Embedded hyperlink structures as a feature for fake news classification

Jamie Coffey

April 19, 2022

A dissertation submitted in partial fulfilment of the requirements for the degree of MCS (Computer Science)
Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: ___________________________ Date: ___________________________
Abstract

Since 2016 fake news classification has risen from an obscure concept to a prominent and dynamic field of research. Much attention has been given to the very natural idea of classifying articles based on their textual contents, but as we push against the limits of this technique other features to continue improving machine learning models are being investigated.

One technique that has been looked at in particular has been the use of network graph analysis to examine the spread of fake news across social media platforms. Twitter in particular has been a popular target of this research. However online news articles already have a built in propagation technique in the form of hyperlinks embedded within the article. This feature could potentially hold the same kind of classification capabilities as social media networks, however until now this has went uninvestigated.

This research project attempts to fill this hole in the research by proposing a technique to develop a data set for this purpose. It proposes a web scraping process that can be applied to almost any website with minimal adjustments. It also describes an algorithm for converting these websites into network graphs. It then applies the popular PageRank algorithm to these graphs and uses this feature to compare a purely linguistic model to a novel linguistic-network hybrid model. This research contributes its novel data set and model to the literature in the hope of making future research in this area easier going forwards.
Acknowledgements

I would like to thank Dr. Owen Conlan for his constant support and guidance on this project over the year as my supervisor. I would also like to thank Dr. Brendan Spillane for his always helpful advice on scientific writing and process.

I would also like to thank my friend Jack Engels, a fellow computer scientist with whom I had conversations that undoubtedly shaped the direction this research eventually took.

Finally I want to thank my parents Nigel and Sharon Coffey for their ceaseless caring and support over the past five years of college.
Contents

1 Introduction .......................................................... 1
  1.1 Motivation ......................................................... 1
  1.2 Objectives ......................................................... 2
  1.3 Contributions ..................................................... 3
  1.4 Challenges ........................................................ 3
  1.5 Dissertation Overview .......................................... 4

2 Literature Review .................................................. 5
  2.1 Technical Background ........................................... 5
    2.1.1 Machine Learning ........................................... 5
    2.1.2 Network Graphs .............................................. 8
  2.2 Fake News ........................................................ 8
    2.2.1 What is Fake (or False) News? .......................... 8
    2.2.2 The Challenge of Fake News ............................... 10
  2.3 Fake News Classification ....................................... 11
    2.3.1 Content Based Approaches to Fake News Classification .... 11
    2.3.2 Network Based Approaches to Fake News Classification .... 12
  2.4 Conclusion ......................................................... 14

3 Methodology .......................................................... 15
  3.1 Experiments ........................................................ 15
  3.2 Obtaining a Source data set .................................... 16
    3.2.1 FakeNewsNet Data set Analysis .......................... 17
  3.3 Experiment 1: Synthesising a Novel Network-Based data set .... 18
    3.3.1 Web Parsing .................................................. 19
    3.3.2 Graph Representation and Generation ...................... 21
    3.3.3 Synthesising the Final Data set ........................... 24
  3.4 Experiments 2 and 3: Training Models .......................... 26
    3.4.1 Content-Based Model Development ......................... 26
    3.4.2 Network-Based Model Development ......................... 26
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>2-class confusion matrix with formulas for calculating evaluation metrics</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Dr J. L. Moreno’s sociogram depicting social relationships between 8th grade students at the New York Training School for Girls.</td>
<td>9</td>
</tr>
<tr>
<td>3.2</td>
<td>Screenshot of html corresponding to Figure 3.1</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>Directed network graph generated from hyperlink structure of <a href="https://people.com/tv/scandals-bellamy-young-talks-being-adopted-landing-her-big-break/">https://people.com/tv/scandals-bellamy-young-talks-being-adopted-landing-her-big-break/</a></td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Graph generation pipeline</td>
<td>25</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Breakdown of data set article quantities based on truthfulness and source of data ........................................ 17
3.2 Number of articles composing each of the top ten sources in the FakeNewsNet data set .................................. 18
3.3 Example DataFrame showing provided contents with room for expanding with additional data ......................... 23
3.4 Example DataFrame showing input DataFrame used for graph generation created from FakeNewsNet data set .......... 23
4.1 Cross-validation evaluation results for purely content-based model ............................................................... 32
4.2 Cross-validation confusion matrix results for purely content-based model ...................................................... 32
4.3 Cross-validation evaluation results for network-content hybrid model ........................................................... 33
4.4 Cross-validation confusion matrix results for network-content hybrid model .................................................. 33
4.5 Comparison of evaluation results for content-based and network-content hybrid models ................................. 34
1 Introduction

This chapter will give a basic overview of the contents of the rest of the dissertation. It will begin by discussing the motivation for pursuing this avenue of research before detailing the exact goals that the rest of this dissertation will fulfil. This will be followed by a brief summary of each of the remaining chapters of the paper and a discussion of the challenges faced throughout the research process.

1.1 Motivation

In recent years fake news has become an increasingly prominent issue in our society. With the rise of social media it has become easier than ever for fake news stories to become widespread. Studies have shown that fake news also spreads much faster than real news, with results finding that “Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information.” (3)

Discussion of fake news rose significantly in the aftermath of both the UK’s referendum on Brexit and the 2016 US presidential election, both of which were frequently perceived to have been influenced by fake news on social media (4). Journalistic organisations such as Snopes and PolitiFact have been attempting to combat this through manual fact-checking, however there is far too much fake news being created for them to correct even a fraction of the falsehoods (5). This has made fake news classification a popular field of research in recent years, however due to both the scale of the task and how recent this surge of interest in the field is there are still many areas that remain largely unexplored.

Much work has been done investigating possible ways of classifying fake news using machine learning. There has been a lot of work done in using NLP techniques to identify fake news stories solely from the text contained within them (6). Researchers have tried training all different kinds of models, from simple logistic classifiers, to support vector machines, to complex neural networks and more. They have tried many different methods of feature engineering with the textual contents, but fake news classification remains an unsolved problem with room for improvement. Others have created network graphs showing the spread of fake news articles through a social media platform. They then use spatial/
spatio-temporal analysis to identify news stories that are likely to be false based on their spread through the network (7).

These network analysis techniques, when combined with a traditional linguistic approach have shown a clear improvement over purely linguistic approaches when detecting fake news in social media posts. However, this same idea has not yet been applied to news articles, which already contain a network structure within them in the form of hyperlinks. It seems worth investigating whether similar improvements could be made to fake news article classifiers by applying this hybrid technique to them.

1.2 Objectives

The research question that will guide the rest of this dissertation is the following:

*Can the hyperlink structure embedded in news articles be used to improve existing fake news classification techniques?*

However, this investigation will require solving several problems along the way, which are as follows:

1. Research current linguistic fake news classification techniques
2. Understand the current work using network graphs to improve fake news classification
3. Either find or create a large, labelled data set of articles for analysis
4. Create a consistent way of representing hyperlink structures in a network graph
5. Implement a linguistic machine learning model
6. Modify model to take advantage of network-based features
7. Evaluate both versions of the model to determine if the network-based features offered any advantages over the traditional model

This paper proposes a web scraping algorithm and graph generation system to dynamically create network graphs showing the cited internal articles that are connected to the main article via hyperlinks. It will then propose a novel fake news classification model, augmenting traditional linguistic approaches by taking spatial data into consideration too. Finally, it will evaluate the performance of this approach against a purely linguistic model to see if network graph modelling can increase the efficacy of our classifications.
1.3 Contributions

Major Contributions

This work provides future researchers with access to a novel data set allowing for the study of an entirely new kind of fake news classification features. Legal restrictions prevent the distribution of pre-generated data scraped from websites but the tools to exactly regenerate the data used here are provided on github (https://github.com/JamieJNPC/Hyperlink-FakeNewsNet). It then trains a novel classifier using an established network graph analysis feature and evaluates its performance relative to a similar linguistic classifier.

Minor Contributions

This paper contributes a semi-algorithmic web scraping technique to allow for the automatic parsing of content data from online news articles. It also proposes a network generation algorithm for interconnected pieces of online media with applications beyond news articles.

