The Title of Your Dissertation

Jacob Barrow - 40337360

Submitted in partial fulfilment of
the requirements of Edinburgh Napier University
for the Degree of
BSc (Hons) Computing Science

School of Computing

October 10, 2020

Authorship Declaration

I, Jacob Barrow, confirm that this dissertation and the work presented in it are my own achievement.

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 dissertation is entirely my own work;

I have acknowledged all main sources of help;

If my research follows on from previous work or is part of a larger collaborative research project I have made clear exactly what was done by others and what I have contributed myself;

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have obtained informed consent from all people I have involved in the work in this dissertation following the School's ethical guidelines.

Signed:

Date:

Matriculation no:

General Data Protection Regulation Declaration

Under the General Data Protection Regulation (GDPR) (EU) 2016/679, the University cannot disclose your grade to an unauthorised person. However, other students benefit from studying dissertations that have their grades attached.

Please sign your name below one of the options below to state your preference.

The University may make this dissertation, with indicative grade, available to others.

The University may make this dissertation available to others, but the grade may not be disclosed.

The University may not make this dissertation available to others.

Abstract

Contents

List of Tables

List of Figures

Acknowledgements

Insert acknowledgements here

I would like to thank...

1 Background Research

1.1 Types of Incongruence

Chesney, Liakata, Poesio, and Purver (2017) talks about the differences between clickbait, fake news, sensationalism and incongruent headlines.

1.1.1 Clickbait

Potthast, Köpsel, Stein, and Hagen (2016) define clickbait as a kind of "web content [...] designed to entice its readers into clicking an accompanying link". Clickbait uses exagerated language, outright fake information and can be accompanied by graphics designed to entice a reader. Figure 1 shows an example of clickbait, sourced from a Natural Health website [1].

RELATED POSTS

12 Personal Hygiene Mistakes Almost Everyone Makes (Mom Never Told You About #4!)

12 Mind-Blowing Benefits Of Drinking Coconut Water During Pregnancy

Do You Have Gluten Sensitivity? 10 Common Symptoms

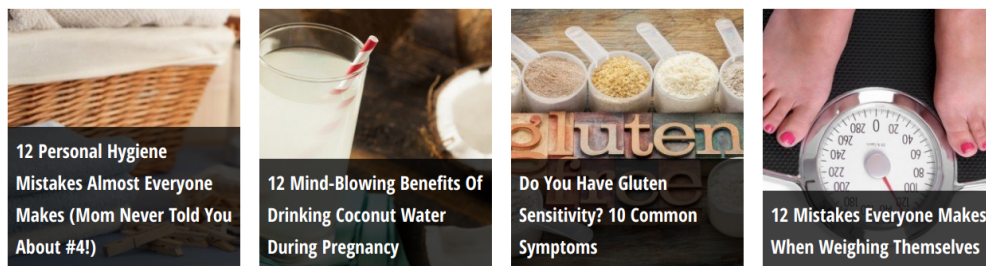12 Mistakes Everyone Makes When Weighing Themselves

Figure 1: Several clickbait articles in a 'chum box'

Mahoney (2015) terms a collection of clickbait stories as a 'chum boxes' - chum being dead fish used as bait for other fish. Mahoney goes on to examine how clickbait uses psychological methods to manipulate, an how they can have an unconcious effect on an individual.

1.1.2 Fake News

Allcott and Gentzkow (2017) defines fake news to be "news articles that are intentionally and verifiably false, and could mislead readers". For example, a fake news conspiracy theory claimed that a pizzeria, Comet Ping Pong, in Washington ran a child sex ring in its basement. Figure 2 shows a news article from 2016 from Your News Wire[2] (now News Punch).

---

[1] https://naturalon.com/
[2] https://archive.is/YTk3n

Figure 2: A fake news story

This lead to a man walking into Comet Ping Pong with an assult rifle and firing several shots. The restauraunt's owner and staff also recieved several death threats (Lopez, 2016).

