbait and non-clickbait headlines,

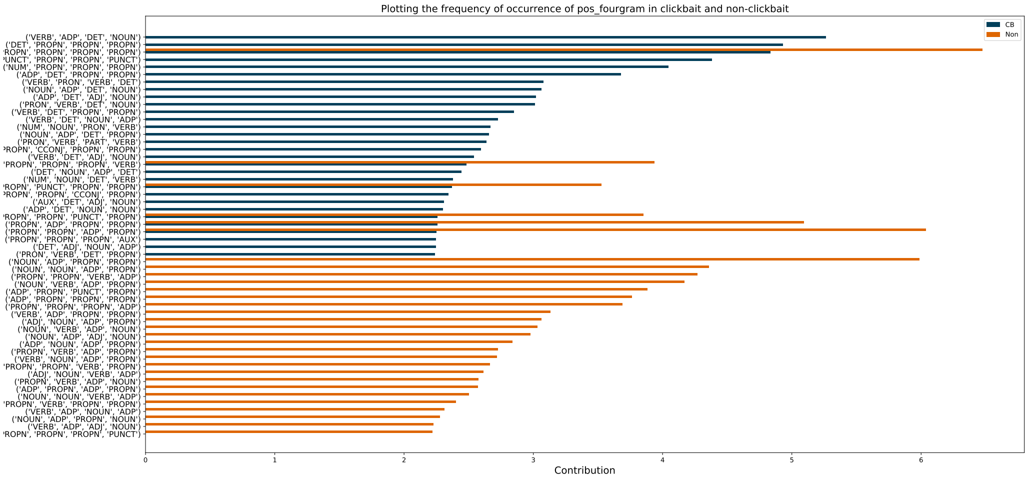


Figure 4.14: Plotting the frequency of each part-of-speech four-gram (Chakraborty et al., 2016)

but their frequencies of occurrence in each type of headline are different. For example, the tri-gram *proper noun-proper noun-proper noun PROPN-PROPN-PROPN* shows up in 35% of non-clickbait headlines but only in about less than 20% of clickbait ones. The tri-gram *proper noun-proper noun-verb *PROPN-PROPN-VERB)* and *verb-adposition-proper noun (VERB-ADP-PROPN)* and the four-gram *proper noun-proper noun-proper noun-proper noun (PROPN-PROPN-PROPN-PROPN)* and *noun-adposition-proper noun-proper noun (NOUN-ADP-PROPN-PROPN)* occur twice as many times in non-clickbait than in clickbait headlines. Additionally, the distinction between the use of four-grams in clickbait and non-clickbait is much clearer than the use of tri-grams, since there are more four-grams that most frequent appear in either clickbait or non-clickbait than tri-grams, indicating that it is better to differentiate clickbait and non-clickbait using four-grams.

Comparing the frequencies of tri-grams and four-grams in the two corpora, we can observe some repetitions of syntactic structures in both corpora. There are a tri-grams and four-grams that frequently appear only in either clickbait or non-clickbait headlines, which is consistent across all datasets. For examples, four-grams *determiner-adjective-noun-adposition (DET-ADJ-NOUN-ADP)*, *verb-determiner-adjective-noun (VERB-DET-ADJ-NOUN)*, *adposition-determiner-adjective-noun (ADP-DET-ADJ-NOUN)*, and *determiner-noun-adposition-determiner (DET-NOUN-ADP-DET)*

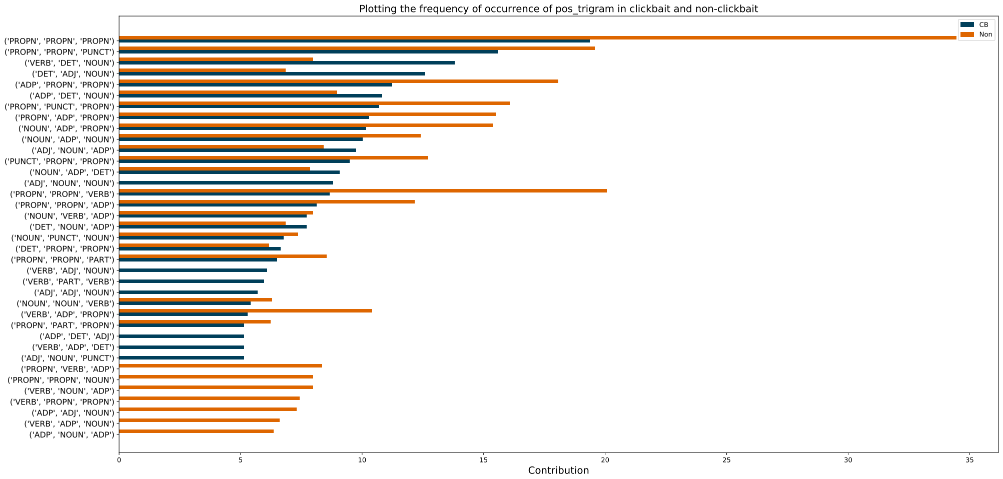


Figure 4.15: Plotting the frequency of each part-of-speech tri-gram (Potthast et al., 2018b)

are found much more frequently in clickbait headlines, while *proper noun- proper noun-verb-adposition (PROPN-PROPN-VERB-ADP)*, *verb-adposition-proper noun-proper noun (VERB-ADP-PROPN-PROPN)*, *verb-particle-verb (VERB-PART-VERB)*, *adposition-determiner-adjective (ADP-DET-ADJ)*, and *verb-adposition-determiner (VERB-ADP-DET)* tend to appear in non-clickbait headlines.

The analysis of tri-grams and four-grams also uncovers a wide discrepancy between the two corpora. Overall, clickbait and non-clickbait headlines in Chakraborty corpus are more discrete than in Potthast corpus in term of syntactic structures for the reason that there are less tri-grams and four-grams that appear in both clickbait and non-clickbait headlines in Chakraborty corpus than in Potthast corpus.

There are several structures that occur quite frequent in only one type of headline in Chakraborty corpus, but not as frequent in both type of headlines in Potthast corpus. For example, the tri-gram *determiner-proper noun-proper noun (DET-PROPN-PROPN)* are used in about 20% of clickbait headlines and less than 5% in non-headlines in Chakraborty corpus, but the percentage of non-headlines using this tri-grams stay the same while the number of clickbait headlines using this tri-grams only makes up 5%. The same happens with the four-gram *noun-noun-adposition-proper noun (NOUN-NOUN-ADP-PROPN)*, only that it is used more commonly in non-clickbait headlines.

In vice versa, there are structures that are not popular in both type of headlines

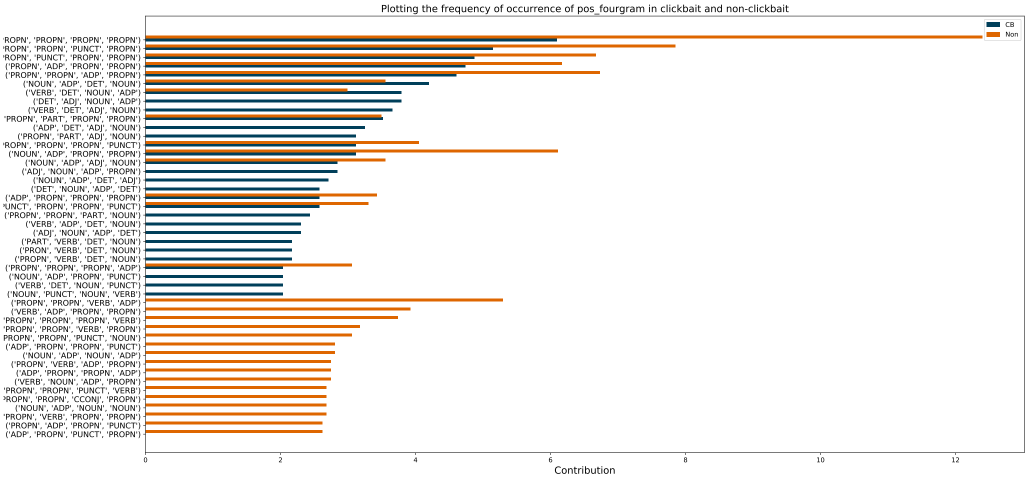


Figure 4.16: Plotting the frequency of each part-of-speech four-gram (Potthast et al., 2018b)

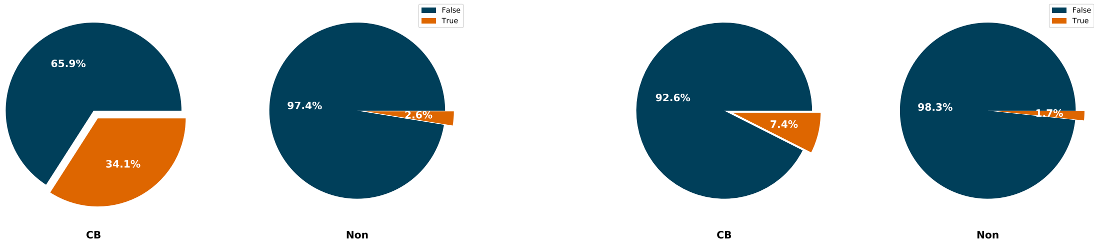
in Chakraborty corpus, but become more popular in one or another in Pothast corpus. The example for this is the four gram *noun-adposition-determiner-adjective (NOUN-ADP-DET-ADJ)* which is not among the 30 most frequent four-grams in Chakraborty corpus, but is the 17th most frequent used four-grams in clickbait headlines in Potthast corpus.

Furthermore, there are n-grams that occur more frequent in one type in Chakraboty corpus, then become more frequent in the other type in Potthast corpus, such as the tri-gram *noun-adposition-proper noun (NOUN-ADP-PROPN)* is 4 times more common in non-clickbait than in clickbait headlines in Chakraborty while in Potthast corpus, the percentage of clickbait headlines containing the tri-gram double and become only 1,5 times less than the number of non-clickbait headlines containing the tri-gram.

In all, based on our observation of the similarities and differences between clickbait and non-clickbait headlines from Chakraborty and Potthast corpus, we form an impression that the structures of clickbait headlines differ substantially from the structures of non-clickbait headlines in Chakraborty corpus, whereas the structures of the two type of headlines in Potthast corpus become more similar to each other, for non-clickbait headlines use structures that are more popular to clickbait

4.2.1.2.5 Sentence types Even though Chakraborty et al. (2016) do not discuss the use of different types of sentence in their analysis, it has been the focus on various studies on clickbait. As the result, we carry out a statistical summary of what types of sentence are used in clickbait and non-clickbait headlines in the two corpora based on the findings of other literature including Paulau-Sampio (2016), Scacco and Muddiman (2016), and Vijgen et al. (2014), as well as our observations.With regard to sentence types, we do not go deeply into the syntax, but rather look at the appearance of some representative structures.

Start with a number The first type of sentence we would like to exam is sentences that start with a number. This type of headlines are listicle headlines which generally announce the number of items in the content (Vijgen et al., 2014). An illustration for this kind of sentence is *"17 Times Kourtney Kardashian Shut Down Her Own Family"*. As seen in the example, this type of headline is not a completed sentence, but a fragment in the form of a large noun phrase. The head of the phrase is a general noun describing a collection which is, hence, usually in plural form. The pre-modifier of this noun phrase includes a cardinal (mandatory) and adjective or adverb (optionally). The post-modifier is usually a relative clause or a prepositional clause.



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.17: Plotting the percentage of listicle headlines

Figure 4.17a demonstrates the percentage of headlines that use listicle sentence type in Chakraborty corpus. It can be seen that there is a large proportion of clickbait headlines use listicle type of sentence structure, while there is only about 2 percent of non-clickbait headlines do. This result is consistent with the previous conclusion made by Vijgen et al. (2014). Examining the non-clickbait headlines that are detected as listicle headlines, we can see the clear differences between them and the clickbait ones. The non-clickbait headlines, even though start with a number, are completed sentences. Some of the examples are *"6.4 magnitude earthquake hits Taiwan"*, and *"2 Somali-Americans Charged With Aiding Terror"*. The former is a normal completed single sentence. The auxiliary is elided from the latter, however, this sentence still conveys a completed meaning.

The percentage of listicle headlines in Potthast corpus is set out in Figure 4.17b. Surprisingly, there is a dramatic decline in the number of clickbait headlines that use listicle structure from 34.1% to only 7.4%. Another interesting result is that there are some listicle headlines classified as non-clickbait in Potthast corpus. These are three examples: *46 Photos Of Sasha Obama Through The Years*, *5 Things About China-Taiwan Relations*, and *5 ways to not freak out on a plane*. Comparing to the examples from Chakraborty corpus, the three examples from Potthast corpus differs not only in terms of structure but also purposes. The headlines from Chakraborty reports on the events that happened while those from Potthast corpus do not report on any events, but rather give information.

When studying the examples mentioned above, we notice that the contents of non-clickbait articles that have listicle headlines are dissimilar to the contents of typical clickbait ones. One of the main characteristic of a listicle is that its content is arranged in the form of a list, which means there are a number of items and each item is represented in clear seperation. Table 4.3 is an example of a typical clickbait listicle that is taken from Potthast corpus.

Headline	*29 Gifts That Even The Most Heartless People Would Adore*
Content	*Kawaii!! Kawaii!! We hope you love the products we recommend! Just so you know, BuzzFeed may collect a share of sales from the links on this page. 1. Pillow rolls to supplement your love for corgis and kittens. Get them from ThinkGeek for $20. 2. Marshmallow mugs that don't need a fire to melt your heart. Get them from Amazon for $37. 3. A happy hedgehog to keep your brushes nice and clean. Get it from Amazon for $7. 4. Sushi pillows that would do anything to be squeezed. Get the pillow on the left from Amazon for $47 and the one on the right for $18.5...*

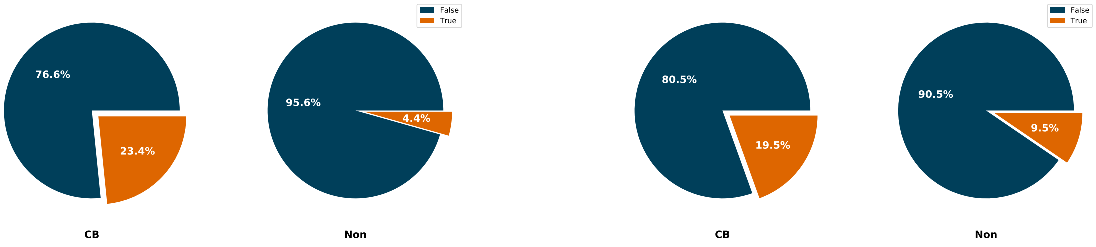
Table 4.3: Example of the headline and content of a listicle from Potthast et al. (2018b)

Meanwhile, the contents of some non-clickbaits that have listicle headline maintain the traditional format of a news article. An example shows in Table 4.4. This could mean that even though traditional journalism is borrowing some strategies of clickbait, they still try to maintain some qualities.