1.4 Challenges

Due to the novelty of this work there were several major problems encountered over the course of this project. To begin with, finding a large, high-quality data set was difficult due to the fact that journalistic experience would be required to create hand labelled data and it is such a laborious process to do so that few appropriate data sets were available for use.

Another problem was the fact that different websites do not follow a standard HTML format exactly, meaning that the web scraping had to be implemented differently for several of the different websites that were used.

The size and scope of these problems fundamentally shaped the direction that this dissertation eventually took, since the lack of useful data in the field was identified as a major blocker of research in this area. This led the project to focus on ensuring the development of a competent and versatile data set that would be usable beyond merely the scope of this dissertation, since this would allow future researchers to build off of this groundwork more easily.
1.5 Dissertation Overview

Literature Review

This chapter will begin by explaining the fundamentals of machine learning and network graphs that provides important context for the rest of the dissertation. It will then detail some important background knowledge about the topic of fake news, before discussing the work done with traditional linguistic models in this field. After this it will then review literature about the implementation of network graph analysis techniques with social media networks and discuss their findings.

Methodology

This chapter will begin by discussing the specific requirements of the data set needed for these models. It will then explain the source and details of the obtained data set, as well as how the obtained data is processed to make it usable for both models. It will then propose two fake news classification models. The first model will be a traditional purely linguistic based classifier whilst the second one will be a novel approach using a network of hyperlinks. Finally it will discuss the training process for each of these models.

Evaluation

This chapter will contain two major elements evaluating the success of both our major and minor contributions. It will begin by critically examining the output of the data set synthesis process with a number of novel calculations and observations. It will then show the evaluation results of each of our models when applied to a testing set of the data. Finally it will compare these results to evaluate if this simple but novel technique offers any advantages over the more traditional model.

Conclusion

This chapter will explain the main takeaways from this project, the potential limitations of its findings, and propose some ideas for where future research could go with these findings.
2 Literature Review

In this chapter the background research conducted for this dissertation will be highlighted and discussed. It is intended to provide a strong background context for the rest of the research contained within this work. Both the public and academic interest in fake news is relatively new, meaning that large gaps in the literature still remain. However generally speaking there are four major topics of research that are of interest to us. The first is the technical background information about machine learning and network analysis that is needed to understand the rest of the dissertation. Second is to do with the nature of fake news; what it is, and why it’s such a problem. Third, we need to understand the work done in textual, or content-based fake news classification. Finally, we can then take a look at the even newer field of network-based fake news classification.

2.1 Technical Background

This section will provide an introduction into some of the technical and algorithmic fundamentals that will be important for later discussion regarding the processing of fake news. It will highlight foundational concepts in the fields of machine learning and network graph analysis.

2.1.1 Machine Learning

Machine learning as a field of computer science first took off in the scientific conscience as far back as the 1960s (8). There was a large push to try and model the human neural learning process in computers, since this would then allow them to learn from any set of data in a general way and could allow them to solve inconceivably complex problems without ever being specifically programmed to handle it. However eventually these lofty goals pushed up against mathematical and computational power limits and what followed was a long period of neglect towards the field. However in the 21st century with more powerful computers and the emergence of GPUs, the field saw a new boom in funding and interest. This time around there was a better understanding of the abilities and limitations of the technology, and the idea was championed not only by computer scientists, but even among
the general public as the future solution to many of the most complicated computational problems of modern times. Today, most people are aware that machine learning powers everything from content-recommendation systems on platforms such as Netflix and YouTube to the image processing pipeline behind modern smartphone cameras.

Machine learning problems can be split into one of two categories depending on the kind of data being worked with. If the problem requires categorising something into one of two or more classes then it is considered to be a classification problem. Machine learning algorithms of this kind are called classifier models. There are also problems that involve calculating a value such as a price or a size where the answer can be any value on a continuous scale. These are known as regression problems. However since fake news detection (as considered by this dissertation) is explicitly a classification problem, the rest of this dissertation will therefore focus on this area.

There are many different types of classifier models, from simple classifiers such as logistic regression and SVMs (Support Vector Machines) to more complex neural networks. However the process of training each these models typically all involves a similar process.

Feature Engineering

Firstly feature engineering usually has be performed. This involves taking the raw data and processing it into a format that is appropriate for the model to interpret. Depending on the kind of data being worked with, this step can involve many different possible actions being taken. Since fake news classification almost always involves text processing it is worth understanding the standard process by which this is done. Typically when a string of text, whether it’s a sentence, paragraph or more, has to be encoded for a machine learning model to understand it, the first technique usually employed is to represent it as a Document-Term Matrix (9). This involves creating a bag-of-words representation of the text by forming a large matrix where each column represents a single word that appears in one of the documents being encoded, and said documents being represented in rows. Then the value within each cell represents a words usage within the document. There are different methods of calculating this value. The easiest method is to simply count the number of times the word appears in the document. However the most popular technique is known as TF-IDF (Term Frequency-Inverse Document Frequency). The details of how it works are not important here but it is worth highlighting that this has become the standard go to approach when investigating text as a feature in machine learning.

Cross-Validation

Most machine learning models feature at least one hyper-parameter. A hyper-parameter is a tuning value that affects the performance of the model. An example of this is the k of k-NN
(k-Nearest Neighbours). A k-Nearest Neighbours model classifies the input value as whichever class several of the most similar training inputs belong to, the number of nearest values being called k. In this case changing the value of k can make the model fit more or less towards small clusters of values.

Another example that is important to explain explicitly (but briefly) for this dissertation is logistic regression, which has two hyperparameter values that can be tuned. The first is the penalty type which can be either L1 or L2. In L1 regularisation, the model is penalised based on the amount of non-zero weighted features it uses, encouraging the model to ignore features that aren’t important to the final prediction. L2 regularization by contrast penalises the model based on the weighting of each feature, encouraging the model to avoid relying on a couple features primarily and to spread out the weighting more evenly. The strength of the penalisation is determined by the second hyperparameter, C. C is also known as inverse regularisation strength, and the larger the value of C, the lower the effects of the penalty type are on the model. Decreasing C serves to make the regularisation effect more pronounced on the model.

### Evaluation Metrics

In order to perform cross-validation or simply to evaluate the effectiveness of any machine learning model it is necessary to select a metric or metrics for evaluation. For classification problems it is possible to create what is known as a confusion matrix. There are also a series of possible metrics that can be evaluated from a confusion matrix, such as accuracy, precision, recall and f1-score. Figure 2.1 showcases the exact layout of a 2-class confusion matrix and how to calculate each of these metrics.

![Confusion Matrix](image)

- **Correctly classified:**
  - True Positives (TP)
  - True Negatives (TN)

- **Incorrectly classified:**
  - False Positives (FP)
  - False Negatives (FN)

**Column totals:**
- Positive (P)
- Negative (N)

<table>
<thead>
<tr>
<th>Hypothesized class (Y)</th>
<th>True class (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (p)</td>
<td>True Positives (TP)</td>
</tr>
<tr>
<td>Negative (n)</td>
<td>False Negatives (FN)</td>
</tr>
</tbody>
</table>

- **Fp rate:** $\frac{FP}{N}$
- **Tp rate:** $\frac{TP}{P}$
- **Precision:** $\frac{TP}{TP+FP}$
- **Recall:** $\frac{TP}{TP+FN}$
- **Accuracy:** $\frac{TP+TN}{P+N}$
- **F-measure:** $\frac{2}{1/\text{precision}+1/\text{recall}}$

Figure 2.1: 2-class confusion matrix with formulas for calculating evaluation metrics (1)
2.1.2 Network Graphs

Network graph analysis is a fundamental technique for representing and understanding data that has been used in countless ways to provide analytical solutions to all sorts of problems. It has been particularly effective in the social sciences, where its use dates back as early as the 1930s. In 1934, Dr J. L. Moreno published what has since become considered the foundational text in social network analysis (10). In it, he explains a surge in runaway students from the New York Training School for Girls by creating network graphs showing the social relationships between the students. He demonstrated that this model could effectively explain why and when certain groups of girls ran away. These specific kinds of network graphs are known as sociograms, and an example of one of Moreno’s sociograms is depicted in Figure 2.2, with remarkable similarity to the directed network graphs that are now commonly used to solve all kinds of problems.

