Detecting Incongruous News Articles

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 28, 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

1 Introduction

1.1 Project Outline

The project will consist of a range of deliverables and milestones.

Background Research A literature review will be undertaken to provide context to the project and define its scope

Data collection In order to provide the algorithm with articles to classify, a dataset is required. It will be sourced from a range of publishers, and ideally cover a large timescale to allow trends to be identified.

Data Labelling A raw dataset can be made more valuable by creating a training set from it. This will be a small subsection of the articles that volunteers will label with their perception of the incongruence.

Algorithm Creation Once a dataset is generated and labelled, the algorithm to classify the articles can be implemented.

Analysis The algorithm will create a range of datapoints that can be analysed. Trends will be spotted in the dataset used, and comparisons made with different implementations of the algorithm.

Discussion ...

2 Background Research

2.1 Introduction

This study aims to create a method of detecting incongrouence in news articles. Before the implementation begins, it's important to review the existing literature to give the study context.

First, this research begins by defining different types of incongruence and specifying the bounds of incongruence applicable to this study.

Several existing approaches are then evaluated and discussed to give a clearer picture of both what's already been done, but also to gain some insight into a possible approach to tackle the problem.

Then, natural language processing (NLP) is defined and different features reviewed.

2.2 Types of Incongruence

Incongruence is a broad term that, when applied to news media, covers a lot of different forms of deception and misleading information. Chesney, Liakata, Poesio, and Purver (2017) classifies three different types of incongruent news articles: clickbait, fake news, and sensationalism.

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 exaggerated 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].

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, and how they can have an unconscious effect on an individual.

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`

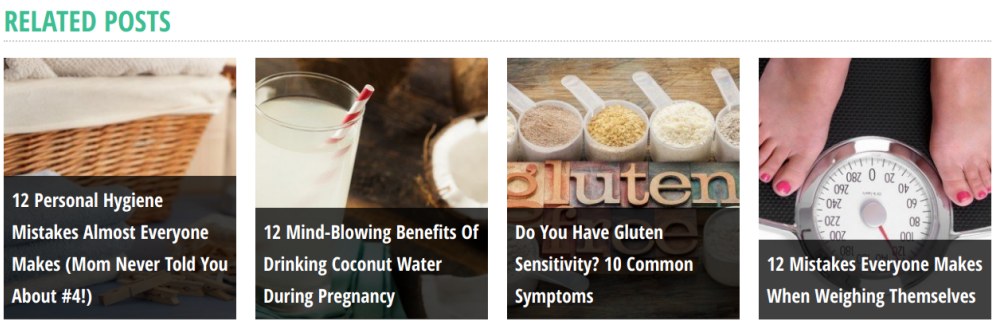

Figure 1: Several clickbait articles in a 'chum box'

Figure 2: A fake news story

This lead to a man walking into Comet Ping Pong with an assault rifle and firing several shots. The restaurant's owner and staff also received 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 propaganda. While the intention of satire and parody is not to deceive but to criticise, the other classifications have more subversive aims, such as misinforming people or gaining as many clicks as possible.

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 superficial and quick-paced.

Below are examples of sensationalised headlines, sourced from The Sun:

- 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'

These headlines 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. 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.

This sensationalism is intended to provoke and entertain, at times at the expense of accuracy (Chesney et al., 2017).

2.2.1 Project scope

This project will not consider fake news - by its nature, the entirety of a fake news article will be false, not just the headline. Therefore, 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.

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

2.3 Impact of Incongruence

Karlsson and Strömbäck (2010) categorise online news both by its immediacy and interactivity, which has shortened the news cycle and increased the competitivity between publishers. Therefore, publishers have to make the news more appealing to potential consumers.

In the information-overload arena of online news reporting, the body of a news article is less read than the headline (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016).

For those that read beyond the headlines, incongruent articles can still be problematic; it's a well-established theory in psychology that first opinions matter (Digirolamo & Hintzman, 1997). Ecker, Lewandowsky, Chang, and Pillai (2014) ran a study that investigated how headlines affect the processing of the facts in news "Information that is initially accepted as valid but is later found to be incorrect can have a persistent influence on people's memory and reasoning". Publishers can seek to sway individuals by using choice phrases to influence their mindset, which means that the same content could be interpreted in many ways depending on its headline (Reis et al., 2015).

This means that if a headline is incongruent, even if the individual reads the whole article, there's a real possibility they will be left with a false impression of the facts.

2.4 Existing Approaches

Manjesh, Kanakagiri, Vaishak, Chettiar, and Shobha (2017) used a range of different techniques to identify clickbait and were able to achieve a 98% accuracy with a deep learning approach. However, they only analysed the article's headline and disregarded the body text. They found that clickbait headlines tend to have elaborate sentences with various linguistic nuances, such as "21 Pics Of Celebs Photoshopped In The Best Way Ever. These Are EPIC". There's also a statement at the end to further strengthen the main claim of the headline.

Park, Kim, Yoon, Cha, and Jung (2020) used a deep learning approach to create a web interface for detecting incongruent articles. They managed to gain an accuracy of 86%. However, for their dataset, they generated incongruent articles by swapping a completely different article's text in for the original. For example Headline A would have a section of Article B's body. They then considered congruent headlines to be those with the original body text in place. This could lead to false positives, and as the manufactured dataset does not reflect the incongruence in real-world articles their algorithm's output lacks validity.