Headline	*5 ways to not freak out on a plane*
Content	*(CNN) The seats are too small, the meals are nonexistent, and the people across the aisle have taken off their shoes and socks to share their stinky feet with everyone. It's no wonder we're all a little infected with air rage. What is the weary traveler to do? If you've got enough room in your coach seat for yoga, that's awesome. If yoga's not your style and you don't have a Xanax prescription, here are more ways to survive short- and long-haul flights. Frequent traveler Arabella Bowen, editor in chief of Fodor's Travel, recommends combining strategies for greater impact. "Noise-canceling headphones, an iPad loaded with a favorite TV series or recent documentaries and a window seat are essential ingredients to my remaining calm," Bowen said...*

Table 4.4: Example of the headline and content of a non-clickbait with a listicle headline from Potthast et al. (2018b)

Question It has been reported that headlines in interrogative forms are more appealing to readers (Lai and Farbrot, 2014). Therefore, it is expected that question headlines are exploited frequently in clickbait headlines, which is shown in Figure 4.18a and 4.18b.

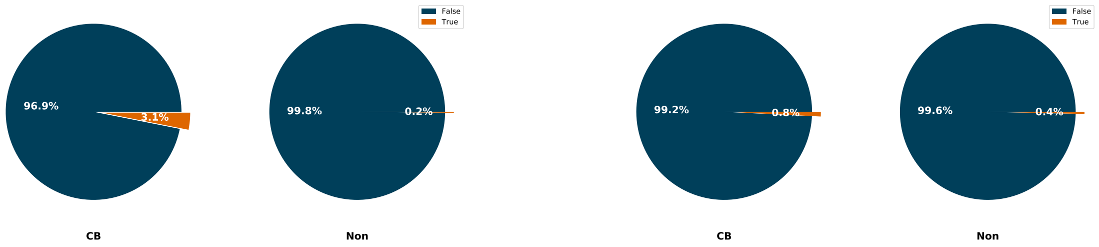


(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.18: Plotting the percentage of questions

In both corpora, about 1 in 4 clickbait headlines adopt the question forms, while only 4.4% of non-clickbait headlines in Chakraborty corpus are questions. The most interesting aspect of this graph is that the percentage of non-clickbait question headlines in Potthast corpus actually double than in Chakraborty corpus, while the percentage of clickbait question headlines in Potthast corpus is slightly smaller than in Chakraborty corpus.

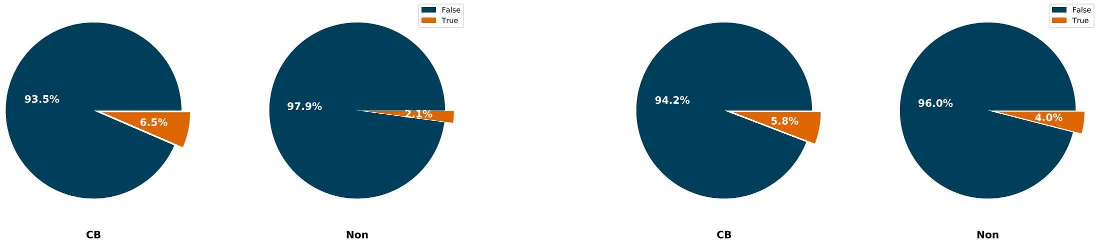


(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.19: Plotting the percentage of conditionals

4.2.1.2.6 Conditionals, superlatives, passives What has emerged from our exploration of the Chakraborty corpus is the recurring use of conditional, superlative and passive structures in headlines.

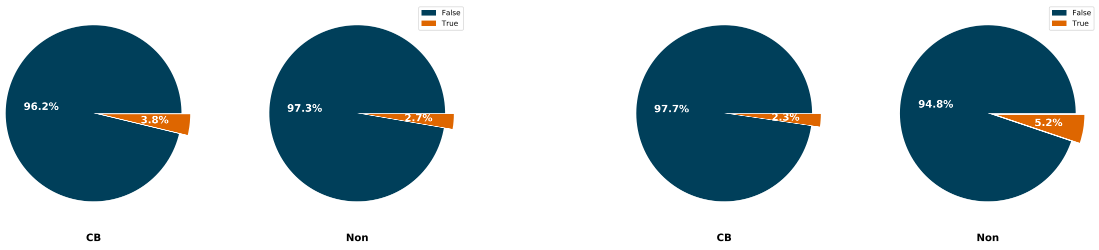


(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.20: Plotting the percentage of superlatives

Manually checking, conditionals are found quite often in clickbait headlines from Chakraborty corpus. Conditional headlines can take the form of a conditional sentence with two clauses: the conditional clause which includes *"if"* or *"unless"* and the main clause, for examples *21 Words You Won't Truly Understand Unless You're From Miami*, or *27 Pictures That Are Way Too Real If You Have A Four-Person Family*. Also, conditional headlines can be a fragment that only includes the conditional clause with "if" like *If Disney Princesses Were From Florida*. However, with the automatic checking system, we are able to find all headlines that use conditionals. Conditional headlines only contribute 3.1% to the total of clickbait headlines in Chakraborty corpus as shown in Figure 4.19a. The percentage of conditional headlines drop to under 1% in Potthast corpus. Meanwhile, in both corpora, there is only an extremely small percentage of non-clickbait headlines that has conditional structures.



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

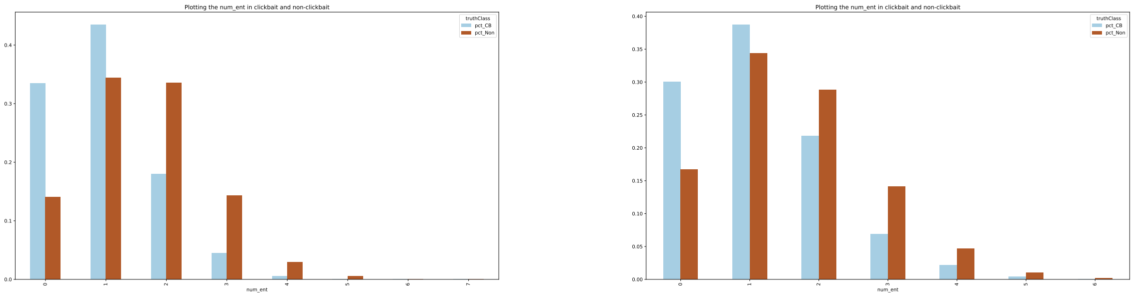
Figure 4.21: Plotting the percentage of passives

We start our analysis with an assumption that passives are more likely to be found in non-clickbait headlines as passives are more favourable in academic writing, and superlatives tend to used in clickbait headlines because of their high sentiment. The analysis shows that the ratio of clickbait headlines that use superlatives is quite the same in the two corpora, but the ratio of non-clickbait headlines using superlatives in Potthast corpus is twice as many as in Chakraborty corpus. Meanwhile, passives are used more in clickbait than non-clickbait headlines in Chakraborty, and twice as many in non-clickbait as in clickbait headlines in Potthast corpus.However, the frequencies of both superlatives and passives in the two corpora are relatively low in comparing to listicles or questions.

4.2.1.3 Semantic Level

4.2.1.3.1 Named entities Named entities are able to tell what is talked about in the headlines. As shown in 4.22a and 4.22b, a clickbait headline tends to have a lower number of named entities than a non-clickbait one.

Looking into the frequency of each type of entities as in 4.23a and 4.23b, we pick up some very interesting results about the main topics of clickbait and non-clickbait. What is striking in 4.23a is that *CARDINAL* (numer) and *WORK-OF-ART* type

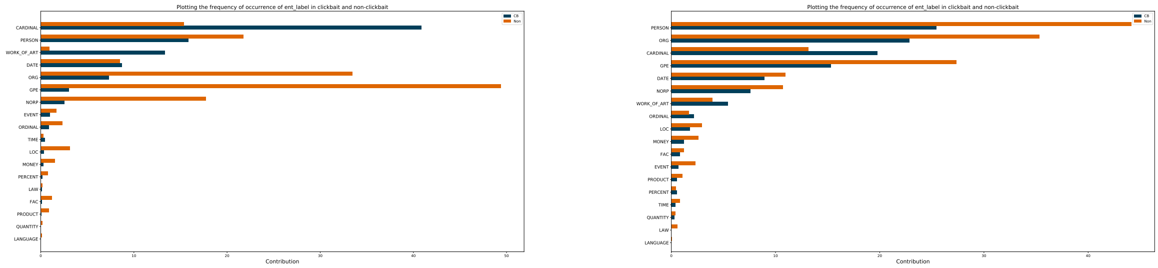


(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.22: Plotting the distribution of the number of entities in clickbait and non-clickbait headlines

of entities occur much more frequent in clickbait than in non-clickbait while the occurrences of *ORG*, *GPE* and *NORP* which talk about organisations; countries, or cities, or states, and nationalities or religious or political groups are extremely high in non-clickbait headlines. *LOC*, *MONEY*, and *FAC* which are about locations, currencies, and facilities (such as buildings, airports, highways, bridges, etc) are also more common in non-clickbait headlines. It is quite obvious that clickbait headlines in Chakraborty corpus focus on the topic of art, or entertaining while non-clickbait headlines concern more serious topics of politics, news, or events.



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.23: Plotting the distribution of each type of entities in clickbait and non-clickbait headlines

On the other hand, 4.23b shows several changes that reverse the above observation. What can be seen is a shape rise in the use of *ORG* category in clickbait, from about 8% in Chakraborty corpus to about 35% in Potthast corpus.*GPE* also appear three times more frequent in clickbait headlines in Potthast corpus than in Chakraborty. At the same times, we observe a dramatic decrease in the appearance of *WORK-OF-ART* category in clickbait headlines in Potthast corpus, even though it still occurs more frequent in clickbaits than non-clickbaits. These changes above suggest a shift in the topic of clickbaits: clickbaits are focusing more on the same serious topics as non-clickbait instead of soft topics.

4.2.1.3.2 Subjects In order to gather evidence to support our claims about the change in the topics of clickbait, we examine the 20 most frequent subjects which includes subjects, objects, direct objects, indirect objects from clickbait and non-clickbait headlines in the two corpora. The reason for our investigation in subjects is that an subject talks about the participant of an event, or the holder of a state of affair, which actually makes up the topic of a sentence. Consider an example from the Potthast corpus *"London's newest chicken takeaway is 100 percent vegan"*, if we only consider the named entities *"London"* then we can assume that this sentence talks about London city which is actually not the real topic of this sentence, but *"takeaway"* which is the subject of the sentence.

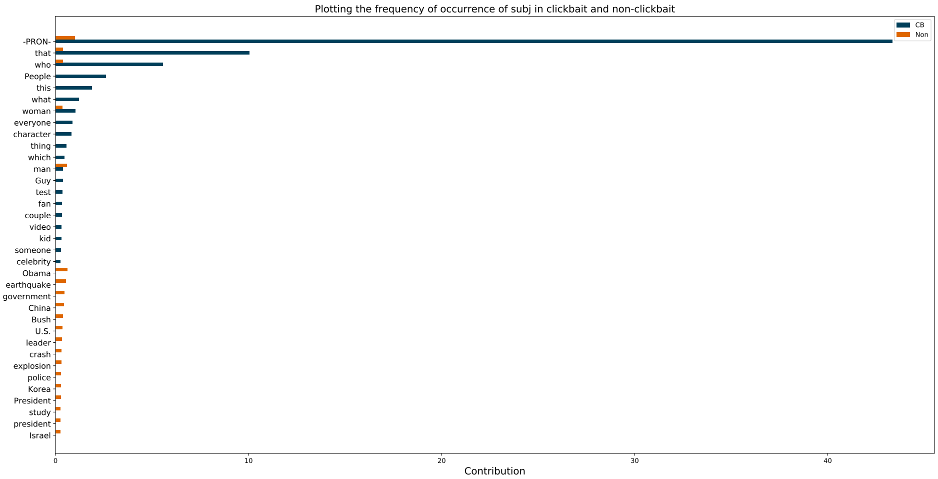


Figure 4.24: Plotting the most frequent subjects in clickbait and non-clickbait headlines (Chakraborty et al., 2016)

Figure 4.24 demonstrates the 20 most frequent subjects of the two types of headline in Chakraborty corpus. It is quite obvious that non-clickbaits aim at reporting on the real-life events since the subjects of non-clickbaits are mostly about real people like *"Obama"* [5], *"Bush"* [6] or *"President"* [7], real countries like *"China"* or *"Korea"* and serious events and phenomena including *"earthquake"*, or *"election"*. On the contrary, clickbaits anonymously talk about people using pronouns, wh-adverbs *"who"*, or generic noun including *"man"*, *"women"*, *"people"*, *"everyone"*, *"life"*, *"couple"*, *"kid"*, and ect. No specific event or phenomenon is frequent discussed in clickbaits. Events and phenomena are generically referred as *"thing"*, *"what"*, *"this"*, or *"that"*. This strategy of using generic terms and pronouns in combination with determiners indicates the existence of forward referencing mentioned in several studies on clickbaits.