Since then many network graphs have been created to help analyse all kinds of complex problems, and many different algorithms have been developed to understand these graphs. One of the most famous algorithms that has been developed in the past few decades is Google’s PageRank algorithm. PageRank is a backlink assessment algorithm that attempts to quantify the significance of a node within a network (11). It was originally developed to help Google’s then new search engine in identifying useful links on the internet that people would actually want to find in a way that would be harder to exploit than simpler approaches. This algorithm has since become popular in network analysis due to it’s ability to identify important nodes in all kinds of networks.

2.2 Fake News

This section will take the previously discussed concepts and apply them to the specific domain of fake news classification. It will begin with a general introduction to the topic of fake news before specifically discussing the current work being done in fake news classification linguistically. It will conclude by highlighting the promising role that network graph analysis has in offering potential additional features for fake news classifiers.

2.2.1 What is Fake (or False) News?

When fake news can convince people of its truthfulness it can result in major consequences for the world at large. For example, when The Associated Press’s twitter account was hacked in 2013 and somebody tweeted from it that then US president Barack Obama had been injured in an explosion, $130 million were wiped out of the stock market before people realised that they were trading on faulty information (12).

As an organisation that has found itself at the centre of discussions around the
dissemination of fake news, Facebook published an in-depth paper discussing the spread of fake news on its platform (13). It provides us with some helpful definitions and categorisations that can help to untangle the different ways in which bad actors can intentionally spread misinformation. The first of these is an “Information operation”, which it defines as “Actions taken by governments or organized non-state actors to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome. These operations can use a combination of methods, such as false news, disinformation, or networks of fake accounts (false amplifiers) aimed at manipulating public opinion.” This then introduces the terms “disinformation”, “false amplifiers” and “false news”. “Disinformation” is defined as “inaccurate or manipulated information/content that is spread intentionally. This can include false news, or it can involve more subtle methods, such as
false flag operations, feeding inaccurate quotes or stories to innocent intermediaries, or knowingly amplifying biased or misleading information. Disinformation is distinct from misinformation, which is the inadvertent or unintentional spread of inaccurate information without malicious intent."

Finally, "False news" is simply the papers preferred term instead of “fake news”. It is not alone in this preference (3) and papers that use this term often justify it with the same reason. Their preference for false news is due to the fact that the language and discourse around fake news has become increasingly politicised and the term is frequently used incorrectly. Sometimes this is done innocently, however it is often misused intentionally to hurt political opponents or to dismiss valid criticisms. However although arguably a matter of personal taste, this paper doesn’t see the non-expert misuse of the term as a strong enough justification to change the scientific language around it. As such this paper will continue to use the more common term “fake news”, but highlights this issue to ensure that the terminology is understood in the correct context.

2.2.2 The Challenge of Fake News

Due to the primarily text-based nature of news articles, language is obviously the primary factor to consider when classifying it into most sets of categories. However fake news is usually created with the intention of misleading the reader into believing it is genuine. For this reason, the language of fake news articles often mirrors the language of real articles. (4) investigates the structural differences between real and fake news articles by evaluating both according to traditional markers of news worthiness. It found that “the majority of the [fake news] articles we studied included the news values of timeliness, negativity, and prominence; were about government and politics; and were written in an inverted pyramid format.” All of these are attributes that traditionally identify genuine news articles too. However, the distinguishing aspect of fake news is the presence of opinion in the text. The paper highlights that whilst only 24.2% of articles in The New York Times expressed the authors opinion 64.3% of the fake news articles examined contained opinion. The paper doesn’t make this connection but these content based similarities highlight the need for non-content based detection methods.

Whilst high quality journalism is written to inform rather than to entertain, fake news is more often written with the intent of being spread quickly regardless of quality. This results in fake news spreading differently than typical news articles. As mentioned earlier, (3) found that “Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information” in their paper studying the spread of true and false posts on twitter. The paper makes useful structural observations about the difference in spreading patterns between true and false posts by examining the degree and depth of nodes in a
network. The paper also identified that false tweets typically contained more novelty (quantified as the information distance between the tweet and all other tweets the user had recently been exposed to) than truthful tweets. On top of this it found that the replies to false tweets exhibited greater disgust, whilst truthful tweets inspired more sad responses. Finally it used a bot detection algorithm to remove suspected bots from the dataset, and interestingly found that none of the above conclusions were changed by their absences, suggesting that contrary to popular belief the rise of fake news has not been brought about by the presence of social media bots, but rather human nature.

(14)) looks into the rumour diffusion surrounding a particular rumour which circulated during the Great East Japan Earthquake in 2011. Shortly after the incident had occurred, a false rumour began spreading through social media that dangerous chemicals would soon be raining from the sky, and that people should use coats and umbrellas to protect themselves from it. Whilst the rumour picks up steam quickly and the corrections begin pouring in almost immediately, it took over a day (and until the government stepped in with an official announcement) for the corrections to reach a similar audience. Importantly this again indicates a noticeable, structural difference between the spread of true and false narratives.

This section has highlighted several societal problems around fake news. However most importantly it highlights that several significant pieces of research have indicated a structural difference between the spread of real and fake news. Although the literature has been slow to pick up on the idea, it indicates a potential method to identify fake news articles directly from structural network patterns.

2.3 Fake News Classification

This section will build on the previous discussion of foundational machine learning and network analysis techniques, alongside the highlighted problem of fake news by applying these ideas to the area of fake news analysis. It will explain several approaches that have been investigated by other works within this area. Some of these will be purely linguistic, or content based in nature, whilst others will feature some aspect of network analysis on top of the traditional content based approach. It will conclude by highlighting the exact gap in this field that the rest of this work will focus on filling.

2.3.1 Content Based Approaches to Fake News Classification

It may at first seem trivially easy to apply standard machine learning text classification techniques such as (9) to fake news. However fake news has proven to be a significantly more complex and nuanced problem than many standard classification problems. One such
issue that has been highlighted is the problem of classifiers overwhelmingly trained on US centric news performing significantly worse on non-US media. (15) investigated this issue by training an LSTM (Long-Short Term Memory) model on exclusively US news, and another model with a South African corpus of articles. It found that this discrepancy caused the US-based model to perform significantly worse despite it’s significantly larger corpus size than its SA trained counterpart. Although this is just one instance, it highlights the importance of carefully selecting the data that is used in this field, since it is easy to become blinded by personal biases, convenience or both.

In 2017, the first Fake News Challenge (FNC-1) was held with the aim of developing more effective fake news stance detection methods (16). It consisted of a dataset of news article headlines and contents, and the goal was to create a multi-class classifier to determine whether the contents of the article either agrees, disagrees, discusses or is unrelated to the headline. This multi-class model provides an interesting contrast to the typical purely agrees or disagrees binary the task is usually presented in. The scoring system allocated some marks for correctly identifying whether or not the variables are related, with the rest of the marks going for a correct classification within a related instance. The top performing model, SOLAT in the SWEN was a combination of gradient boosted decision trees and a deep convolutional neural network.

A retrospective analysis was later completed to critically examine both the event and the results of it (6). It criticised the scoring methods of the event, since it’s weighted scoring system could be easily exploited, and proposed using the traditional and multi-class aware F1-score to evaluate the systems instead. Using this new metric another model actually comes out on top of SOLAT in the SWEN, with the second place Athene model taking the lead. Athene used five MLPs with six hidden layers in each. However, it is worth noting that this model was created by the same people that authored this retrospective. It seems reasonable to question the impartiality of this retrospective despite its valid arguments since it’s newly proposed approach clearly favours their model compared to the old metric. One point that is worth highlighting is that this retrospective proposes a dummy classifier under the old evaluation metric, whereby the model always predicts “discusses”. This dummy classifier achieves a higher FNC score than any of the actual models submitted to the challenge. This does showcase the problems with the scoring metric but also serves to highlight the limitations of the linguistic based approaches used and the need to develop better models capable of beating this baseline.