Allcott and Gentzkow go further in their definition, and give the following sub-categories for fake news: satire, parody, fabrication, manipulation, advertising and propoganda. While the intention of satire and parody is not to decieve but to criticise, the other classifications have more subversive aims, such as misinforming people or gaining as many clicks as possible.

1.1.3 Sensationalism

Molek-Kozakowska (2013) defines sensationalism as "a specific discourse strategy aimed at channeling audience's attention, which may well be resorted to by both popular and quality outlet". They suggest that media fails to provide important and valuable news, in preference for that which is superfical and quick-paced.

Unlike clickbait headlines, information is not withheld but rather dramatised - while the aim is still to get as many clicks as possible, this is achieved through different means.

These headlines, all sourced from The Sun, are examples of sensationalised headlines:

- "DOOMSDAY DISEASE FEARS Terrorists could turn 'sniff and die'

virus that kills victims in 24 hours into a BIO-WEAPON"

- "SPICE UP YOUR LIFE Chilli and ginger 'slash the risk of cancer – stopping tumours growing'"

- "JAB DEBATE As Melinda Messenger slams the HPV jab the parents of two teenagers blame their daughters' 'paralysis on vaccine'"

- "'I KNOW WHO KILLED JONBENET' Juror from the JonBenet Ramsey case gives sensational interview revealing he 'knows who killed six-year-old'"

They use dramatic language ('slams', 'sensational', 'slash') to evoke a sense of urgency and excitement in the reader, urging them to click through to the rest of the article.

1.1.4   Project scope

This project will not consider fake news - by it's nature, the entirety of a fake news article will be false, not just the headline. Therefore, in order to determine whether an article is fake, external sources would have to be consulted. Creating an algorithm for the truth, while an open problem in computer science[3], is considered out of the scope of this project.

Instead, this study will seek to evaluate to what extent a headline represents an article's body. This could identify sensationalism, over exaggerated news stories and potentially some types of clickbait.

1.2   Impact of Incongruence

1.2.1   Manufacturing Outrage

1.2.2

For people who read beyond the headlines - Ecker, Lewandowsky, Chang, and Pillai (2014) ran experiments that investigated how headlines effect the processing of the facts in news "Information that is initially accepted as valid but is later foundto be incorrect can have a persistent influence on people's memoryand reasoning"

---

[3]`https://www.youtube.com/watch?v=leX541Dr2rU`

1.3   Existing Approaches

A range of studies have already considered several aspects of this project.

Manjesh, Kanakagiri, Vaishak, Chettiar, and Shobha (2017) used Natural Language Processing (NLP) to identify 'clickbait' headlines, however they did not consider the headline's relationship to the article.

2 Data Collection

Chesney et al. (2017) covers the need for a decent labelled dataset.

2.1 Attributes

Before collecting the data, it's important to decide what form it'll take and what attributes will be stored.

As the aim of the project is to identify incongruence between an articles headline and body, these two attributes will be included in the dataset. In order to identify trends and allow for further analysis, the article's date of publication and the publisher (e.g. BBC, The Guardian, etc.) will also be stored.

Collection could have gone further and retained the articles category (e.g. 'politics', 'sport' etc.), but different publishers categorise articles in different ways - for instance, the BBC has a combined 'Science and Environment' category, whereas The Guardian splits these into two distinct categories. Additionally, similar news articles can be filed under different categories, depending on the publisher. As this project's focus is on the article's content, and not categorisation, it can be considered out of scope to investigate the interplay between different publisher's approach to categorising articles.

2.2 Sources

The Independent is one of the only online publishers to make available their entire archives. Using the methods mentioned in Section 2.3, XXX articles were collected, from 2011 to 2020. This 9-year period should prove a useful dataset to analyse a potentially changing landscape in the congruity of news headlines.

The BBC has an 'On This Day' page[4] that has a very select archive from 1950-2005, and analysing these articles could produce some interesting results. However, each of these articles will have been hand-picked (as evidenced by the 'In Context' notes alongside each article), and only represent historic world news events. Therefore, these articles will not be a suitable representation for the average of the time period they are from.