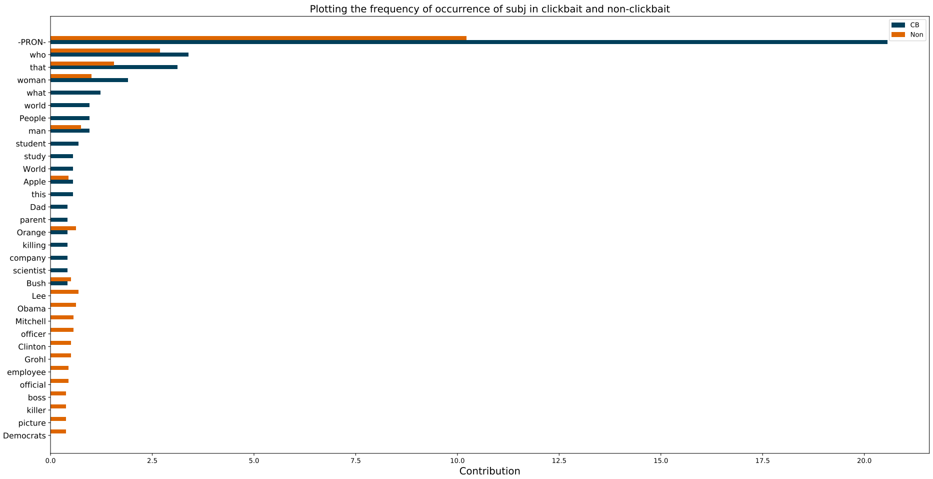


Figure 4.25: Plotting the most frequent subjects in clickbait and non-clickbait headlines (Potthast et al., 2018b)

From 4.25, we notice the similarity between the two corpora in what is used as subjects of clickbait and non-clickbait: clickbait is more likely to use generic terms while non-clickbait tend to use specific terms as subjects or objects. However, the occurrence of pronouns in non-clickbait goes up from about 2% in Chakraborty corpus to more than 10% in Potthast corpus even though there is still a large percentage of clickbait headlines with the appearance of pronouns in Potthast corpus.

[5] Barack Obama is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.

[6] George Bush is an American politician and businessman who served as the 43rd president of the United States from 2001 to 2009.

[7] "President" in capital imply The president of the United States who is the head of state and head of government of the United States of America.

By the same token, there is also a slight increase in frequency of generic terms such as *"who"*, *"that"*, and *"women"* in non-clickbait headline in Potthast corpus, in comparing to in Chakraborty corpus. In addition, a few proper names show up quite often in clickbait headlines including *"Bush"*, and *"Apple"* [8]. As we can see that the observation on the occurrence of named entities in the two headline is quite applicable with the analysis of subjects. This means that the topic of clickbaits is shifting toward the topic of serious news reports. Moreover, we can see that with the higher frequency of generic terms, non-clickbait headlines in Potthast corpus also take advantage of forward referencing which is only used in clickbaits in Chakraborty corpus. Even though the difference between the subjects of clickbaits and non-clickbaits seem to get smaller, it is still critical for the separation between clickbaits and non-clickbaits.

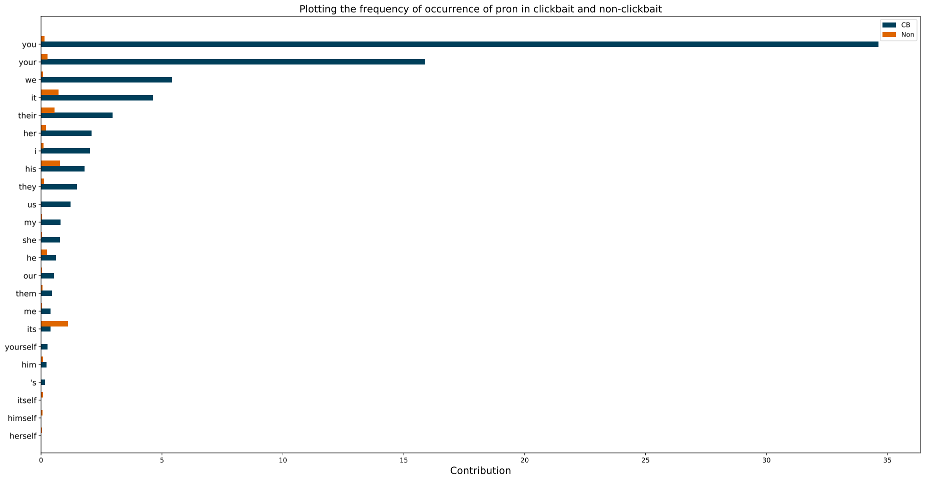


Figure 4.26: Plotting the most frequent pronouns in clickbait and non-clickbait headlines (Chakraborty et al., 2016)

4.2.1.3.3 Pronoun Due to the five-fold increase in the frequency of pronouns in non-clickbait headlines in Potthast corpus, we carry out a further investigation into the use of each pronoun. Figure 4.26 and 4.27 shows the frequency of each pronoun occurring in the two corpora. What is significant is that second person pronouns *"you"* and *"your"* appear remarkably more frequent in clickbait than in non-clickbait headlines in both corpora. Likewise, first person pronouns including *"we"*, *"our"*,

[8] Apple Inc. is an American multinational technology company

"us", *"I"*, *"my"* are also more popular in clickbait headlines. Simultaneously, third person pronouns including *"he"*, *"his"*, *"she"*, *"her"*, *"its"*, *"their"* are even more common in non-cickbait then in clickbait headlines in Potthast corpus. Reflexive pronouns consistently appear more frequent in non-clickbaits than clickbaits, while the third-person possessive pronoun *"it"* singly occur much more often in clickbaits than non-clickbaits.

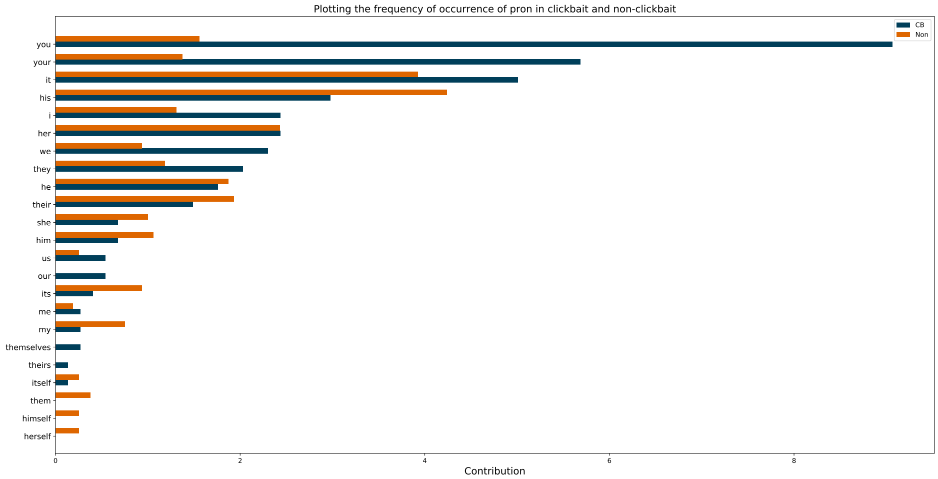
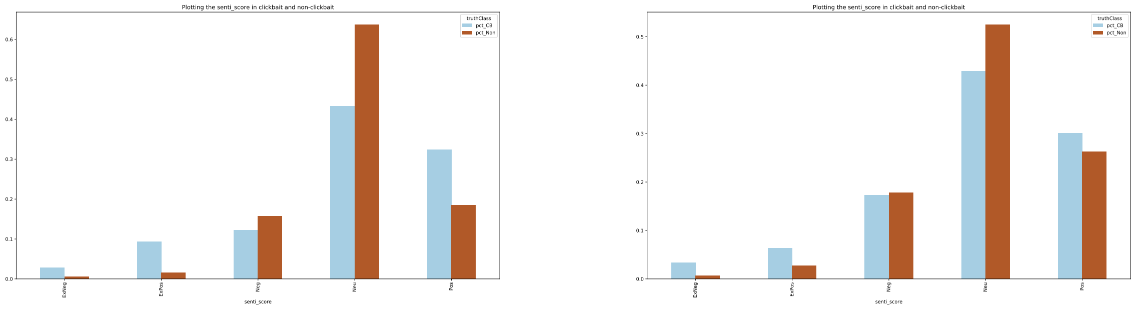


Figure 4.27: Plotting the most frequent pronouns in clickbait and non-clickbait headlines (Potthast et al., 2018b)

4.2.1.3.4 Sentiment Sentiment has been considered one of the characteristics of clickbaits according to Chakraborty et al. (2016). In their study, Chakraborty et al. (2016) find a considerable fraction of clickbait headlines consisting of words having 'Very Positive' sentiments like *"Awe-inspiring"*, *"gut-wrenching"*, and *"soul-stirring"*. Our analysis focuses on the calculation of the sentiment score of each headline instead of searching for only a few individual words. A headline can have the sentiment from the scale of -1 to 1 which is later translated in to a nominal scale including five different levels of sentiments: *extreme negative* (ExNeg) which has the sentiment score from -1 to -0.5, *negative* (Neg) with the sentiment scores between -0.5 and 0, *neutral* (Neu) which has the sentiment equal to 0, *possitive* (Pos) which is from 0 to 0.5 and *extreme positive* (ExNeg) for the sentiment from 0.5 to 1.

Figure 4.28a and 4.28b compare the distribution of sentiment scores among clickbait and non-clickbait headlines in Chakraborty corpus and Potthast corpus. The



(a) Chakraborty et al. (2016)

(b) Potthast et al. (2018b)

Figure 4.28: Plotting the distribution of sentiment scores in clickbait and non-clickbait headlines

result shows a great similarity between the two corpora. A major part of both clickbait and non-clickbait headlines in the two corpora are in neutral tone, meaning that their sentiment scores are 0. Non-clickbait headlines in both corpora tend to have a more negative tone with the sentiment scores from -05.to 0, while there are slightly more clickbait headlines have positive tone with the sentiment scores from 0 to 0.5. What is highlighted in the two charts is that there are a larger portion of clickbait headlines in both corpora being either extreme positive or extreme negative. As stated in Chakraborty et al. (2016), words with very high sentiment are used in clickbaits delivering a promise of sensational information. However, sensational feelings in clickbait can also be over negative, not only over positive.

On the one hand, sentiment scores can be an obvious indicator for clickbaits since there are many more clickbaits with extremely high or low sentiment scores, one the other hand, it could be difficult to tell if a headline has a neutral or slightly positive tone.

4.2.2 Analysis of Clickbait Contents

The similarity between clickbait headlines and clickbait contents has been considered a characteristic of clickbait in many studies including Biyani et al. (2016), Cao et al. (2017), Wei and Wan (2017), Wei and Wan (2017). However, these studies only focus on extracting features for machine learning systems, instead of exploring the characteristics of clickbaits. Therefore, we include the analysis of clickbait contents which compare the similarity between clickbait headlines and contents, justifying the hypothesis that clickbait contents do not align with their headlines. At the same time, we also include the extraction of some basic stylometric characteristics from the contents.

4.2.2.1 Stylometric characteristics

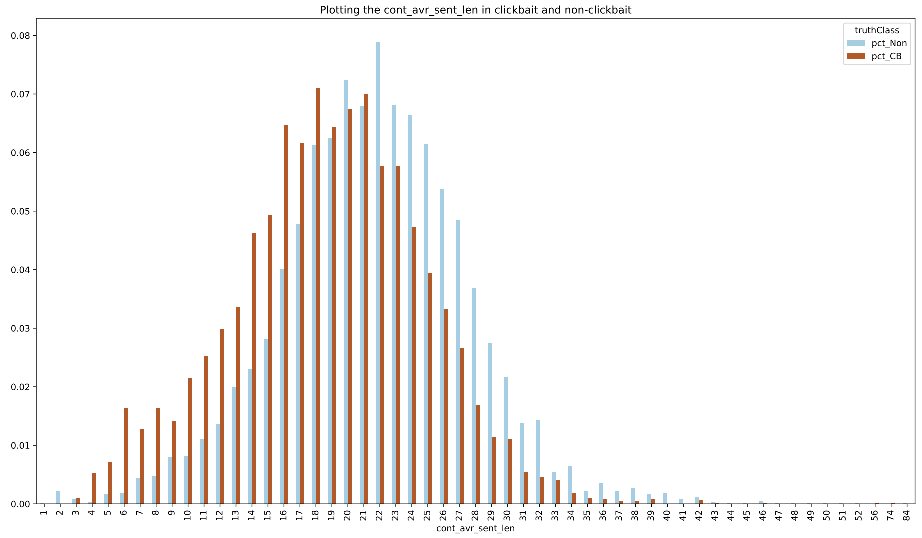


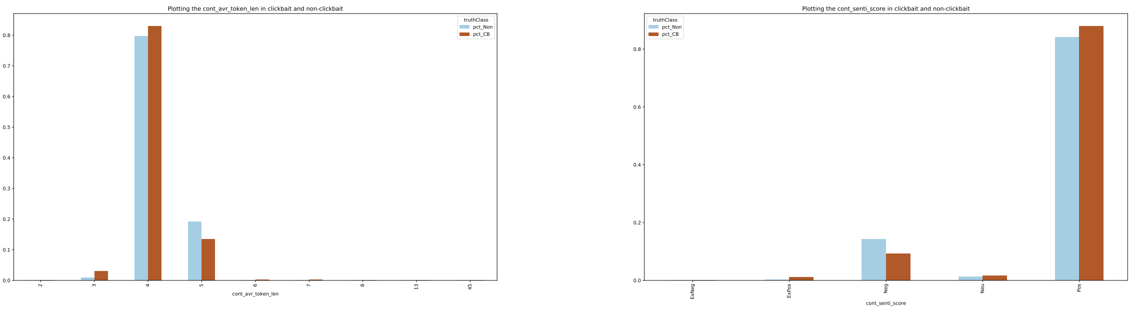
Figure 4.29: Plotting the average sentence length in clickbait and non-clickbait contents

The content stylometric characteristics examined in this analysis include the average number of sentences, the average sentence length, the average number of tokens, the average token length, the sentiment scores.

In average, the content of a clickbait is slightly longer than non-clickbait with the average number of sentences in a clickbait content is 10 sentences more than the average number of sentences in a non-clickbait content.

Since the average number of tokens in clickbait content is also higher than the average number of tokens in non-clickbait content (750 tokens in clickbait contents and 705 in non-clickbait contents), the average sentence length of clickbait contents is also slightly is higher than the average sentence length of non-clickbait contents. As can be seen from Figure 4.29, clickbait contents contains many more really short sentences while quite long sentences tend to appear more in non-clickbait contents.

Similarly, longer tokens also tend to occur more in non-clickbait contents, as shown in 4.30a. However, we can see that the difference between the average length



(a) Average token length

(b) Sentiment scores

Figure 4.30: Plotting the distribution of the average token length and the sentiment scores in clickbait and non-clickbait contents

of tokens in the two types of contents is not so significant.