2.3.2 Network Based Approaches to Fake News Classification

(17) Explores two different network relationships in the context of fake news classification. It identifies the relationship between the source and the piece itself, and the relationship
between the piece and the user. It then proposes and evaluates a tri-embedding framework known as TriFN. This model importantly captures the standard textual contents of the news articles alongside the connected social links and engagements, and the publisher of the piece of content. It found that even using just the non-linguistic features alone was more effective than the linguistic feature was on its own. On top of this, when considering both textual and spatial features, the model achieved even better performance than either of the elements individually could allow.

(7) Builds on the work of (3) by building a network based model off of that papers observations about the structural differences between the spreading of real and fake news. Specifically remember that it's observation was that “Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information.” These could be elaborated on and developed into the following rules to tell real and fake news articles apart:

1. (More-Spreader Pattern) More users spread fake news than true news.
2. (Farther-Distance Pattern) Fake news spreads farther than true news.
3. (Stronger-Engagement Pattern) Spreaders engage more strongly with fake news than with true news.
4. (Denser-Network Pattern) Fake news spreaders form denser networks compared to truth spreaders.

(18) Takes the idea of using the structural fact that fake news spreads more effectively than real news through a social network and applies Google’s PageRank algorithm to identify news articles that have been more successful in spreading news to users.

Hyperlink analysis within the context of fake news classification is a subject that has hardly been investigated at all so far. One of the few available papers beginning to look into the idea gives the entire topic a mere paragraph near the end of the paper (19). It studies the hyperlink citation patterns of nine new sources ranging from right wing to far-right in their coverage of the 2016 US presidential election. In order of extremity from most to least mainstream it investigates Fox News, the Washington Examiner, The Daily Caller, The Gateway Pundit, Breitbart, Infowars, Vdare, American Renaissance and the Daily Stormer. It found that while the far-right news sites frequently cited the more mainstream sources (and Fox News in particular was referenced often by all other sources) the more mainstream news sources did not cite the far-right articles nearly as often. This points to a one-sided structural difference that network graph analysis will hopefully be able to identify, since intense partisanship is often correlated with falseness of news.
2.4 Conclusion

This literature review has highlighted the significance of the problem of fake news. It is capable of appealing to humans in ways truthful news cannot. It has also highlighted the issues and limitations of our current linguistics first approach to fake news classification, showing that contrary to expectations a social network-based approach isn’t only helpful, but is more helpful on its own than content-based detection. Furthermore combining these two approaches offers the most effective classification techniques of all. It describes the structural reasons behind the efficacy of network-based approaches and highlights a couple of the recent network-based implementations of note. The current literature lacks any real study of hyperlink analysis beyond the single cursory article described above. This is despite the article being over two years old and presenting a promising possibility for an additional feature for fake news classification. In the following sections this dissertation will begin the work of filling this substantial gap in the literature.
3 Methodology

This chapter will discuss the methodology that was followed in the process of this research. It will lay out the series of experiments carried out in the course of this project and detail the implementation of them. Much of this section will focus on the development of the eventual data set, before detailing the work of training and evaluating both a simple linguistic classifier and more sophisticated novel network-based classifier.

3.1 Experiments

The rest of this project can be split into three major experiments which are as follows:

1. Design a suitable method for understanding and representing the hyperlink structure of news articles from a labelled data set
2. Develop a purely linguistic fake news classifier
3. Develop a fake news classifier that uses the network graph data from experiment 1 and the linguistic technique from experiment 2 to determine if the network features offer any improvement

It is worth expanding on each of these experiments to understand what a successful outcome to each will look like. For experiment 1 the ability to perform experiment 3 will be considered a minimum requirement for success. Although a truly successful experiment 1 will also provide a network graph data set with the textual contents of each article in the graph provided.

Experiment 2 will be the most straightforward of the three since this is well trodden territory, however it is a crucial step to get right in order to ensure experiment 3’s feasibility.

Finally experiment 3 will be considered a success if it can offer a meaningful improvement over experiment 2’s results. However if this improvement cannot be found then this experiment will have been inconclusive. A failure of this experiment would involve failing to develop the model at all due to one of many factors.
3.2 Obtaining a Source data set

Initially it was investigated to see if an appropriate data set already existed. There were a few requirements for the data set that would have to be met:

1. The data set must contain hand labelled data classifying each article as either true or false
2. The data set must provide in some form the textual contents of each article
3. The data set must provide a network graph containing the connected articles for each source article

Unfortunately, whilst data sets of this kind for social media posts are abundant these kinds of data sets do not appear to have yet been synthesised for hyperlinks. This meant that the decision had to be made to synthesise the necessary data set for this project. However, the creation of this novel data set would still require a source data set to build from. The requirements for this source data set were as follows:

1. The data set must contain hand labelled data classifying each article as either true or false
2. The data set must provide either a url for each source article or the html from the webpage to allow the connected articles to be discovered and parsed too

Although there exists techniques for automatically extrapolating more data from a small hand-labelled fake news data set, such as (20), it is still relatively crucial for our data set to be of an appropriate size. This is because the need to manually tune the parser for each individual source limits the number of sources our data set can use to a low enough number that the data set requires nuance within each source (ie. true and false articles within each source). The technique that is suggested in (20) involves treating every article from a trustworthy or untrustworthy source as true or false respectively to create a noisy but large and somewhat effective data set.

As mentioned previously, FNC-1 offered a data set containing a series of news article headlines and their corresponding textual contents (16). However, this makes it unfit for this research paper’s purpose since this format fails to meet our second criteria for a viable data set.

CREDBANK (21) is a popular data set for fake news classification, and particularly for fake news network analysis. It contains a corpus of over 60 million tweets which crowd sourced the work of classifying them. However whilst this data set is great for social media analysis the nature of tweets makes them unsuited to this papers needs since they already use a non-hyperlink based spreading structure. The way in which tweets spread socially through


retweets and replies has already received appropriate attention (22)

Finally, FakeNewsNet (23, 24, 25) is a data set that was also developed to investigate the spread of fake news on twitter. It contains a data set mostly focused on this aspect, however instead of studying fake news contained directly within tweets, it is focused on actual news articles that have been linked within tweets. This manages to meet the second criteria as these hyperlinks are provided by the data set. On top of this each article is hand-labelled by journalists from either PolitiFact or GossipCop based on its truthfulness, meeting the first criteria as well.

3.2.1 FakeNewsNet Data set Analysis

This section will discuss the in-depth analysis that was performed to ensure that the FakeNewsNet data set is sufficient for our use case, alongside relevant information worth noting about it. To begin with, the data set was generated in its most up to date state as of January 2022. Due to potential legal issues with the distribution of copyrighted material the data set is not provided directly. Instead, the code for generating it is provided with run time instructions. The data set has also been updated significantly from the initial version discussed in the cited work.

After generating a local copy of the data set, the data was inspected in a number of ways. To begin with, the total data set was loaded and discovered to contain 22,866 articles. These could be broken down into 17,371 True articles and 5,495 False articles as seen in Table 3.1. The url’s for the articles were then parsed to determine the publishing source by extracting the section of the string between a potential "http://" or "https://" and the following "/" character. This revealed a total of 2,442 separate sources, with an average of 10 articles per source. However, as Table 3.2 demonstrates, these articles are not shared evenly between the sources. This helpfully shows that parsing even just a few of the largest sources in the data set can provide a sufficiently large data set for experimentation and classifier training. Because of this, and the inviability of meaningfully working with 2,442 unique websites, the rest of the research contained within this paper will focus exclusively on the data from these 10 websites.