2.5 Natural Language Processing

Natural language processing (NLP) is a method of extracting information from a spoken or written language. 'Natural' here means the more free and less well defined human language, as opposed to strictly interpreted programming and mathematical notation. (Jackson, 2002)

As natural language is filled with a range of nuances, assumptions and relies heavily on context, codifying it into a standardised, programmatic output poses a range of difficulties. For example, consider the following two sentences:

- Apple's shares fell by 10% in the last quarter
- An apple a day keeps the doctor away

The word 'apple' appear in both sentences, but in one it refers to a multinational company, and in the other a tasty fruit. It is only by using the context clues in the surrounding sentence that the meaning of the word can be deduced.

There are a range of different approaches that seek to tackle the problems inherent in determining the meaning and sentiment of natural language, each with their characteristics, strengths and limitations.

2.5.1 Statistical NLP

Statistical NLP creates metadata from a sentence and aims to extract meaning by using statistical inference (Manning & Schutze, 1999). Several techniques and models are used to create and interpret the metadata.

Tokens A token is typically an alphanumeric string or a punctuation mark. For instance, the sentence "Is this the way to Amarillo?" could be tokenised (represented as a list of tokens) like so: `"Is", "this", "the", "way", "to", "Amarillo", "?"`.

n-grams An n-gram is a subsection of a tokenised sentence, where `n` represents the number of tokens in a subsection. An n-gram of length 3 (also known as a tri-gram) of the above sentence could be `"way", "to", "Amarillo"` The location of these n-grams, their frequency and their composure all provide data points that can provide insights into the meaning of a body of text (Banerjee & Pedersen, 2003).

Colocations Manning and Schutze describe a colocation as "an expression consisting of two or more words that correspond to some conventional way of saying things". For example, "around about", "stark naked" and "stiff upper lip" are all colocations. In a colocation, the subsequent parts make up a whole and lose some of their independent meaning - "fool hearted" makes sense to an English speaker's ear, but "idiot hearted" could sound offensive, or cause a misunderstanding. One way of identifying colocations is to count the frequency of bigrams in a body of text - a high number of two words occurring next to each other could indicate a colocation.

Hidden Markov Models A Hidden Markov Model (HMM) is a system with hidden states. The system is comprised of a Markov chain, a model describing the probabilities an event occours. In NLP, HMMs can be used to identify how often a token or n-gram

2.5.2 Structured NLP

2.5.3 Machine Learning

3 Methodology

3.1 Data Collection

In order to create an algorithm to detect incongrouence in news articles, first news articles have to be collected. While there are already existing datasets, they tend to be incomplete, out of date, or in innacessiable formats, as they've been made to tackle different problems.

To overcome these obstacles, a bespoke dataset needs to be created that can be classified and analysised by an NLP algorithm.

3.1.1 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.

3.1.2 Sources

The Independent is one of the only online publishers to make available their entire archives. Using the methods mentioned in Section 3.1.3, 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

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

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

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**	**Raw**	**Cleaned**
BBC On This Day	1950-01-21	2005-12-11	1857	1857
The Independent	2011-01-01	XXX	XXX	XXX
BBC	2019-04-18	XXX	XXX	XXX
Daily Mail	2019-09-24	XXX	XXX	XXX
The Guardian	2020-06-25	XXX	XXX	XXX
Huffington Post	XXX	XXX	XXX	XXX

Table 1: Extents of the data sources collected

3.1.3 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.

3.1.4 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

[5] `https://github.com/jacobbarrow/honours/tree/master/data-collection`

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 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.

3.1.5 Cleaning

While a bespoke scraper was created for each site, on some articles publishers used different page structures or included certain elements (such as inforgraphics) that the scraper didn't know how to handle. As a result, a portion of the articles in the dataset have erroneous text in them, such as unformatted lists of tweets or social media comments.

When creating the subset of articles for labelling, out of the 300 records 36 (12%) were corrupt or included content not part of the article's body text. Extrapolating this to the rest of the dataset, this means approximately XXX of the collected articles are 'dirty'.

3.2 Data Labelling

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

Need for labelled data, website design and implementation, crowdsourcing labelling, ethics, analysis of labelled data

[6] `https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research`

4 Analysis and Discussion

5 Conclusion

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

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. In *International conference on intelligent text processing and computational linguistics* (pp. 370–381).

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).

Digirolamo, G. J., & Hintzman, D. L. (1997). First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin & Review*, *4*(1), 121–124.

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.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 acm sigmetrics international conference on measurement and modeling of computer science* (pp. 179–192).

Jackson, P. (2002). *Natural language processing for online applications : text retrieval, extraction, and categorization.* Amsterdam Philadelphia, PA: John Benjamins Pub.

Karlsson, M., & Strömbäck, J. (2010). Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. *Journalism Studies*, *11*(1), 2–19.

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 so-*

lution (csitss) (p. 1-5).

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

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

Park, K., Kim, T., Yoon, S., Cha, M., & Jung, K. (2020). Baitwatcher: A lightweight web interface for the detection of incongruent news headlines. *arXiv preprint arXiv:2003.11459*.

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

Reis, J., Benevenuto, F., de Melo, P. O., Prates, R., Kwak, H., & An, J. (2015). Breaking the news: First impressions matter on online news. *arXiv preprint arXiv:1503.07921*.

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.