Regarding to the sentiment score, it is believed that non-clickbait contents, like non-clickbait headlines, are more prone to a neutral or slightly negative tone. However, a large proportion of non-clickbait contents are observed to have quite positive sentiment. Still, like in the headlines, extreme positive sentiment is used more in clickbait than in non-clickbait contents.

4.2.2.2 The similarity between the contents and headlines

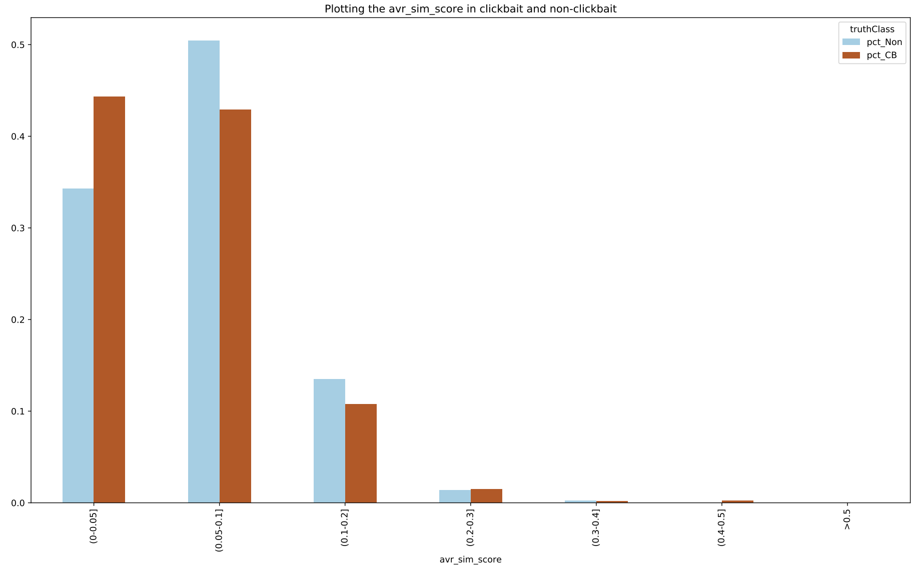


Figure 4.31: Plotting the average similarity scores between headlines and contents

Several studies above claim that the content of a clickbait is usually not related to its headline. We calculate the similarity between the headlines and contents of both clickbaits and non-clickbaits by building a vector space of keywords and calculate the average score of the similarity of keyword vectors in the headline and keyword vectors in each sentence of the content. The score is in the range between 0 which is not similar at all and 1 which is exactly the same. The result is demonstrated in Figure 4.32.

Even though there is a bigger percentage of non-clickbaits has the higher similarity scores, the majority of both clickbaits and non-clickbaits have the similarity score between headlines and content in the range of 0 to 0.1 indicating very low similarity. It is understandable that headlines are usually much shorter than content, and the information provided in headlines is much less than in contents, therefore, most of the time, there are many parts in the content do not align with what is in the headline.

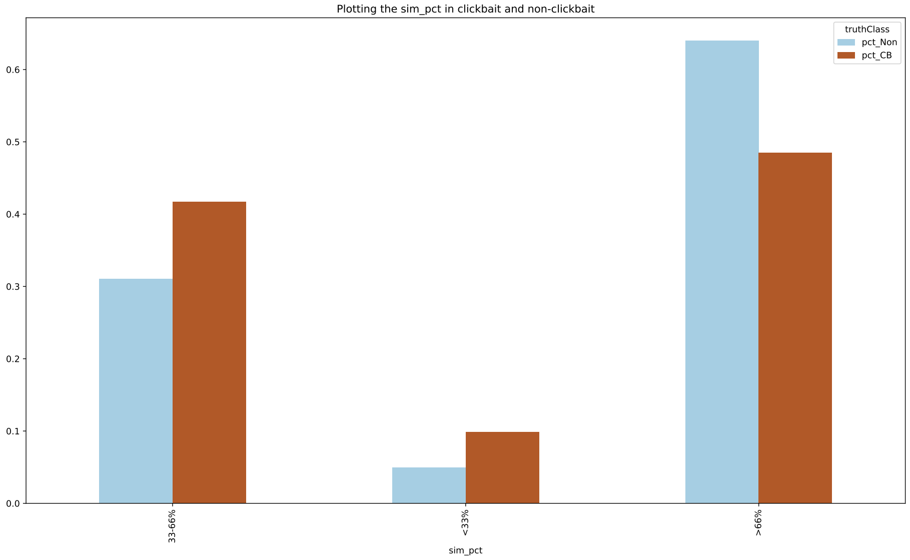


Figure 4.32: Plotting the proportion of sentences in contents that are similar to headlines

Nevertheless, when calculating the number of sentences in the contents that is similar to the headlines, we can see that frequently, more than 65% of non-clickbaits that have the contents that are 60% and more similar to their headlines, while the number for clickbaits is only about 45%. Meanwhile the percentage of clickbaits that have less than 33% of the contents that is similar to the headlines is double the percentage of non-clickbait ones. This means that the content of a non-clickbait tends to have a larger proportion of sentences that are similar to the headlines. If we only rely on the average similarity score between the headlines and the contents, we are not able to make a big distinction between clickbaits and non-clickbaits.

Looking into a few example from the Potthast corpus, we are able to come to an understanding about the relatedness of the content and headline in clickbaits and non-clickbaits. Consider the two examples in Appendix A, the first example is a non-clickbait and the second example is a clickbait. It is easily can be seen that the contents of both examples are related to their headlines. Yet, the non-clickbait content directly talk about the subject that is mentioned in the headline. The first sentence of the content is just a paraphrase of the headline stating the main point of the content. Other parts of the content are to elaborate on the main point by reporting about related information about the main event, such as who the victims of the kidnap were, or how the victims were transported, or why ISIS captured these people. The content also talks about the more general background of the event, however, that part stays at the end of the text body. After reading the whole article, we are able to get a whole of a coherent and informative story around the main event mentioned in the headline.

Meanwhile, it is much more difficult to find the main idea of the second article due to the organisation of the information in the content. First of all, the main point of the whole article is only described in only two clauses *"the art work for the single is being featured on a credit card"* and *"the artwork for their infamous album* Never Mind The Bollocks, Here's The Sex Pistols*, to illustrate consumer credit."*. These two clauses are parts of two very lengthy sentences that talk about a music album and a contract with a music company. Then the content goes on describing the band that owns the art-work used on a credit card. In addition, instead of summarising the main point, the article presents some of the quotation from Branson whose identity is unknown. The content of the clickbait does mention some keywords from the headlines like *"credit card"*, *"Sex Pistols"* or *"Anarchy"*, but there is not a completed story about them. We do not gain much new information about the card such as whether we need to pay more money for this special card, or what the art work is about, or how to obtain this card. In fact, the mentions of keywords from headlines are purely repetitive. Nevertheless, we suppose it is unlikely that machines are able to make such interpretation which requires discourse level of understanding.

4.3 Discussion

4.3.1 On feature selection

Based on the findings of the analysis, we summarise the importance of each parameter to the characterising of clickbait in Table 4.5. As can be seen from the table, any parameters that are very crucial for the differentiation of clickbaits and non-clickbaits in Chakraborty corpus become less but still crucial in Potthast corpus. Those which are already less important in Chakraborty corpus hardly have much impact in Potthast corpus. Therefore, we decide to keep only those which are important in both corpora to be engineered into features for the automatic detection. We also realise there are grey areas where an individual characteristic cannot help to tell a clickbait and a non-clickbait apart. Therefore, we think it is important to combine these characteristics together in order to form a better representation of clickbaits.

Parameter		**Chakraborty**	**Potthast**
Word/token level	Average word/token length	+	-
	Contraction	+	-
Syntax level	Sentence length	+	+
	Punctuation	++	+
	Maximum length of dependency path	++	+
	POS	++	+
	POS n-grams	N/A	+
	Start with number	++	+
	Question	++	+
	Conditional	-	-
	Superlatives	-	-
	Passives	-	-
Semantic level	Named entities	++	+
	Subjects	++	+

	Pronoun	++	++
	Sentiment scores	+	+
Content	Average sentence length	N/A	+
	Average token length	N/A	-
	Sentiment	N/A	-
	Similarity score between headline and content	N/A	+

Table 4.5: The summary on the importance of each parameter in differentiating clickbait from non-clickbait based on the results of the analysis (++: quite important, +: important, -: not important)

From the analysis, we also found some evidence for the changes in the use of strategy in clickbait construction from Chakraborty corpus to Potthast corpus in terms of syntax and semantic/topic. These changes are indicated by the difference in the occurrence of named entity types, subjects, punctuation, and especially syntactic structures in the headlines. With respect to the the feature selection, count and frequency features are no longer informative since clickbaits seem to acquire a new format resembling non-clickbaits, instead, we need to extract semantic and syntactic features since the differences between clickbaits and non-clickbaits are underlying. Therefore, in order to represent the semantic information, we use word embeddings and document embeddings for their ability of capturing the semantic similarity between two words or two documents. For the syntactic representation, we rely on the dependency relation of each word with its governor and its part-of-speech.

4.3.2 On the nature of clickbait

As stated about, the analysis reveals a remarkable development in the characteristics of clickbaits and non-clickbaits. Our hypothesis for this change is that people start to recognize the patterns in the clickbait headlines and stop clicking, therefore it is necessary to apply new strategy in in order to adapt to user preferences. Psychology studies on clickbaits from reader's perspective include Beleslin et al. (2017); Karaca (2019); Munger et al. (2018, 2020) all come to the conclusion that the participants of their studies are able to recognise clickbait structures in headlines and tend to avoid to view articles that have clickbait headlines. There is a time span of one to one and a half year different between the time the two corpora collected, so the changes between the two corpora can actually be chronicle. This could also explain the unremarkable results for machine learning system in the Clickbait Challenge since many of them are built to recognise old features that do not occur so frequently in the new corpus.

Another possibility is that the characteristics of each corpus are only significant within the corpus itself due to the data collection procedure. It cannot be denied that the subjective view of the authors of these corpora can strongly influence the method of collecting data. As for the Chakraborty corpus, Chakraborty et al. (2016) collect the headlines basing on the reputation of the publishers in the practice of clickbaiting, meaning that there is a presumption of the authors about the category of the headlines written by these publishers.

In addition, the lack of comparison between the headlines and the content in Chakraborty corpus could also lead to mislabeling. As presented in 4.2.1, even non-clickbait headlines are also changing with the more frequent use of some clickbait strategies such as using question form. What could happen is that non-clickbait headlines are marked as clickbaits since they apply some clickbait techniques in constructing headlines, but the contents are still able to deliver a good deal of information. Potthast et al. (2018b), in their attempt to create a guideline for anotating clickbait, also emphasise on the relevant of the content in the process, as they discuss the intention of authors to lure as many people as possible to a web page, disregarding the content's target audience. Based on the analysis of the contents and some examples in Potthast corpus, we truly believe that the contents can provide substantial cues in order for human annotators to recognize clickbaits. Therefore, in our opinion, a clickbait should be regarded as the whole article including both its headline and content.

The new findings about the characteristics of clickbait raises a question the procedure for collecting and annotating data. There should also be a change in our methods to handle clickbait which allow more interpretation of the clickbait concept for it is changing. Also, we need to collect more data from different periods and different sources in order to formulate some ideas of what is current and what changes in the clickbaiting practice. Only then, will we be able to extend our understanding of the clickbait concept and find the best solutions to prevent clickbaits.

Chapter 5

Automatic Detection Systems

5.1 Experiment Settings

5.1.1 Feature Engineering

The features for the automatic detection systems are extracted according to section 4.3.1. Based on the list of parameters in Table 4.5, we divide the features into four types, each of which are engineered with a different technique as in Table 5.1

5.1.1.1 Sylometric Features

Stylometric features includes all statistical features and syntactic features. Statistical features include count and binary features from the quantitative analysis such as the number of each part-of-speech tag, sentence length or if a headline is a question or not. Count features are numerical while binary features have a value of either *True* or *False*. These features can be computed easily using built in function of Python or the libraries for text analysis. The features are later represented by a sparse vector using DictVectorizer from Scikit-learn library.

In order to represent the syntactic structures of the headlines, we represent each headline as a sequence of part-of-speech and dependency tags. Each tag is encoded with a particular number, turning a sentence in an array of numbers. To do so, we use SpaCy vector representation for dependency and part-of speech tags. For instance, the sentence *"The global refugee crisis, region by region"* is represented by the vector [90, 84, 92, 92, 97, 92, 85, 92], and each dimension of this vector

Types	**Features**	**Method of engineering**
Stylometric	Sentence length	Number of token in a sentence
	Punctuation	Assign value to each token, 0 if not punctuation, else 1
	Maximum length of dependency path	The length of each dependency to the root
	POS	Number of each POS in the headline
	POS n-grams, pronoun	Sequence of numerical values assigned for each POS tag
	Start with number	Assign value to each token, 0 if not punctuation, else 2
	Question	Boolean value (True or False)
	Named entities	Assign value to each token, 0 if not named entity, else numerical value assigned for each type of named entity
	Subjects	Sequence of numerical values assigned for the dependency of each token
	Sentiment scores	Sentiment scores
Word embedding	Tokens	100-dimension word embeddings
	Syntactic relation	numerical values of each token's POS tag, dependency relation, and word embedding of the governor
Combined	Combined features	Combine stylometric features and word embedding features
Document embedding	Similarity score between headline and content	Similarity scores between headline and content
	Content	100-dimension document embeddings

Table 5.1: List of features with the methods of engineering

is the encoded part-of-speech tag of each token in the sentence with a numerical value provided by SpaCy. In addition, we also mark punctuation, numbers, and named entities with different numbers as described in Table 5.1. For example, the sentence *"6 new Frapp flavors at Starbucks."* is encoded as [2, 0, 0, 0, 0, 0, 0] for the information about numbers, as [0, 0, 0, 0, 0, 0, 1] for the information about punctuation and [0, 0, 383, 0, 0, 383, 0] for the information about named entities. In the end, each token in a sentence is represented by a number with is the sum of the number represent its part-of-speech and dependency tags, named entities type, and the marks for punctuation and numbers. As each sentence has a different length, we have to resize the vector representing them to the same length of 30. If the length of the vector is less than 30, we add 0 to it at the end, and if the length excess 30, we slice the vector from the beginning to the 30th element. The syntactic feature vectors are concatenated with the feature vectors from the quantitative analysis and both are generally referred as stylometric features.