Table 3.1: Breakdown of data set article quantities based on truthfulness and source of data

<table>
<thead>
<tr>
<th>Source</th>
<th>True</th>
<th>False</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolitiFact</td>
<td>567</td>
<td>428</td>
<td>995</td>
</tr>
<tr>
<td>GossipCop</td>
<td>16804</td>
<td>5067</td>
<td>21871</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17371</strong></td>
<td><strong>5495</strong></td>
<td><strong>22866</strong></td>
</tr>
</tbody>
</table>
Table 3.2: Number of articles composing each of the top ten sources in the FakeNewsNet data set

<table>
<thead>
<tr>
<th>Source</th>
<th>True Articles</th>
<th>False Articles</th>
<th>Total Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>people.com</td>
<td>1570</td>
<td>216</td>
<td>1786</td>
</tr>
<tr>
<td><a href="http://www.dailymail.co.uk">www.dailymail.co.uk</a></td>
<td>770</td>
<td>194</td>
<td>964</td>
</tr>
<tr>
<td>en.wikipedia.org</td>
<td>618</td>
<td>123</td>
<td>741</td>
</tr>
<tr>
<td><a href="http://www.usmagazine.com">www.usmagazine.com</a></td>
<td>562</td>
<td>147</td>
<td>709</td>
</tr>
<tr>
<td><a href="http://www.etonline.com">www.etonline.com</a></td>
<td>585</td>
<td>81</td>
<td>666</td>
</tr>
<tr>
<td><a href="http://www.longroom.com">www.longroom.com</a></td>
<td>549</td>
<td>0</td>
<td>549</td>
</tr>
<tr>
<td>hollywoodlife.com</td>
<td>64</td>
<td>460</td>
<td>524</td>
</tr>
<tr>
<td><a href="http://www.usatoday.com">www.usatoday.com</a></td>
<td>300</td>
<td>32</td>
<td>332</td>
</tr>
<tr>
<td><a href="http://www.hollywoodreporter.com">www.hollywoodreporter.com</a></td>
<td>298</td>
<td>32</td>
<td>330</td>
</tr>
<tr>
<td>variety.com</td>
<td>259</td>
<td>45</td>
<td>304</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5575</strong></td>
<td><strong>1330</strong></td>
<td><strong>6905</strong></td>
</tr>
</tbody>
</table>

It is tempting to begin making observations based on the numbers of true and false articles within this data set. An example of this would be calculating the ratio of truthful-to-false articles from each source. However this paper has consciously avoided providing any statistics of this kind. This data set has not been generated with any consistent methodology to determine which articles are or aren’t reviewed from any given source, meaning that it by no means provides articles at a ratio representative of their entire source website. One more observation of note is that due to its larger size, all of the top ten websites are from the GossipCop data set.

### 3.3 Experiment 1: Synthesising a Novel Network-Based data set

This section will describe the process by which this papers proposed data set was generated. It will explain the general procedure that was followed to create a web parsing algorithm for each of the chosen websites, as well as showing specific examples for each website. This process will be reproducible for most online news sources and other kinds of websites that meet certain criteria.

This section will also explain the exact algorithmic process that is used to generate the graph structures. It will detail how the web scraping parser is integrated with this algorithm to create an entire data set generation pipeline.
3.3.1 Web Parsing

Most online news publications do not have the author of an article format the HTML of the webpage themselves for obvious reasons. Instead they typically rely on internal publication tools that generate the webpage from the provided article contents. This allows for the dynamic insertion of digital ads, dynamically generated related links and more. It also significantly simplifies the publication process. A helpful benefit of this approach for this research is that all articles on a certain website typically conform to a common HTML layout. This means that by inspecting the HTML structure of a single article from the publishing source of interest it is possible to build a HTML parser that can scrape the contents from any standard article from that website.

Take as an example of this, the website people.com from the FakeNewsNet data set. One of the URLs contained in the data set is https://people.com/tv/scandals-bellamy-young-talks-being-adopted-landing-her-big-break/. As of the publication of this paper, opening this link should display something similar to what is displayed in Figure 3.1. Figure 3.2 then displays the corresponding HTML code for the webpage. The highlighted link in Figure 3.1 corresponds to the highlighted line of HTML code in Figure 3.2. By inspecting the structure of the code in Figure 3.2 it is possible to identify a standard structure for extracting relevant content and hyperlinks from the article, albeit noisily. People.com follows one of the most standard HTML structures this research has identified. It begins by enclosing the entire article and related contents within an HTML article tag. It then stores the textual contents of the article in a series of p tags, with hyperlinks further enclosed with a tags. Understanding this as a common and reproducible example allows this common approach for identification to be applied to the other websites in this paper's chosen data set. Although the exact structures of each website would be too space consuming to include here, this exact process was followed for each of the ten websites. With this structural understanding of each websites standard HTML layout it becomes possible to build a parser to automate the process of extracting the desired features from any given article on the target website. For this research's use case there are only two features of interest. It is necessary to identify both the textual contents of the article and all of the hyperlinks contained within it.

This parser was constructed in Python using the BeautifulSoup4 package for web scraping (26). It works by first of all processing the URL to ensure that it is encoded/decoded correctly and then requests the HTML from that web page. If the URL involves redirection due to URL shortening then the parser will keep retrying the URLs until it is no longer changing between requests. Once the HTML is received the parser then checks to see which website the article is sourced from, and applies the relevant web parsing algorithm based on this. In the people.com example, this means first extracting the contents of the article tag. This can
Figure 3.1: Screenshot of https://people.com/tv/scandals-bellamy-young-talks-being-adopted-landing-her-big-break/ taken on 10/04/2022.

Bellamy Young may be riveting as First Lady Mellie Grant on ABC’s Scandal, but her off-screen story is just as captivating.

“I was adopted, so I was in foster care for six weeks,” the actress, 45, says in the latest issue of PEOPLE. “We only had two lines on my dad and a

be converted straight into the output text using BeautifulSoup4’s built in get_text() function. Following this the list of p tags are identified, and then individually each of them are parsed for their nested a tags.

This process was followed for all ten websites in Table 3.2. Its worth noting however that www.longroom.com is no longer a working domain and so it has been excluded from the remainder of this work. Also hollwoodlife.com has either deleted or changed the domains of most of its articles from the dataset, meaning only 139 of the 524 articles were parsable. Finally www.usatoday.com was parsable, however suffered from a problem where the article urls in FakeNewsNet simply redirect the request towards the actual url, which involves functionality

The result of all of this is a parser which returns a string containing the contents of the
article, and a list of the embedded hyperlinks, which will be used in the graph generation algorithm detailed in the following section.

3.3.2 Graph Representation and Generation

This section will detail the work surrounding the creation of a graph generation algorithm. It will explain and defend the decision to use the python networkX package to represent the structure of the final graph. This will then lead into discussion of the exact process by which the graphs are created.

Graph Representation

The first decision that needed to be made surrounds the method of representing the graph structure. The Python programming language is perfectly capable of accommodating the manual development of a graph structure thanks to its support for the object oriented programming paradigm. It is also easily possible to conceive of a graph-like representation entirely through the use of standard Python data structures such as tuples and lists. However these would be cumbersome to apply more advanced analysis techniques to as everything would need to be implemented specifically for this use.

This is where the python package networkX comes in. It is a popular and established package for network analysis in the scientific community (27). This is due to several of the
packages features. It is very flexible, since it allows for both weighted and unweighted graphs, for nodes to be any kind of object and for edges to contain any kind of additional data. Also a feature that makes this package extremely useful for this work in particular is its built-in implementation of many standard network analysis algorithms, including Google’s PageRank algorithm discussed earlier.

It is worth returning to the data that the web parser is collecting, since this is the data that the graphical representation must ultimately store. The web parser collects two important pieces of data, the textual contents of an article and the list of embedded hyperlinks. It intuitively makes sense to store both of these features directly within the graph itself. This is easily accommodated by the networkX package, which allows any additional data about a node to be stored in custom attributes within the node object. However using this functionality here can result in increased complexity and inefficiency further in the data processing pipeline. Almost any analysis of the contents of these articles is going to involve applying some common process/ pipeline/ algorithm to every article in the graph. Also linguistic model training would require extracting the textual contents and combining them into a single list or DataFrame. Instead this work proposes splitting the textual feature into a separate DataFrame at the point of graph generation. Doing this means that the graph can represent the spatial data connecting the articles whilst the DataFrame can be used to store not just the textual data used here, but all kinds of metadata about the article. Non-spatial additional features are beyond the scope of this work, but this method of splitting the data will allow for easier processing and modelling of these features in future works.