As well as archives, current news was also collected from a range of publishers. A varied range of UK publishers were selected, in order to create a cross

---

[4]`http://news.bbc.co.uk/onthisday`

Table 1 shows the full list of data sources collected, as well as the time range they cover and the total records obtained.

Publisher	Earliest	Latest	Total
BBC (Archive)	1950-01-21	2005-12-11	1857
The Independent (Archive)	2011-01-01	XXX	XXX
BBC (Daily)	2019-04-18	XXX	XXX
Daily Mail (Daily)	2019-09-24	XXX	XXX
The Guardian (Daily)	2020-06-25	XXX	XXX
Huffington Post (Daily)	XXX	XXX	XXX

Table 1: Extents of the data sources collected

2.3   Obtaining the Data

To collect the data, several Python scripts were created. For the daily news, the publishers' various RSS feeds were consulted, and for the archives a more customised approach was taken.

These scripts utilise the BeautifulSoup library to parse each article's webpage and scrape them for the headline, date and body text. As each publisher builds their websites using different design patterns and with different technologies, each script had to be tailor made to fit the page structure. All the scripts used are available in this project's GitHub repository[5].

In addition, some sites implemented a strict rate-limit on requests - to make a copy of The Independent's archive took around XXX days to complete, scraping one article every 15 seconds.

2.4   Ethics

Across a variety of datasets, XXX articles were collected for analysis. This is a substantial amount of data, and represents the work of many individual journalists and news publishers.

While automated techniques were used to collect the data, everything collected was publically accessiable. In addition, it is legal to make a digital copy of copyrighted data for non-commercial research [6]. Even so, care still needs to be taken in the obtainment of the data in order to avoid overloading or altering the regular service of these archives. As mentioned above, requests

---

[5]https://github.com/jacobbarrow/honours/tree/master/data-collection
[6]https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research

were rate-limited to avoid inadvertant denial of service attack, and spread out over a long period of time. Additionally, the rolling news was only collected once per day, at midnight, in order to minimise the impact of the scraping.

References

Allcott, H., & Gentzkow, M. (2017, May). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. Retrieved from `https://doi.org/10.1257/jep.31.2.211` doi: 10.1257/jep.31.2.211

Chesney, S., Liakata, M., Poesio, M., & Purver, M. (2017). Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 emnlp workshop: Natural language processing meets journalism* (pp. 56–61).

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

Lopez, G. (2016, Dec). *A man fired shots at a dc pizzeria while "investigating" a bizarre fake news conspiracy theory.* Vox. Retrieved from `https://www.vox.com/policy-and-politics/2016/12/5/13839178/comet-fake-news-pizzagate-gunman`

Mahoney, J. (2015, Jun). *A complete taxonomy of internet chum.* Retrieved from `https://www.theawl.com/2015/06/a-complete-taxonomy-of-internet-chum/`

Manjesh, S., Kanakagiri, T., Vaishak, P., Chettiar, V., & Shobha, G. (2017). Clickbait pattern detection and classification of news headlines using natural language processing. In *2017 2nd international conference on computational systems and information technology for sustainable solution (csitss)* (p. 1-5).

Molek-Kozakowska, K. (2013). Towards a pragma-linguistic framework for the study of sensationalism in news headlines. *Discourse & Communication*, *7*(2), 173–197.

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait detection. In *European conference on information retrieval* (pp. 810–817).

Appendices

A   Project Overview

A.A   Example sub appendices

...

B   Second Formal Review Output

Insert a copy of the project review form you were given at the end of the review by the second marker

C   Diary Sheets (or other project management evidence)

Insert diary sheets here together with any project management plan you have

D   Appendix 4 and following

insert content here and for each of the other appendices, the title may be just on a page by itself, the pages of the appendices are not numbered, unless an included document such as a user manual or design document is itself pager numbered.