5.1.1.2 Embedding Features

The embedding features include word embeddings and document embeddings. These two features are learned using Word2vec and Doc2vec models [1] built using Gensim. We train our models using vocabularies from both Chakraborty et al. (2016) corpus and Potthast et al. (2018b) corpus. We try to use pre-trained embeddings GloVe, however, the two corpora contain words that is not included in the vocabulary of GloVe. Even though training our own embeddings models do take a lot of times and resources, it ensures that all words are represented. We believe that it is quite important to reserve the vocabulary of the corpus, since the frequency of stop words, symbols, punctuation, numbers or uncommon words are quite high. If we use filtered vocabulary, then some important information can be lost.

Since Gensim already provides a completed pipeline for training, our job is only to input the corpus to build a vocabulary and set parameters for the algorithms. For word embeddings, we select the skip-gram model with a 5-word window and the dimension set to 100. We iterate learning process 10 times with the learning rate starting at 0.025. The syntactical features are also integrated in word embeddings

[1] For the detailed description of the architectures of the two embedding models, refer to 2.2.1

in the way that a token is represented by its embeddings in combination with its part-of-speech and dependency tags and the embedding of its governor creating a 200 dimension vector. A sentence is represented by the sum of vectors of tokens in the sentence.

The combined features between stylometric features and word embeddings features are represented by 250-dimension vectors that consist of stylometric feature vectors and word embeddings vectors.

Document embeddings are obtained using 100-dimensional distributed bag of words (DBOW) mode. This mode, which is similar to the Skip-gram model in word vectors, tries to predict words that can appear in a collection of text windows providing the learned paragraph representation vectors from previous sections. The advantage of this model is that it is quite simple and requires much less memory. The embeddings learned from the training are used to represent the contents since each content can contain a large number of tokens, which can create a lot of noise if we use the sum method for word embeddings or consume much computational resources if we use the concatenation of syntactic representation. We also set the learning rate at 0.025 and the epoch at 5. The embedding of each content is also combined with the stylometric and word embedding vector of its headline creating a 350 dimension feature vector.

5.1.2 Machine Learning Architectures

We select three supervised machine learning algorithms for this project are basic algorithms for classification including Logistic Regression, Random Forest and Support Vector Machine [2] with Stochastic Gradient Descent. They are suggested from the literature as the most popular algorithms for this task. The algorithms are applied using Scikit-learn [3], one of the most popular machine learning libraries for the Python programming language. This library provides already built pipeline for each model with vector transformation functions that can easily be used to represent features. In order to apply the code correctly, we refer to the documentation of each function provided on Scikit-learn official release website. The documentation

[2] For the detailed description of the architecture of each algorithm, refer to Table 2.1

[3] https://scikit-learn.org/

is always accomplished with implementing examples which provides ease for users. We also refer to other websites, forums, or discussion groups for the optimization of each algorithm. All references are documented with each function is the code. Each algorithm is optimised using grid search which is a function provided in Scikit-learn library allowing the algorithm to be trained with different parameters and then choosing the best optimisation with the best result.

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

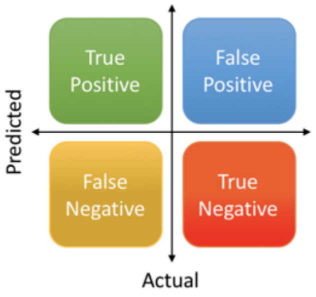


Figure 5.1: Formula for each performance evaluation metric

We use both corpora from Chakraborty et al. (2016) and Potthast et al. (2018b) for extracting features for headlines, and only corpus from Potthast et al. (2018b) since the corpus from Chakraborty et al. (2016) does not include clickbait contents. The data is shuffled and randomly split for training and testing at a 7:3 training-test ratio. To avoid over-fitting, we set the cross-validation generator parameter integrated in the grid search optimisation function at 5, creating 5-fold cross validation. We notice the unbalance in the number of clickbait and non-clickbait samples in the corpus of Potthast et al. (2018b). Even though we use the whole dataset for the analysis, for the training of machine learning systems, we decide to reduce the number of non-clickbait samples which is almost four times more than the number of clickbait samples in the corpus of Potthast et al. (2018b). In addition, we also reduce the number of data from Chakraborty et al. (2016) so there are equal numbers of data from each corpus. In the end, we have a combined dataset with a size of about 20000 data points in which 10000 are from the corpus of Chakraborty et al. (2016) and the rest are from the corpus of Potthast et al. (2018b)

The performance of the classifiers was evaluated in terms of the area under accuracy, precision, and recall. Accuracy calculates the ratio of total true predictions and overall predictions made by the model. Precision is the ratio of true positives

	Precision	**Recall**	**Accuracy**
sty[1]+ LR	0.74	0.74	0.74
sty + RF	0.75	0.75	0.75
sty + SVM	0.75	0.75	0.75
w2v[2]+LR	0.80	0.80	0.80
w2v+ RF	0.80	0.79	0.79
w2v+SVM	**0.82**	**0.82**	**0.82**
cmb[4]+LR	0.81	0.81	0.81
cmb +RF	0.80	0.80	0.80
cmb + SVM	0.81	0.81	0.81
d2v[4]+ LR	0.70	0.70	0.70
d2v + RF	0.70	0.70	0.70
d2v + SVM	0.70	0.70	0.70

[1] Stylometric features

[2] Word embedding features

[3] Combined features

[4] Document embedding features

Table 5.2: Performance of different classifiers with respect to features.

to the sum of true and false positives while recall is defined as the ratio of true positives to the sum of true positives and false negatives. The formulas to calculate each score can be referred to in figure 5.1 [4].The evaluation is carried out using functions provided by Scikit-learn.

5.2 Results

In general, our systems are able to achieve quite strong performance with the minimum scores for all evaluation metrics of 0.7. The best system is SVM with word embedding features achieving a precision score as well as a recall score of 0.82, and 82% accuracy. The results for each classifier are summarized in Table 5.2.

[4] Picture borrowed from https://bit.ly/32zk4zs

	sty[1]	**w2v**[2]	**cmb**[3]	**d2v**[4]
LR	00:04	00:13	00:26	00:19
RF	00:44	03:31	03:19	02:23
SVM	02:52	11:02	21:14	38:29

[1] Stylometric features

[2] Word embedding features

[3] Combined features

[4] Document embedding features

Table 5.3: Training time for each classifier with respect to features.

The worst performance is from all classifiers with document embeddings with the precision, recall and accuracy scores are stuck at 0.7. None of the classifier can actually outperforms others when it comes to document embeddings. We note that the performance of all classifiers are quite uniform with document embedding features.

Stylometric features are a little bit more informative since the performance of classifiers with stylometric features is improved for about 5% in comparing to the performance with document embeddings. LR performs only about 1% worst than the other two classifiers on stylometric features.

Word embedding features and combined features are able to contribute most to the performance of all classifiers since using the performance of all classifiers using word embedding and combined features are improved about 5% in comparing to stylometric features and 10% in comparing to document embeddings. With both type of features, RF seems to slightly perform less effectively than other the two classifiers for about 1 or 2%. With word embeddings, LR and RF achieve the same precision of 0.80, while LR has a little better recall of 0.80 than RF. With combined features, the results from LR and SVM are the same.

So far, SVM has the best performance on all type of features, however, this algorithm requires quite an excessive amount of training time. From Table 5.3, we can see that the longest training is SVM with document embeddings. LR are really fast while still able to achieve fair results that can match the results of SVM.

5.3 Discussion

It is beyond our expectations that the classifiers are able to achieve such results since our focus is on the experimentation with new features and feature representation, rather than optimising the algorithms as we only perform basic optimisation for all algorithms.

In comparing to the results from the Clickbait Challenge, our results are able to excess the baseline created by Potthast et al. (2018a) with only 0.43 recall, 0.75 precision and 0.83 accuracy. Our best classifier outperforms the best system at that moment by Zhou (2017) using a RNN with bidirectional gated units (biGRU) and a self-attention mechanism trained on Glove word embeddings in terms of precision and recall. We notice that the system from Zhou (2017) also applies word embeddings as features for their system, however, what could make our system perform better is that we do not use pre-trained word embedding vectors. For their system,Zhou (2017) uses word embeddings trained from Wiki data which is quite disparate in comparing to the news headline data, leading to the fact that the representation in pre-trained vectors are not compatible with the context of the targeted data. Moreover, we encode the syntactic information into our word embedding representation by including the vectors of part-of-speech, and dependency tags and the word embeddings of governors, which create another layer of syntactic representation on top of word embeddings.

We also notice that many of the systems participating in the Cickbait Challenge focus on sparse features especially word-ngrams, character-ngrams, or tf-idf. The analysis of the data helps us realise that such features can be marginally uninformative for this task. We think that by presenting the headlines in a sequence of part-of-speech and dependency, as well as encoding the part-of-speech tags and dependency relation to each token representation vector, we are able to preserve the syntactic information which is proven to be essential for the clickbait detection. The successful performance of all classifiers on word embedding features also proves the power in distributional representation of texts.

What is most surprising to us is the lower results of all classifiers with document embeddings This indicates that document embeddings actually provide the least information in comparison to other features even though both stylometric and word embedding features of the headlines are included in the vectors, meaning that document embeddings outweigh other two features when combined. What could be the explanation for this plummet is that there are many noises in the contents. Examining the corpus again, we can see that other types of text are also included in the contents of both clickbait and non-clickbait, not just the bodies of the article. These texts can be headlines of other articles, or captions of pictures and videos, or comments of readers on the articles.

As we can see from Table 5.2, LR performs slightly poorer than RF with stylometric features, on the other hand, LR performs better than RF with word embeddings, and with the combined features, LR outperforms RF. We think the explanation for this behaviour is that RF performs better with discrete features, while LR performs better with embeddings which are continuous. However, the embeddings features seem to out-weight the discrete features, therefore with combined features, LR still has more advantages.

As mentioned before, there have not been any analyses on Potthast corpus which can be one of the reasons for the average performance of many systems on the corpus. We can affirm the relevance of studying and understanding the data before building any predicting systems. This part gives the answer to the last research question is that the combined of syntactic and semantic representation in word embedding features provide the most information to the system for the task of automatic clickbait detection.

Chapter 6

Conclusion

6.1 Summary

The main goal of the current study is to build an machine learning system to automatically detect clickbait using the characteristics of clickbait as features. These characteristics are selected based on the linguistic analysis of headlines and contents of clickbaits and non-clickbaits in two clickbait corpora, then the performance of there different classifiers LR, RF, and SVM with each features are evaluated in order to find which is are the most informative for the task.

This study aims at giving the relevant answers to these research questions: (1) What is clickbait and what are the linguistic characteristics that differentiate clickbait and non-clickbait articles? (2) Among these characteristics of clickbait, which potentially can become features for automatic system and how can they be represented into computational features for machine learning? (3) Which among these features is the most informative for the clickbait detection task using machine learning method?

From our analysis we conclude that a clickbaits should be a news article with a headline and content. What makes a clickbait different from a serious news article is not only use of special the syntactic structures and lexicon in its headlines but also its content. There is a chronological change in the linguistic form of clickbait in terms of semantics and syntax. Clickbaits are adapting the form of non-clickbait serious news since people might avoid some old typical clickbait topics and structures. In addition, non-clickbaits are also borrowing some characteristics of clickbaits such as using question headlines. This finding provides an explanation to unsuccessful attempt to build automatic systems to detect clickbaits using previous clickbait representation on top of the recent data. It also suggests the use of new approaches to extract features from clickbait data. Another finding from our analysis is that the contents of clickbaits are quite relevant to distinguish clickbaits from non-clickbaits. Clickbait contents might repeat some keywords from headlines leading to the decent similarity scores calculated based on term frequencies, but on a discourse level, clickbait contents fail to present more new informative and relevant information about the main idea discussed in headlines. Even though the incoherence in the contents can provide crucial cues for human to differentiate clickbaits, it might still a great challenge for machines.

Within the scope of our abilities, we experiment with different methods to represent features of clickbaits. The headlines are represented by a sequence of part-of-speech tags and dependency relations in order to preserve syntactic information. Each token of a headline is represented by a 100 dimension word embedding vector integrated with its part-of-speech tag, the embedding of its governor and their dependency relations. Finally, the contents are represented by a 100 dimension document embeddings with the representation vectors of the headlines. These vectors are fed into three different machine learning systems: Linear Regression, Random Forests, and Support Vector Machines. The results of the experiment are quite satisfactory with the best result achieved by Support Vector Machines (SVM) with word embedding features. The results also show that word embedding and encoded syntactic features can be very informative for machine learning system to learn about clickbait.

6.2 Suggestions for Further Research Work

Due to the limitation of computational power and human resources, we are not able to expand our study. Our suggestions for future researches on clickbait are: (1) building a new clickbait corpus with the most updated data and perform the same analysis on that corpus to find out if characteristics of clickbait continue to change, (2) train higher dimension word embeddings and document embeddings to see if the number of dimensions of the embeddings are able to effect the performance of the classifiers.

Bibliography

Adelson, P., Arora, S., and Hara, J. (2017). Clickbait; didn't read: Clickbait detection using parallel neural networks.

Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer*, 10:978–3.

Agrawal, A. (2016). Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 268–272. IEEE.

Alekseev, A. and Nikolenko, S. (2017). Word embeddings for user profiling in online social networks. *Computación y Sistemas*, 21(2):203–226.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anand, A., Chakraborty, T., and Park, N. (2017). We used neural networks to detect clickbaits: You won't believe what happened next! In *European Conference on Information Retrieval*, pages 541–547. Springer.

Bagosy, K., Bish, R., and Schneider, A. (2018). Baited by clickbait: Reading beyond the headlines. *The Journal of Purdue Undergraduate Research*, 8(1):12.

Beleslin, I., Njegovan, B. R., and Vukadinović, M. S. (2017). Clickbait titles: Risky formula for attracting readers and advertisers. In *XVII International Scientific Conference on Industrial Systems (IS'17) Novi Sad, Serbia*, pages 364–369.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Biswas, E., Vijay-Shanker, K., and Pollock, L. (2019). Exploring word embedding techniques to improve sentiment analysis of software engineering texts. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 68–78. IEEE.