Another decision that had to be made was whether the network graph should be directed or undirected. It was decided to make the graph directed since it can be easily converted to an undirected graph if needed, whilst the opposite transformation is not possible. Also although the investigation of directed graph analysis techniques is beyond the scope of this paper, including this feature in the data set will allow for this kind of research to be completed in the future using this data set.

To summarise the end result of these decisions for the resulting data set, two objects will need to be returned to represent the data in its entirety. A DataFrame with a layout as shown in Table 3.3, and a directed networkX graph where the nodes contain the url of the corresponding article and the nodes represent it’s hyperlinked connections to other articles.

**Graph Generation**

Now with the exact method for representation decided upon, it is possible to construct a specific algorithm for the expansion and parsing of the hyperlinks collected from the web parser to generate this desired output. This part of the work begins by following the
instructions provided by the FakeNewsNet data set (23) to generate the data set locally. This returns four csv files, two GossipCop and two Politifact csv files. Each source has one file containing truthful articles and another file containing the fake news articles. These four csv files contain four fields, an id, url, article headline and related tweet id’s. The only data this work requires from the data set is the url, and to record whether an article came from the true or false csv files. This leads to storing the resulting data in a combined DataFrame, with a url column and a label column as shown in Table 3.4. Each of these articles are then individually passed into the graph generation algorithm. The graph generation algorithm begins by passing the article to the web parser which downloads the HTML for the article and extracts the textual and spatial data from it. This is returned to the graph generator which then adds the article contents to the DataFrame. Following this the graph generator then goes through each of the newly discovered hyperlinked articles. It begins by checking the hyperlink for validity. This involves verifying that it is an internal link and that it is not pointing towards another website. Then if checks to see if the article has already been added to the graph yet. If it has not been, then a new node is created with the url and it is inserted into the graph. Regardless of whether or not the node is newly discovered, the edge to this url from article it was scraped from is also added to the graph. The end result typically resembles something like the directed network graph in Figure 3.3, which shows the hyperlink network generated from the previously analysed people.com article. The benefit of this algorithmic approach is that although this dissertation will only apply it
Figure 3.3: Directed network graph generated from hyperlink structure of https://people.com/tv/scandals-bellamy-young-talks-being-adopted-landing-her-big-break/.

To a small subset of the FakeNewsNet data set, future works will easily be able to apply it to more or different articles as needed.

The end result of all of this is the graph generation pipeline shown in Figure 3.4. It depicts the rendering process described above over the last several pages and is generally applicable to any internally hyperlinked website with just a few tuning steps applied to the web parser internally.

3.3.3 Synthesising the Final Data set

Although it would be great to synthesise the network graph data for the entirety of the FakeNewsNet data set, there are several factors which restrict this work’s ability to do this. For a start the data set contains articles from over 2,000 different sources. Despite the simplicity and standardisation of the process, it would still require an inordinate amount of work to tune the web parser to each one of these. On top of this, each of the over 20,000 articles in the data set would require significant processing time and bandwidth to generate meaningful network graphs for each. For these reasons this work will limit its scope to the top 10 news sources shown earlier in Table 3.2. It will also limit itself to taking no more than 200 articles from each source.
One factor to consider is how many fake news articles in particular should be taken from each source. It would be possible to balance the data set by attempting to take an equal amount of true and false articles, however in practice this is not nearly the sort of ratio of real to fake news encountered on the internet. Most of the news articles that a model like this will actually encounter on the internet are likely to be truthful, and so this reality is worth reflecting in our data set. For this reason, the final decision that was reached by this work was to take 150 truthful news articles, alongside 25 fake news articles from each source. This would result in a data set with a large enough fake news corpus for the models to make meaningful inferences from whilst still maintaining a similar truth to falsehood ratio as that of the source FakeNewsNet data set itself.

Excluded Domains

Unfortunately a unique problem was observed when this pipeline was applied to www.dailymail.co.uk. Inspecting the results of the graph generation process on this domain

---

**Figure 3.4: Graph generation pipeline**

---

<table>
<thead>
<tr>
<th>Web Parser</th>
<th>Graph Generator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receive Labelled Articles</td>
<td></td>
</tr>
<tr>
<td>For each labelled article</td>
<td></td>
</tr>
<tr>
<td>Parse Article</td>
<td></td>
</tr>
<tr>
<td>Return article contents</td>
<td></td>
</tr>
<tr>
<td>Return linked articles</td>
<td>Store linked articles and parsed contents in labelled DataFrame row</td>
</tr>
<tr>
<td>For each linked article</td>
<td></td>
</tr>
<tr>
<td>Parse Article</td>
<td></td>
</tr>
<tr>
<td>Return article contents</td>
<td></td>
</tr>
<tr>
<td>Store contents in unlabelled DataFrame row</td>
<td></td>
</tr>
<tr>
<td>Return linked articles</td>
<td></td>
</tr>
<tr>
<td>Article count and store within parameters</td>
<td>False</td>
</tr>
</tbody>
</table>
revealed that although the FakeNewsNet provided articles were parsed without any issues, most of the hyperlinked articles were not parsed correctly. Of the 840 internal hyperlinks discovered, only 44 were successfully parsed. This issue appears to be primarily due to the sheer amount of non-article content on the website.

En.wikipedia.org was also chosen to be removed from the data set due to concerns that the frequent editing inherent to posts on that website mean that the labels could be significantly out of date with false articles being corrected, and truthful articles having falsehoods added to them.

Furthermore www.usatoday.com also failed to be parsed correctly in most instances. This could possibly be caused by the website redirecting European requests to a separate domain. Whilst the parser is specifically designed with functionality to handle this scenario, it also appears to create request encoding issues that make the web parser unable to inspect the HTML correctly. This may be resolvable with more advanced decoding techniques or simply by accessing the servers from within the United States, however this is considered beyond the scope of this work and so www.usatoday.com was also excluded.

### 3.4 Experiments 2 and 3: Training Models

#### 3.4.1 Content-Based Model Development

For the content-based model it was decided to focus on implementing an approach that is traditionally considered to be effective at these kinds of problems. A common approach that has been used before when creating baseline fake news classifiers with the intention of viewing how an augmentation impacts performance is a simple logistic classifier using TF-IDF bag-of-words encoding (15). Although typically less effective than more complex models, they are also more predictable and explainable, allowing for more in depth inference from the results of the augmentation. Due to this being an accepted convention it is the approach that this work has decided to follow as well.

#### 3.4.2 Network-Based Model Development

With a content-based approach decided upon it is now possible to augment this model with a network-based approach as an additional feature. Since the use of a hyperlink feature is entirely novel and unexplored it was decided to use Google’s PageRank algorithm. It has proven its utility in many network graph problems before, including fake news problems as discussed earlier. For each labelled article in the data set, it’s PageRank metric within its graph is calculated and added to the DataFrame to aid in training as an additional feature.
In the following chapter these models will be evaluated and compared, but it’s worth explaining briefly here the methodology by which their evaluation will be conducted. The Cross-validation performed on both of the models will be described. This process involves testing several combinations of hyperparameter values to ensure that each model is performing optimally. The hyperparameters available to cross-validate vary from model to model, however in the case of logistic regression there are two values that can be changed.

3.5 Conclusion

This chapter has detailed the work which was done in the process of generating and designing the data set. It details the process in an exact enough manner for replication and showcases what the end result looks like. It then takes this process and synthesises an 875 article training data set with exact instructions to be reproduced.
4 Evaluation

This chapter will take the experiments performed in the previous chapter and critically evaluate their results so that their efficacy may be understood. It will begin by examining the data set that has been created and testing it with a number of different metrics. It will also then detail the cross-validation and evaluation of both of the ML models proposed in the previous chapter according to standard machine learning procedures.

4.1 Data Set Evaluation

Evaluating the utility of a data set like this is not an easy task at all. Even if there were an accepted convention surrounding data set evaluation it would be hard to apply to the data set presented here. Because of this this section will attempt to investigate the resulting quality of the generated data in a more qualitative manner, by comparing and contrasting the end data with what could be considered ideal examples. It will also discuss the limitations this paper has identified with the data set, since lacking any standard evaluation criteria, this paper must create one for itself.

4.1.1 Developing an Evaluation Criteria

One of the few regular methods for numerically assessing a data set is its size. However this data sets size is inherently the same size as the FakeNewsNet data set it is built on top of. Despite this, a more helpful metric that can actually be identified is the success rate of the web parsing technology this data set is built off of.

Furthermore an important part of any fake news classification data set is the quality of the text parsed from the source websites. Although not assessible numerically, this is an example where comparing the actual parsed text to an ideal alternative can prove to be quite helpful.

Finally although the size of the data set may be fixed by the size of the FakeNewsNet data set, it is worth considering the size of the network graphs, since more complex graphs should allow for more complex analysis to be performed with it going forwards.
4.1.2 Evaluation Results

Web Parsing Efficacy

Firstly it is worth highlighting that the data set generation process was ultimately only effective on four of the ten original domains it was designed to be used on. Part of this problem is the age of the FakeNewsNet data set which has stopped receiving regular updates. However fortunately if a new appropriate data set is developed in the near future which meets the criteria specified in the methodology of this paper then this process could easily be reapplied to the new up-to-date data set which would likely improve the success rate of the web parsing significantly. Also exemptions like en.wikipedia.org had nothing to do with the efficacy of this papers approach and were instead barred due to methodological concerns about the reliability of that data. Of the seven web domains that were supposed to be used in the final data set for this project five of them were parsed successfully.

However the width of domains the parser works effectively on is only one way of measuring its efficacy. Another important aspect of the web parsing is how often it works successfully on supported domains. This web parser achieved a successful parsing rate of over 90% on the 825 labelled articles data set used in the previous chapter. This was calculated by first taking the entire set of parsed articles (including node articles in the graphs), and removing any which did not contain parsed text. After this the total number of articles in the data set fell from 9718 to 8807, a loss of only 911 articles in total. This on its own is considered a pretty successful element of the data set, however it doesn’t mean anything if the text being parsed isn’t of useful quality.

Text Parsing Quality

Since the final synthesised data set contained data parsed from four different domains, and the data within each of those domains was parsed in the same way it seems reasonable to consider each of these instances in turn. Beginning with an example parsed from people.com:

People.com TV Bellamy Young Opens Up About Being Adopted, Her Real First Name and How She Almost Missed Out on Scandal Bellamy Young Opens Up About Being Adopted, Her Real First Name and How She Almost Missed Out on Scandal "I was adopted, so I was in foster care for six weeks," the actress tells PEOPLE By Patrick Gomez Updated September 16, 2015 11:00 AM Advertisement FB Tweet More Pinterest Email Send Text Message Print Image Credit: Desiree Navarro/Wireimage Bellamy Young may be riveting as First Lady Mellie Grant on ABC's Scandal, but her off-screen story is just as captivating. "I was adopted, so I was in foster care for six weeks," the actress, 45, says in the

29
latest issue of PEOPLE. “We only had two lines on my dad and a paragraph on my mom. It said she loved to sing, so my mom who raised me would find any way to let me perform.” Get push notifications with news, features and more. + Follow Following You’ll get the latest updates on this topic in your browser notifications. Young says she spent her youth in Asheville, North Carolina, learning ballet and tap dancing and performing at school and church. After graduating from Yale, Young – who grew up with the first name Amy – spent a few years doing theater in New York before making the move to Los Angeles in 2000 to pursue TV and film work. “When I went to join the Screen Actors Guild, there was already an Amy Young so I had to register under a different name,” says Young. “I tried to become Susanna or Violet or something fabulous, but it just didn’t feel like I could carry the ruse off.” So the actress turned to her family for inspiration. “My first [adoptive] dad died when I was 15, and his best friend, Bill, did all the dad stuff with me, so I did a mushing of our names,” Young says. “I felt like I could get away with it because I’m Southern.” After years of hopping from project to project, Young was given the opportunity to audition for the role of Mellie on Scandal. The character only had a few lines in the pilot episode, but Young says she showed up at the first table read for the season excited to work. “[Creator Shonda Rhimes] came around afterward and told everyone what their arc would be for that first season. She got to me, and I was so happy. Then she told me she thought I’d be in about three episodes. She wanted to write a presidential divorce,” Young recalls. “I died inside when I heard that. But I kept a smile on my face and showed up to work every day. At some point, they decided it was fun to write stuff for me,” continues Young, who has also released an album of cover songs called Far Away So Close. “I’m having such a good time. I don’t want it to ever end.” For more of Young’s story, pick up the latest issue of PEOPLE, on newsstands Friday RELATED VIDEO: Why Bellamy Young Said Olivia Pope Would Be ‘Distraught’ on Scandal Season Finale Scandal – along with Rhimes’ other #TGIT hits Grey’s Anatomy and How to Get Away with Murder – premieres Sept. 24 on ABC.

In this case we can see clear extracts of text throughout the article that should not be part of it. We can see inlaid text from headers and meta data such as image captions that would ideally be parsed out of the article. However the majority of the article is correctly and coherently extracted from the url, with these issues likely not impacting the results of our machine learning models significantly. Although it would take up too much space to elaborate on each of the sources here this pattern was consistent across people, etonline, usmagazine, hollywoodreporter and variety. Variety, hollywoodreporter and usmagazine in particular have an issue with vast amounts of new line characters separating pieces of text.
content, however this shouldn’t be an issue for machine learning models since this kind of thing is frequently removed/ignored by most preprocessing methods applied to text such as removing stop words.

Network Graph Complexity

Finally it is important to consider the sophistication of the network graphs this process generates. The easiest and most telling metric for analysing this is to calculate the average amount of nodes in any given graph. Again the final data set from the previous chapter will be used in this stage of the evaluation. That data set contained 700 labelled articles and after network graph expansion it contained a total of 8126 articles. This leaves the data set with an average of over 11 articles per graph. Considering that each of these nodes also contains parsed text contents it seems reasonable to conclude that these graphs are typically sophisticated enough to contain meaningful data themselves. This part of the project has therefore been considered successful.

4.1.3 Data Set Evaluation Conclusions

From the data and analysis laid out above this project has shown that despite it’s limitations as a first generation technology with nothing even similar to evaluate it against, this data set is robust enough and of sufficient complexity for machine learning work to be performed with it.

4.2 Model Evaluation

This section will detail the cross-validation performed to ensure that the final models were using appropriate hyperparameters in both instances. It then calculates standard evaluation metrics and confusion matrices in order to compare the two models.

4.2.1 Content Model Cross-Validation Results

To begin with k-Fold cross-validation was performed using both L1 and L2 regularisation techniques with a range of inverse penalty weightings $C$. The evaluation metric results from this are shown in Table 4.1. At first these results all appear to be remarkably strong, with the lowest $C$ values providing the best performance. However this is only due to the fact that imbalanced data sets such as this one can be easily skewed by traditional evaluation metrics. In order to properly understand these results in the correct context it is important to look more closely at the confusion matrices generated by each of these models.

Inspecting the confusion matrices shown in Table 4.2 shows us that the highest scores being achieved on the models with the lowest $C$ values are simply playing to the fact that most of
the data set is truthful and always predicting that the article is true. However we see that as the value of C increases (and thus the penalty weighting decreases), that both the L1 and L2 models begin to make more false classifications. However past a certain point they begin to make many more false than accurate predictions of falsehood.

With all of this data it is important to decided which version of the model, with which hyperparameters performs the best overall. The measurement that this work has settled on is the number of false positives, since they represent fake news articles that were mistakenly classified as true. This is because the identification of fake articles is considerably more important than the confirmation of true articles in many use cases, and so the lower the number of false positives, the lower the number of fake news articles going undetected. By this measure the best performing content-based classifier uses L1 regularisation with C=1000.