Biyani, P., Tsioutsiouliklis, K., and Blackmer, J. (2016). " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Blom, J. N. and Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Burke, M., Kraut, R., and Marlow, C. (2011). Social capital on facebook: Differentiating uses and users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 571–580.

Cable, J. and Mottershead, G. (2018). 'can i click it? yes you can': Football journalism, twitter, and clickbait. *Ethical Space*, 15(1/2).

Cao, X., Le, T., et al. (2017). Machine learning based detection of clickbait posts in social media. *arXiv preprint arXiv:1710.01977*.

Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. (2016). 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. IEEE.

Chen, Y., Conroy, N. J., and Rubin, V. L. (2015a). Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Chen, Y., Conroy, N. J., and Rubin, V. L. (2015b). News in an online world: The need for an “automatic crap detector”. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Chen, Y. and Rubin, V. L. (2017). Perceptions of clickbait: A q-methodology approach. In *Proceedings of the 45th Annual Conference of The Canadian Association for Information Science/L'Association canadienne des sciences de l'information (CAIS/ACSI2017), Ryerson University, Toronto, May 31-June 2, 2017*.

Choi, J. D., Tetreault, J., and Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396.

Deng, L. and Liu, Y. (2018). *Deep learning in natural language processing.* Springer.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695–721.

Ecker, U. K., Lewandowsky, S., Chang, E. P., and Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied*, 20(4):323.

Eisenstein, J. (2018). Natural language processing.

Elyashar, A., Bendahan, J., and Puzis, R. (2017). Detecting clickbait in online social media: You won't believe how we did it. *arXiv preprint arXiv:1710.06699.*

Faul, A. C. (2019). *A Concise Introduction to Machine Learning.* CRC Press.

Firth, J. R. (1935). The technique of semantics. *Transactions of the philological society*, 34(1):36–73.

Gairola, S., Lal, Y. K., Kumar, V., and Khattar, D. (2017). A neural clickbait detection engine. *ArXiv*, abs/1710.01507.

Ganegedara, T. (2018). *Natural Language Processing with TensorFlow: Teach language to machines using Python's deep learning library.* Packt Publishing Ltd.

Glenski, M., Ayton, E., Arendt, D., and Volkova, S. (2017). Fishing for clickbaits in social images and texts with linguistically-infused neural network models. *arXiv preprint arXiv:1710.06390.*

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Gómez-Adorno, H., Posadas-Duran, J.-P., Ríos-Toledo, G., Sidorov, G., and Sierra, G. (2018). Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas*, 22(1):47–53.

Grigorev, A. (2017). Identifying clickbait posts on social media with an ensemble of linear models. *arXiv preprint arXiv:1710.00399.*

Gurney, K. (1997). *An introduction to neural networks.* CRC press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Heaton, J. (2015). Artificial intelligence for humans, volume 3: Neural networks and deep learning, 1.0. *Chesterfield, USA: Heaton Research Inc.*

Henderson, J. B. (2010). 9 artificial neural networks. *The Handbook of Computational Linguistics and Natural Language Processing*, page 221.

Indurthi, V. and Oota, S. R. (2017). Clickbait detection using word embeddings. *arXiv preprint arXiv:1710.02861.*

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *EuroVis (STARs)*, pages 83–103.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759.*

Kalveks, T. (2007). Clickbait. *The Blackwell Encyclopedia of Sociology*, pages 1–2.

Karaca, A. (2019). News readers' perception of clickbait news.

Kelleher, J. D., Mac Namee, B., and D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* MIT press.

Knobloch, S., Patzig, G., Mende, A.-M., and Hastall, M. (2004). Affective news: Effects of discourse structure in narratives on suspense, curiosity, and enjoyment while reading news and novels. *Communication Research*, 31(3):259–287.

Kumar, V., Khattar, D., Gairola, S., Kumar Lal, Y., and Varma, V. (2018). Identifying clickbait: A multi-strategy approach using neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1225–1228.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., and Demidov, P. (2019). A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Lai, L. and Farbrot, A. (2014). What makes you click? the effect of question headlines on readership in computer-mediated communication. *Social Influence*, 9(4):289–299.

Lary, D., Nikitkov, A., Stone, D., and Nikitkov, A. (2010). Which machine-learning models best predict online auction seller deception risk. *American Accounting Association AAA Strategic and Emerging Technologies.*

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Lex, E., Juffinger, A., and Granitzer, M. (2010). Objectivity classification in online media. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, page 293–294, New York, NY, USA. Association for Computing Machinery.

Li, Q. (2019). Clickbait and emotional language in fake news.

Lockwood, G. (2016). Academic clickbait: Articles with positively-framed titles, interesting phrasing, and no wordplay get more attention online. *The Winnower*, 3.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitchell, T. M. et al. (1997). Machine learning.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Moisl, H., Dale, R., and Somers, H. (2000). Nlp based on artificial neural networks: Introduction. In *Handbook of Natural Language Processing*. Marcel Dekker Inc.

Morville, P. and Rosenfeld, L. (2006). *Information architecture for the World Wide Web: Designing large-scale web sites*. " O'Reilly Media, Inc.".

Munger, K., Luca, M., Nagler, J., and Tucker, J. (2018). The effect of clickbait.

Munger, K., Luca, M., Nagler, J., and Tucker, J. (2020). The (null) effects of clickbait headlines on polarization, trust, and learning. *Public opinion quarterly*.

Naili, M., Chaibi, A. H., and Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112:340–349.

Obar, J. A. and Wildman, S. S. (2015). Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy*, 39(9):745–750.

Pandey, D., Verma, G., and Nagpal, S. (2018). Clickbait detection using swarm intelligence. In *International Symposium on Signal Processing and Intelligent Recognition Systems*, pages 64–76. Springer.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., and Kompatsiaris, I. (2017). A two-level classification approach for detecting clickbait posts using text-based features. *arXiv preprint arXiv:1710.08528*.

Paulau-Sampio, D. (2016). Reference press metamorphosis in the digital context: clickbait and tabloid strategies in elpais. com. *Communication & Society*, 29(2):63–79.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Potthast, M., Gollub, T., Hagen, M., and Stein, B. (2018a). The clickbait challenge 2017: towards a regression model for clickbait strength. *arXiv preprint arXiv:1812.10847*.

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., Hagen, M., and Stein, B. (2018b). Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1498–1507.

Potthast, M., Köpsel, S., Stein, B., and Hagen, M. (2016). Clickbait detection. In *European Conference on Information Retrieval*, pages 810–817. Springer.

Pujahari, A. and Sisodia, D. S. (2019). Clickbait detection using multiple categorisation techniques. *Journal of Information Science*, page 0165551519871822.

Richert, W. (2013). *Building machine learning systems with Python*. Packt Publishing Ltd.

Rogers, S. and Girolami, M. (2016). *A first course in machine learning*. CRC Press.

Rojas, R. (2013). *Neural networks: a systematic introduction*. Springer Science & Business Media.

Rony, M. M. U., Hassan, N., and Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 232–239.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Scacco, J. M. and Muddiman, A. (2016). Investigating the influence of “clickbait” news headlines. *Engaging News Project Report.*

Sisodia, D. S. (2019). Ensemble learning approach for clickbait detection using article headline features. *Informing Science: The International Journal of an Emerging Transdiscipline*, 22:031–044.

Taylor, M. (2017). Make your own neural network: An in-depth visual introduction for beginners.

Thomas, P. (2017). Clickbait identification using neural networks. *arXiv preprint arXiv:1710.08721.*

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavvaf, N., and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200.*

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.

Vijgen, B. et al. (2014). The listicle: An exploring research on an interesting shareable new media phenomenon. *Studia Universitatis Babes-Bolyai-Ephemerides*, 59(1):103–122.

Wang, S. and Wu, Q. (2017). An empirical study on the clickbait of data science articles in the wechat official accounts. In *International Conference on Frontier Computing*, pages 131–140. Springer.

Wei, W. and Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. *arXiv preprint arXiv:1705.06031.*

Wiegmann, M., Völske, M., Stein, B., Hagen, M., and Potthast, M. (2018). Heuristic feature selection for clickbait detection. *arXiv preprint arXiv:1802.01191.*

Zannettou, S., Sirivianos, M., Blackburn, J., and Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–37.

Zheng, H.-T., Chen, J.-Y., Yao, X., Sangaiah, A. K., Jiang, Y., and Zhao, C.-Z. (2018). Clickbait convolutional neural network. *Symmetry*, 10(5):138.

Zheng, H.-T., Yao, X., Jiang, Y., Xia, S.-T., and Xiao, X. (2017). Boost clickbait detection based on user behavior analysis. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pages 73–80. Springer.

Zhou, Y. (2017). Clickbait detection in tweets using self-attentive network. *arXiv preprint arXiv:1710.05364*.

Appendix A

(Example of Clickbait and Non-clickbait from Potthast Corpus)

1. **"ISIS captures 88 Eritrean Christians in Libya, US official confirms"**

   Label: **Non-clickbait**

   ISIS terror group kidnapped 88 Eritrean Christians from a people-smugglers' caravan in Libya last week, a U.S. defense official confirmed Monday. The defense official confirmed initial reports of the mass kidnapping to Fox News after seeing a recent intelligence report. The independent Libya Herald newspaper reported that the convoy was ambushed by militants south of Tripoli before dawn this past Wednesday morning. Meron Estafanos, the co-founder of the Stockholm-based International Commission on Eritrean Refugees, told the paper that the group of migrants included "about 12 Eritrean Muslims and some Egyptians. They put them in another truck and they put 12 Eritrean women Christians in a smaller pick-up". Estafanos said that the militants had initially stopped the truck and demanded that the Muslims on board make themselves known. Everyone who responded was asked about the Koran and their religious observance in an attempt to catch Christians pretending to be Muslims. The main body of the group was put back on the original truck. As the militants drove the vehicle away, Britain's Daily Telegraph reported that at least nine men attempted to escape by diving off the back of the truck. Estefanos said three of those who had escaped were safe, but still trying to get out of Libya. The fate of the others was not known. Libya has become a jumping-off point for thousands of migrants from the Middle East and sub-Saharan Africa who attempt the dangerous Mediterranean crossing to southern Europe. However, Libya's ongoing instability has led to an increased presence by ISIS and other terror groups, increasing the risk for Christians and other non-Muslims attempting the crossing. In February, Libyan militants proclaiming loyalty to ISIS released a video showing the beheading of 21 Egyptian Coptic Christians at the edge of the Mediterranean Sea. Two months later, another video showed the militants shooting and beheading an indeterminate number of Ethiopian Christians. Estefanos told the Libya Herald that the video released in April had been edited and that 64 people had been massacred, including several Eritreans. "Ever since the kidnapping by ISIS in Libya last February," she said, "many are taking different routes. Some go from Khartoum [Sudan] to Turkey, then Greece. Others are now leaving via Khartoum to Cairo, then Alexandria and from there by boat to Italy. I think we will see an increase towards Turkey and Cairo instead of Libya". ISIS on Tuesday also claimed that it seized a power plant near the Libyan city of Sirte, which supplies central and western parts of the country with electricity, Reuters reports. "The plant ... was taken," ISIS said in a message on social media, while forces loyal to the self-declared government that controls Libya's capital, Tripoli, fled the area, a military source told Reuters. The source said three soldiers were killed in the attack. Libya is divided between rival governments and hundreds of militias in the aftermath of its 2011 civil war that ousted dictator Muammar Qaddafi. The violence has impacted the country's oil revenues heavily. U.N. envoy to Libya Bernardino Leon has warned that the country only has enough money to pay salaries for another six weeks, urging warring parties to agree on a unity government. Negotiators are currently meeting in Morocco to discuss a power-sharing agreement. Fox News' Lucas Tomlinson and the Associated Press contributed to this report.

2. **"Anarchy at the checkout: You can now own a Sex Pistols credit card"**

Label: **Clickbait**

LONDON — “Your future dream is a shopping scheme.” Nearly 40 years after Johnny Rotten spat those angry lines out to an unsuspecting audience as punk's disenfranchised trailblazers unleashed "Anarchy In The UK" on the world, the artwork for the single is being featured on a credit card. See also: Artist re-envisions post-punk rockers as popular Marvel superheroes 38 years after the Pistols signed to Virgin Records, another arm of Richard Branson's sprawling business empire has snapped up the imagery for the single, and also the artwork for their infamous album Never Mind The Bollocks, Here's The Sex Pistols, to illustrate consumer credit. Virgin Money's new cards have even shifted the customer details onto the back, so your name, number and expiry date won't intrude on the classic designs. "The Sex Pistols are an iconic band and an important part of Virgin's history," Branson said. "Even after nearly 40 years, the Sex Pistols' power to provoke is undimmed, and we are still being asked to censor the word 'bollocks' in our advertising. Over the years many things have changed, but in this case some attitudes clearly do not." Rotten's thoughts on the cards are unknown, but the LA-based singer - who famously shilled butter a few years back - is probably more concerned with the forthcoming PiL album and tour. The details of the deal are not known, but The Telegraph reports that the band came to a commercial arrangement with Virgin Money for an undisclosed sum of money. "We don't want Anarchy in banking – but we do want change," Virgin Money's Jayne-Anne Gadhia said with a straight face. "And we want to get rid of the bollocks in banking and to be simple, open, transparent and fair." Most of the cards have a representative APR of around 18.9%. Here's what Twitter thought of the cards. Don't know what I want but I know how to get it and pay it off in monthly installments #sexpistols — Justin Horton (@ejhchess) June 9, 2015 "They're selling hippie wigs in Woolworths, man" pic.twitter.com/xZMpt3jfJb — RamAlbumClub (@RamAlbumClub) June 9, 2015 To be fair, bankers are some of the most legit anarchists out there. #sexpistols @VirginMoney — lesmondine (@lesmondine) June 9, 2015 I AM AN ACTUARIST! #sexpistols @VirginMoney — lesmondine (@lesmondine) June 9, 2015 It's great to see Virgin Money acknowledging the Sex Pistols. Thank you @VirginMoney. http://t.co/EDTe7mvsDn — Sex Pistols Official (@pistolsofficial) June 9, 2015 The cards are available now.