Table 4.2: Cross-validation confusion matrix results for purely content-based model

<table>
<thead>
<tr>
<th>Regularisation</th>
<th>Penalty Weight (C)</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>117</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>13</td>
<td>106</td>
<td>28</td>
<td>705</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16</td>
<td>103</td>
<td>40</td>
<td>693</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>19</td>
<td>100</td>
<td>72</td>
<td>661</td>
</tr>
<tr>
<td>L2</td>
<td>0.1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>7</td>
<td>112</td>
<td>7</td>
<td>726</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>9</td>
<td>110</td>
<td>20</td>
<td>713</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>11</td>
<td>108</td>
<td>45</td>
<td>688</td>
</tr>
</tbody>
</table>
4.2.2 Network Model Cross-Validation Results

The same cross-validation process was applied to the network-content hybrid model and the evaluation data in Tables 4.3 and 4.4 were obtained. Again there is a similar pattern to before with great evaluation metrics until looking at the associated confusion matrices. In this case the best performing model also uses L1 regularisation with C=1000, although with slightly better performance by the False Positives metric than our purely content-based classifier.

Table 4.3: Cross-validation evaluation results for network-content hybrid model

<table>
<thead>
<tr>
<th>Regularisation</th>
<th>Penalty Weight (C)</th>
<th>F1-score</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.92482</td>
<td>0.86030</td>
<td>0.86030</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.92600</td>
<td>0.86264</td>
<td>0.86236</td>
<td>1.0</td>
</tr>
<tr>
<td>L1</td>
<td>10</td>
<td>0.91318</td>
<td>0.84271</td>
<td>0.86967</td>
<td>0.96213</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.90638</td>
<td>0.83215</td>
<td>0.87132</td>
<td>0.94597</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.88688</td>
<td>0.80405</td>
<td>0.87399</td>
<td>0.90407</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.92482</td>
<td>0.86030</td>
<td>0.86030</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.92482</td>
<td>0.86030</td>
<td>0.86030</td>
<td>1.0</td>
</tr>
<tr>
<td>L2</td>
<td>10</td>
<td>0.924220</td>
<td>0.86030</td>
<td>0.86653</td>
<td>0.99060</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.91645</td>
<td>0.84741</td>
<td>0.86672</td>
<td>0.97304</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.89952</td>
<td>0.82046</td>
<td>0.86492</td>
<td>0.93931</td>
</tr>
</tbody>
</table>

Table 4.4: Cross-validation confusion matrix results for network-content hybrid model

<table>
<thead>
<tr>
<th>Regularisation</th>
<th>Penalty Weight (C)</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>117</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td>L1</td>
<td>10</td>
<td>13</td>
<td>106</td>
<td>27</td>
<td>706</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>18</td>
<td>101</td>
<td>41</td>
<td>692</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>22</td>
<td>97</td>
<td>66</td>
<td>667</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>733</td>
</tr>
<tr>
<td>L2</td>
<td>10</td>
<td>6</td>
<td>113</td>
<td>9</td>
<td>724</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>10</td>
<td>109</td>
<td>21</td>
<td>712</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>109</td>
<td>45</td>
<td>688</td>
</tr>
</tbody>
</table>

4.2.3 Comparison

From looking at these two models, it is clear that by all measures the two models are nearly indistinguishable, although the confusion matrix shows a slight but noticable improvement.
There are still many factors left to be investigated with this data set. More sophisticated models need to be used and more sophisticated network analysis needs to be performed but even from this simple example the promise of non-contextual features is visible.

Table 4.5: Comparison of evaluation results for content-based and network-content hybrid models

<table>
<thead>
<tr>
<th>Model</th>
<th>F1-Score</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Based</td>
<td>0.88816</td>
<td>0.80641</td>
<td>0.87515</td>
<td>0.90542</td>
<td>19</td>
<td>100</td>
<td>72</td>
<td>661</td>
</tr>
<tr>
<td>Network Based</td>
<td>0.88688</td>
<td>0.80405</td>
<td>0.87399</td>
<td>0.90407</td>
<td>22</td>
<td>97</td>
<td>66</td>
<td>667</td>
</tr>
</tbody>
</table>

4.3 Conclusion

This chapter has demonstrated the strengths, but also the limitations of the data set synthesised earlier. It investigated this scientifically despite the absence of standard procedure to draw from. It also cross-validated, evaluated and compared the two models discussed in the previous chapter.
5 Conclusion

This chapter will provide a summary of the main findings and contributions of this research. It will also highlight the limitations of this research that need to be considered before suggesting some of the future research that can be done to follow up on what has been done here.

5.1 Main Findings

In this section the original research question behind this research project will be addressed and its success will be evaluated. To do this it is worth revisiting the seven major objectives that were laid out back in the beginning of the dissertation. They were as follows:

1. Research current linguistic fake news classification techniques
2. Understand the current work using network graphs to improve fake news classification
3. Either find or create a large, labelled data set of articles for analysis
4. Create a consistent way of representing hyperlink structures in a network graph
5. Implement a linguistic machine learning model
6. Modify model to take advantage of network-based features
7. Evaluate both versions of the model to determine if the network-based features offered any advantages over the traditional model

Objectives 1 and 2 were detailed in the literature review and showcased via implemented in the methodology of this dissertation. Objectives 3 and 4 were the largest part of the methodology and in the opinion of the author represent the greatest success and contributions of the entire work. Finally objectives 5, 6 and 7 were all focused on the model training and evaluation that took place after the data synthesis had been completed. Whilst 5 and 6 were completed soundly, objective 7, and by extension the research question at the heart of this paper, was inconclusive with the improvement being too small to objectively consider significant. It was perhaps too ambitious to try to definitively prove a matter like
this with a single piece of work. The data set that ended up being created is too versatile to have all of its use cases investigated.

Over the course of this project several significant discoveries were made. The first is that web parsing, although error prone, is absolutely viable on a large scale for news article data collection. Secondly is that access to quality and up-to-date data is crucial, with most of the major data sets reaching several years of age and running into problems with urls or entire domains being changed. Thirdly is that this field remains with large gaps in the research that deserve further exploration in the future, which this work more than anything else aims to facilitate going forwards. Now with the task of data synthesis implemented and enumerated on in this dissertation it should hopefully be viable for more sophisticated machine learning models to be trained using this data.

5.2 Limitations

The ambition of this research topic meant that limitations on its findings were inevitable. However there are four fundamental limitations that should be highlighted here. Firstly, this dissertation is built on top of quite old data that is starting to show its age. Secondly due to processor and time constraints the final models were trained on a mere 875 article data set with only 125 fake news articles contained within. This is simply not enough data for machine learning models to be fully trained on. The third limitation of these findings is that the machine learning models that were used were very simple. This was important to ensure fair comparisons could be made between them however it means that there could still be other kinds of models much more capable of taking advantage of the network features presented here. Fourth and finally, the network features were taken advantage of in a very crude and rudimentary fashion in this work. The data set is capable of much more sophisticated analysis than the simple PageRank algorithm utilised here.

5.3 Future Work

This field of research is so new and so broad that this project was barely able to scratch the surface of what remains to be done. Obviously the data set generated as one of the core scientific contributions of this dissertation can now be used in future research to try and develop more sophisticated machine learning models. There are also many other features that can be investigated alongside network spatial features. It is hoped that the flexibility and modularity of the web parsing and graph generation processes proposed in this dissertation should help to accommodate more work in this field of research going forwards. An example of the kind of additional feature that could possibly help to improve fake news classification performance would be the collection of publication dates. This would allow for
full spatio-temporal analysis to be performed which could be an effective evolution on the ideas presented here, although it was beyond the scope of this dissertation. Moreover it is hoped that this model of web parser and graph generation algorithms can have applications entirely beyond fake news classification. This model, with slight if any modifications can be used anywhere on the internet that internal hyperlink structures are prevalent. Also due to the fact that most news sources only embed links to other news articles also from their website this project made the decision early on to focus on internal links. However it could possibly be even more useful to somehow incorporate external urls if a robust enough web parser was developed. This could even open up the possibility of creating a large interlinked network graph of fake news across the internet, however this was obviously beyond the scope of this single dissertation.
Bibliography