Appendix B

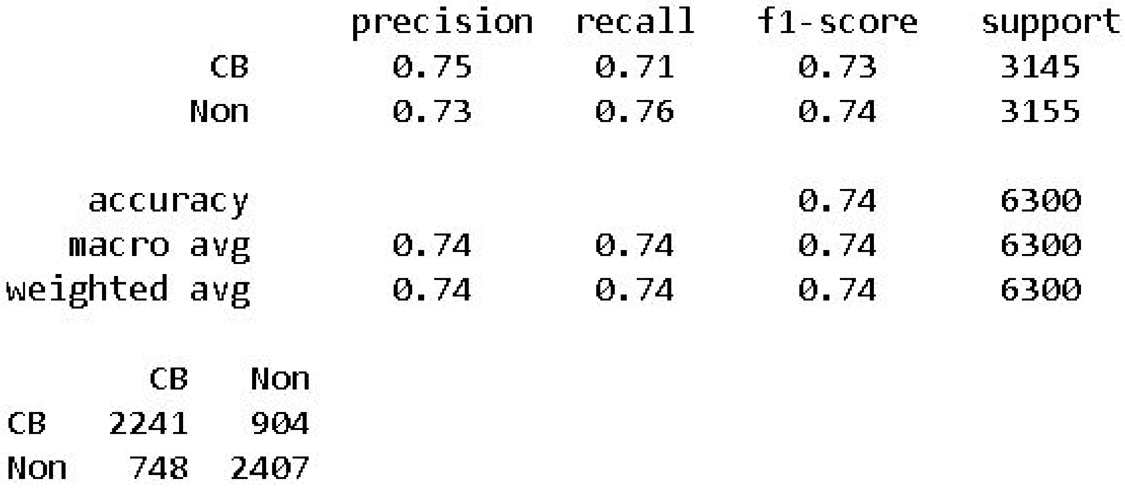
(Report on performance results of each classifier in respect to features)

B.1 Stylometric Features

B.1.1 Logistic Regression

Best parameters: C: 10, solver: liblinear

Accuracy: 0.7377777777777778



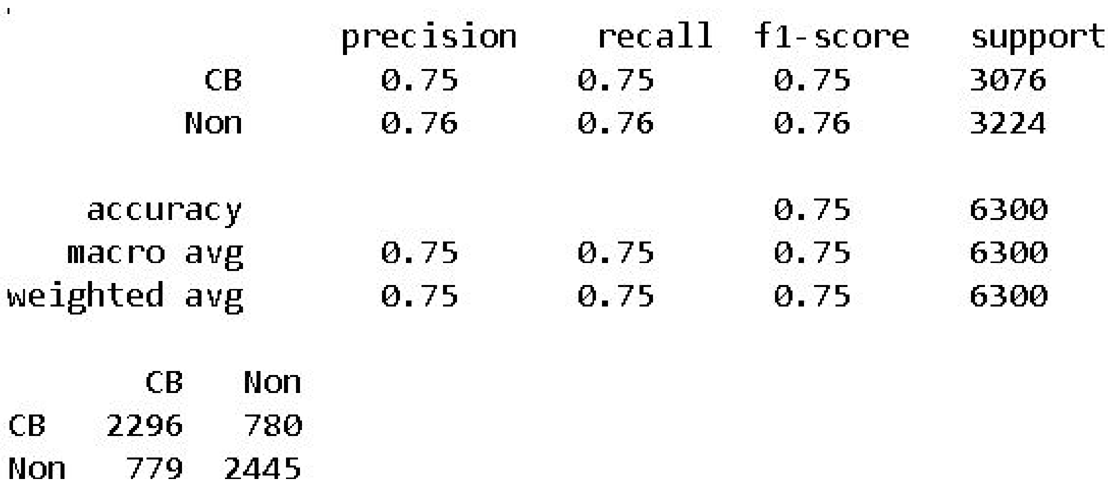
```
              precision    recall  f1-score   support
          CB       0.75      0.71      0.73      3145
         Non       0.73      0.76      0.74      3155

    accuracy                           0.74      6300
   macro avg       0.74      0.74      0.74      6300
weighted avg       0.74      0.74      0.74      6300

       CB   Non
CB   2241   904
Non   748  2407
```

Figure B.1: Result report for LR classifier and stylometric features

B.1.2 Random forest

Best parameters: criterion: entropy, max-depth: 6, max-features: sqrt, estimators: 50

Accuracy: 0.7525396825396825



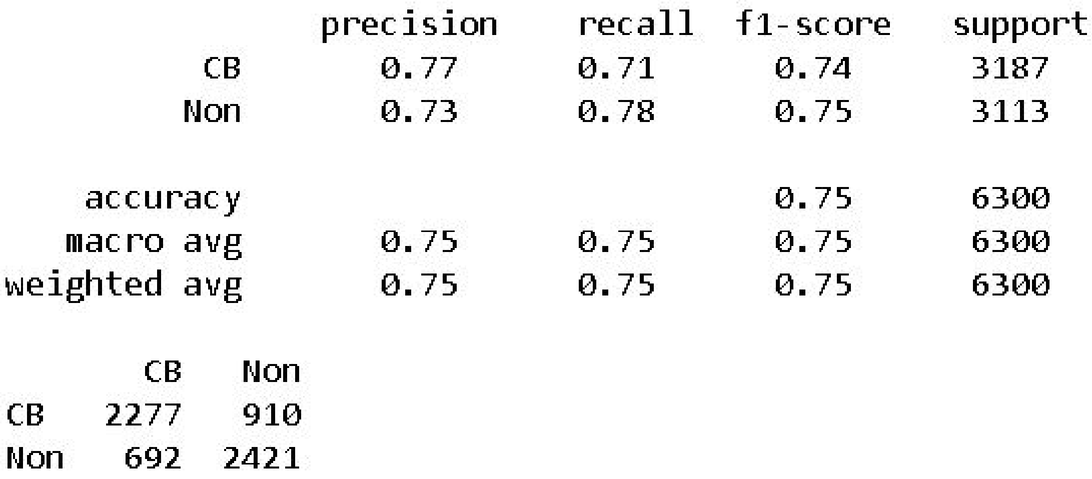
```
              precision    recall  f1-score   support
          CB       0.75      0.75      0.75      3076
         Non       0.76      0.76      0.76      3224

    accuracy                           0.75      6300
   macro avg       0.75      0.75      0.75      6300
weighted avg       0.75      0.75      0.75      6300

       CB   Non
CB   2296   780
Non   779  2445
```

Figure B.2: Result report for RF classifier and stylometric features

B.1.3 Support Vector Machine

Best prarameters: eta0: 0.1, learning-rate: adaptive, loss: hinge, penalty: l2, tol: 0.001

Accuracy:0.7457142857142857



```
              precision    recall  f1-score   support
          CB       0.77      0.71      0.74      3187
         Non       0.73      0.78      0.75      3113

    accuracy                           0.75      6300
   macro avg       0.75      0.75      0.75      6300
weighted avg       0.75      0.75      0.75      6300

       CB   Non
CB   2277   910
Non   692  2421
```

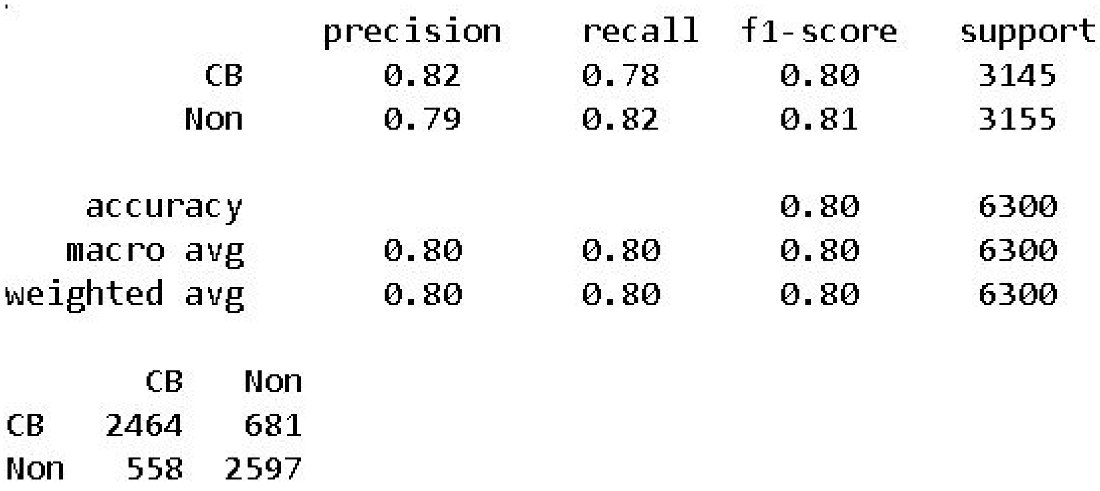
Figure B.3: Result report for SVM classifier and stylometric features

B.2 Word embedding Features

B.2.1 Logistic Regression

Best parameters: C: 10, solver: liblinear

Accuracy: 0.8033333333333333



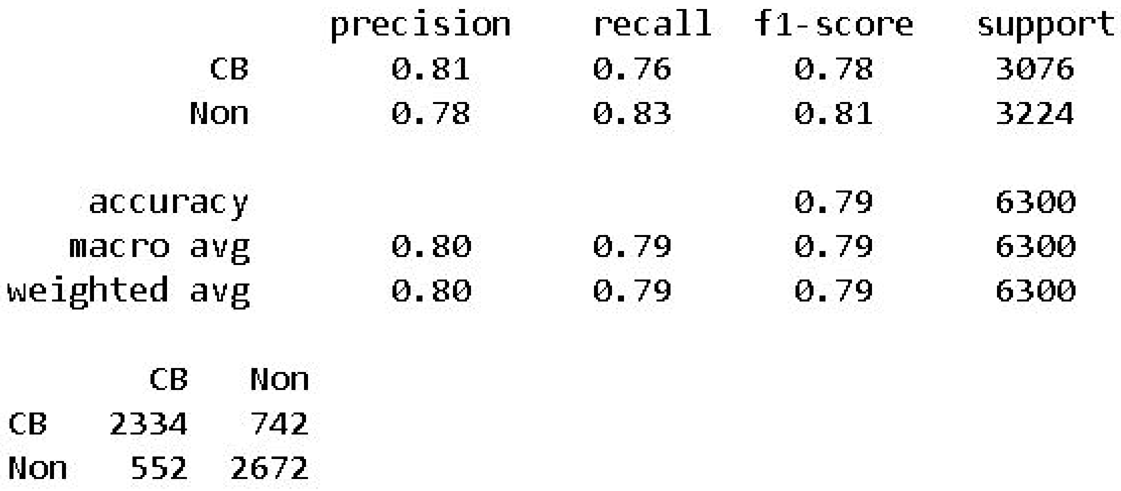
```
              precision    recall  f1-score   support
          CB       0.82      0.78      0.80      3145
         Non       0.79      0.82      0.81      3155

    accuracy                           0.80      6300
   macro avg       0.80      0.80      0.80      6300
weighted avg       0.80      0.80      0.80      6300

       CB   Non
CB   2464   681
Non   558  2597
```

Figure B.4: Result report for LR classifier and word embedding Features

B.2.2 Random forest

Best parameters: criterion: gini, max-depth: 6, max-features: log2, estimators: 100

Accuracy: 0.7946031746031746



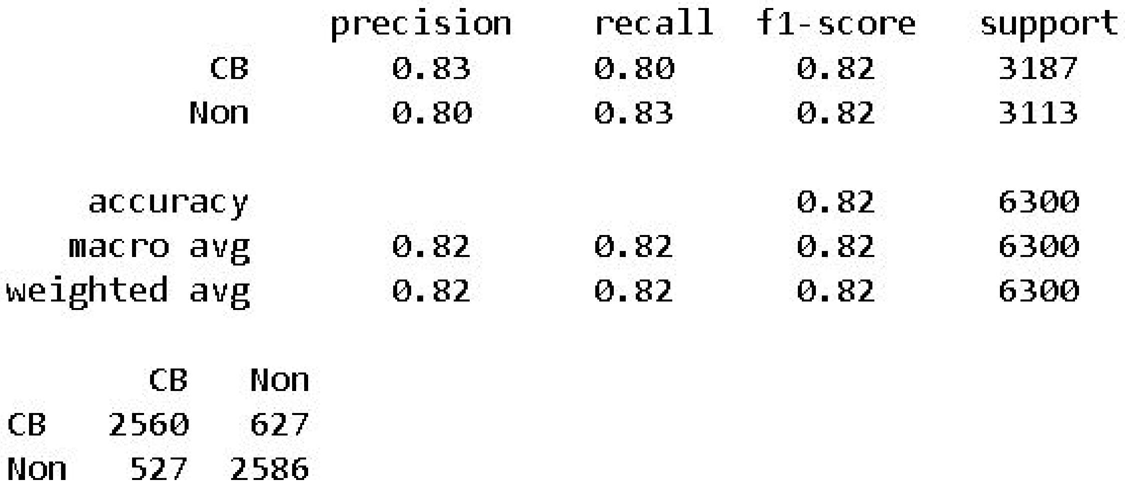
```
              precision    recall  f1-score   support
          CB       0.81      0.76      0.78      3076
         Non       0.78      0.83      0.81      3224

    accuracy                           0.79      6300
   macro avg       0.80      0.79      0.79      6300
weighted avg       0.80      0.79      0.79      6300

       CB   Non
CB   2334   742
Non   552  2672
```

Figure B.5: Result report for RF classifier and word embedding Features

B.2.3 Support Vector Machine

Best parameters: eta0: 0.1, learning-rate: adaptive, loss: hinge, penalty: l1, tol: 0.1

Accuracy:0.8168253968253968



```
              precision    recall  f1-score   support
          CB       0.83      0.80      0.82      3187
         Non       0.80      0.83      0.82      3113

    accuracy                           0.82      6300
   macro avg       0.82      0.82      0.82      6300
weighted avg       0.82      0.82      0.82      6300

       CB   Non
CB   2560   627
Non   527  2586
```

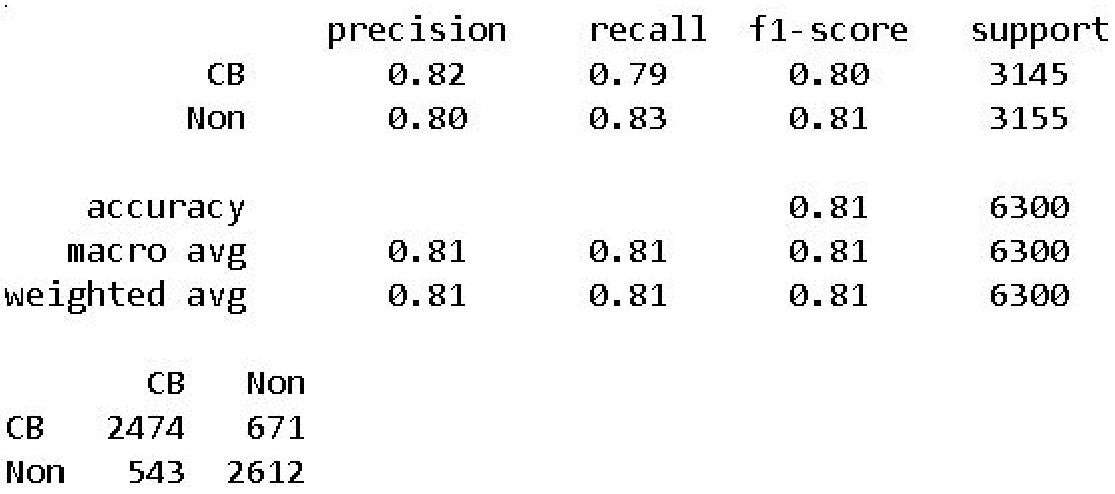
Figure B.6: Result report for SVM classifier and word embedding features

B.3 Combined features

B.3.1 Logistic Regression

Best parameters: C: 0.1, solver: liblinear

Accuracy: 0.8073015873015873



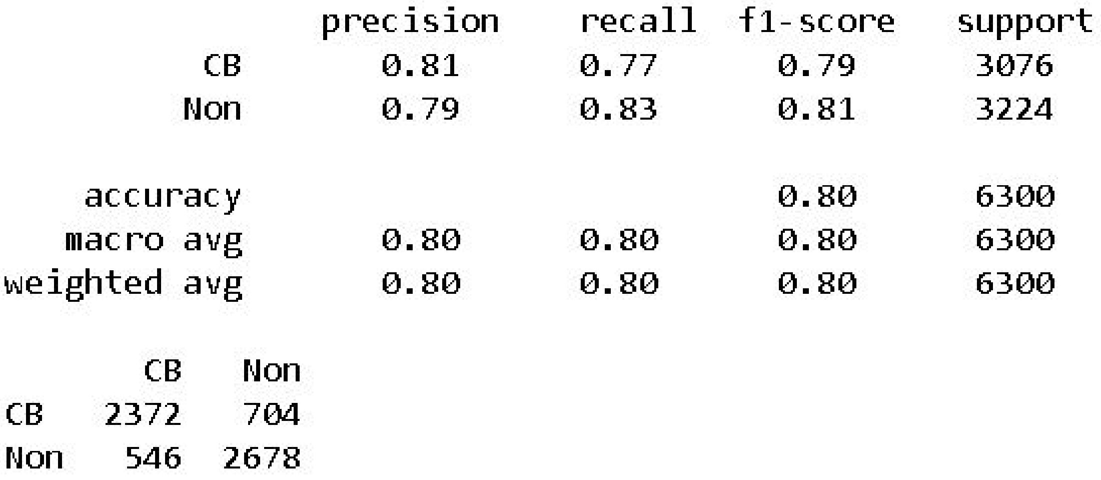
```
              precision    recall  f1-score   support
          CB       0.82      0.79      0.80      3145
         Non       0.80      0.83      0.81      3155

    accuracy                           0.81      6300
   macro avg       0.81      0.81      0.81      6300
weighted avg       0.81      0.81      0.81      6300

       CB   Non
CB   2474   671
Non   543  2612
```

Figure B.7: Result report for LR classifier and combined features

B.3.2 Random forest

Best parameters: criterion: entropy, max-depth: 6, max-features: log2, estimators: 150

Accuracy: 0.8015873015873016



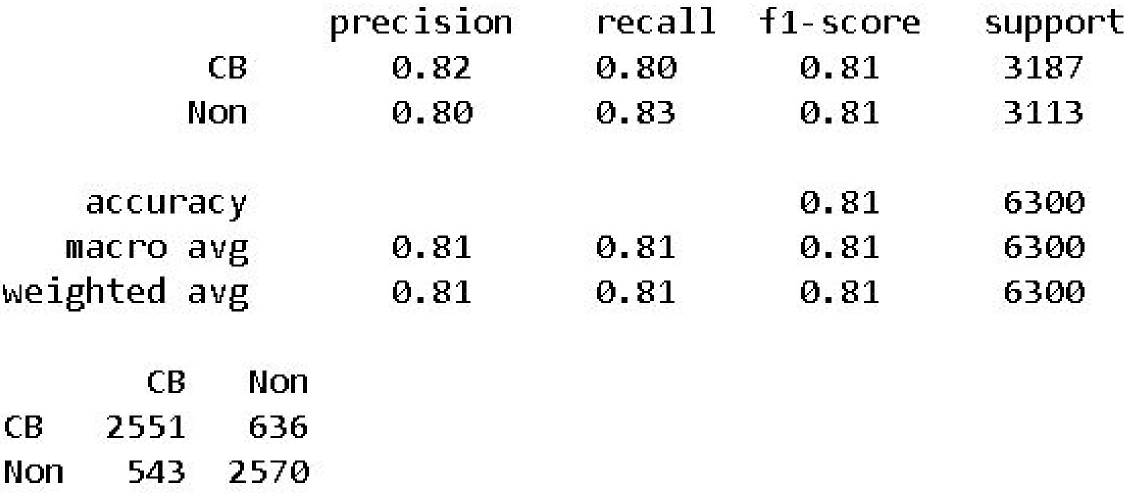
```
              precision    recall  f1-score   support
          CB       0.81      0.77      0.79      3076
         Non       0.79      0.83      0.81      3224

    accuracy                           0.80      6300
   macro avg       0.80      0.80      0.80      6300
weighted avg       0.80      0.80      0.80      6300

       CB   Non
CB   2372   704
Non   546  2678
```

Figure B.8: Result report for RF classifier and combined features

B.3.3 Support Vector Machine

Best prarameters: eta0: 0.001, learning-rate: constant, loss: hinge, penalty: l1, tol: 0.001

Accuracy:0.8128571428571428



```
              precision    recall  f1-score   support
          CB       0.82      0.80      0.81      3187
         Non       0.80      0.83      0.81      3113

    accuracy                           0.81      6300
   macro avg       0.81      0.81      0.81      6300
weighted avg       0.81      0.81      0.81      6300

       CB   Non
CB   2551   636
Non   543  2570
```

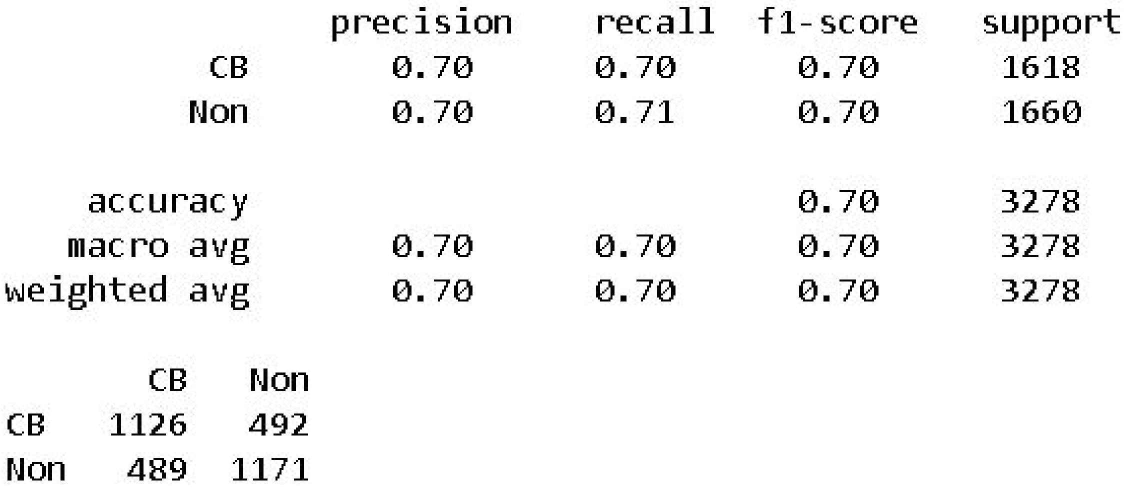
Figure B.9: Result report for SVMne classifier and combined features

B.4 Document embedding features

B.4.1 Logistic Regression

Best parameters: C: 100, solver: lbfgs

Accuracy: 0.700732153752288



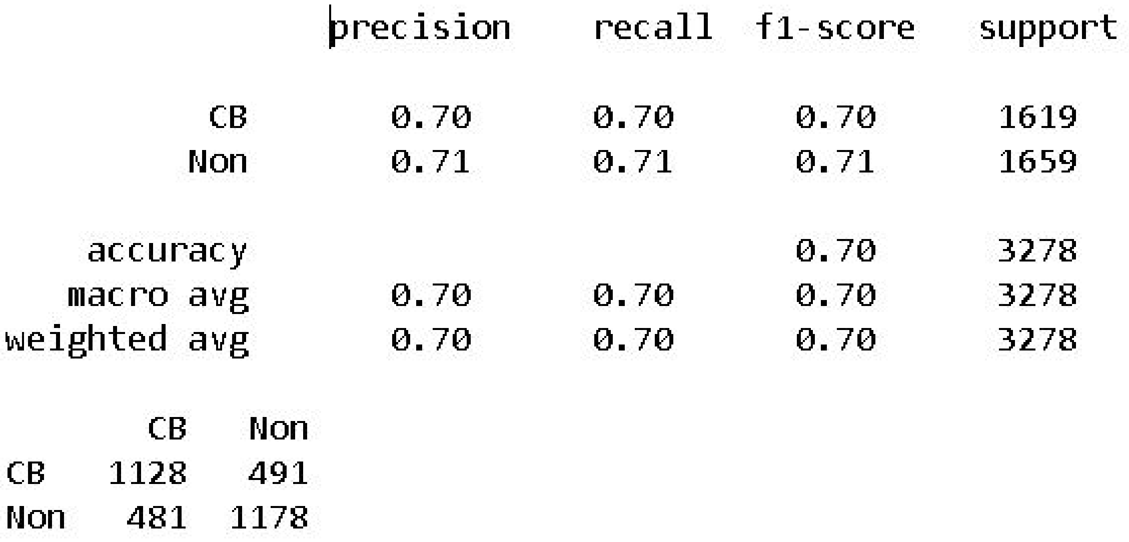
```
              precision    recall  f1-score   support
          CB       0.70      0.70      0.70      1618
         Non       0.70      0.71      0.70      1660

    accuracy                           0.70      3278
   macro avg       0.70      0.70      0.70      3278
weighted avg       0.70      0.70      0.70      3278

       CB   Non
CB   1126   492
Non   489  1171
```

Figure B.10: Result report for LR classifier and document embedding features

B.4.2 Random forest

Best parameters: criterion: entropy, max-depth: 6, max-features: sqrt, estimators: 150

Accuracy: 0.7034777303233679



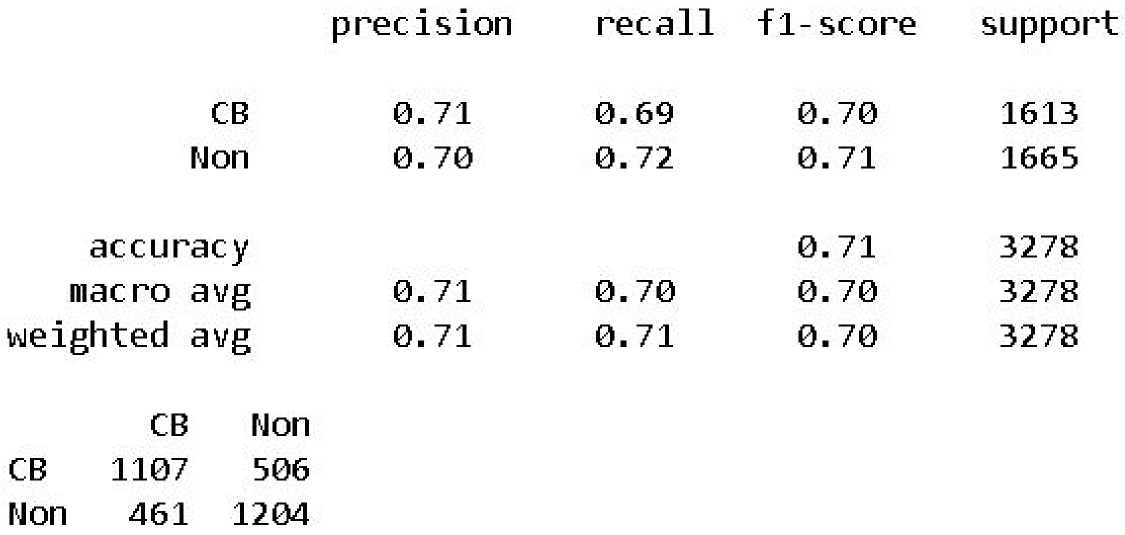
```
              precision    recall  f1-score   support

          CB       0.70      0.70      0.70      1619
         Non       0.71      0.71      0.71      1659

    accuracy                           0.70      3278
   macro avg       0.70      0.70      0.70      3278
weighted avg       0.70      0.70      0.70      3278

       CB   Non
CB   1128   491
Non   481  1178
```

Figure B.11: Result report for RF classifier and document embedding features

B.4.3 Support Vector Machine

Best prarameters: eta0: 0.01, learning-rate: invscaling, loss: hinge, penalty: l2, tol: 0.01

Accuracy:0.7050030506406345



```
              precision    recall  f1-score   support

          CB       0.71      0.69      0.70      1613
         Non       0.70      0.72      0.71      1665

    accuracy                           0.71      3278
   macro avg       0.71      0.70      0.70      3278
weighted avg       0.71      0.71      0.70      3278

       CB   Non
CB   1107   506
Non   461  1204
```

Figure B.12: Result report for SVM classifier and document embedding features