Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Determining the features most influential in predicting user engagement in news articles

Jack Mac Namara

April 30, 2021

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

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: _____

Summary

The research objective of this thesis is to determine the features that are most influential in predicting different user engagement metrics in order to provide a foundation of knowledge upon which to base future research. As the current emphasis in the area is based around creating optimal models for predicting the user engagement metrics, it is important that this foundation-level research is performed so that researchers know where to begin model creation. Most of the features used in this analysis are available for extraction on other websites using only basic libraries and so the work can be easily replicated.

As stated, the current research environment is heavily model-focused in this area. The research that does focus on which features determine user engagement metrics tends to use a small number of features or a specific measure (for example, only looking at the effect of topics or using the comments in an initial time-span to predict future comments). As such, although this is useful research, there is scope for a broader look at the features affecting user engagement metrics, which is what this thesis aims to address.

The data used in this research was scraped from the-journal.ie, the most popular Irish online news website. The features extracted were analysed before the experimentation and determined to have some probable link to the user engagement metrics. The features were cleaned so they could be used for machine learning in the logistic regression model chosen. The model allowed for the analysis of the features in predicting the most popular articles, in this case the top 10% of articles for each user engagement metric were isolated. This allowed experimentation to be performed producing the coefficients for each of the features corresponding to their relationship to the user engagement metrics. The magnitude of these coefficients was used to determine the relative influence of these features in predicting the specific user engagement metrics.

Identifying the importance of these features is the major contribution of this research. The research shows that there is a strong relationship between many of the features extracted from the articles and the user engagement metrics. In particular, the title and blurb seem to have a larger effect on many of the forms of interaction when compared to the content of the article. This leads to the conclusion that many of the readers decide on interacting before reading the article, which has implications for future research in this area. Other notable findings are that figure-based information in the title and blurb are negatively correlated to views and comments, and that time-related features have a strong predictive relationship to sharing and viewing while almost no relationship to commenting or the polarity therein.

These relationships show the strength of the connection between the features used and the user engagement metrics, but due to the performance metrics it can be concluded that other features which were not analysed also play an important role. This research can be used

to create more optimal models for predicting the user engagement metrics and also provides a platform for further research into the relationship between these and other features in determining these user engagement metrics.

Acknowledgements

Firstly, I would like to thank my supervisor Owen Conlan for all his help and guidance throughout my research. He provided assistance whenever necessary and without his support this project would not have been possible.

I would also like to thank my friends and family for their continued support throughout this year. In particular my dad who, through his experience and guidance, helped me over the last few months, and Mark Doody, Áine Harte, Brendan O'Connell, and Gabriel O'Rourke who, through their advice and support, helped me to produce my thesis.

Abstract

User engagement plays an important role in the consumption of news media. Research in this area has focused on creating optimal models to predict this user engagement but has placed less emphasis on the features used to produce these models. In this research the features will be analysed to determine the correlation found between the selected features and the chosen user engagement metrics on an article and the reasons behind these correlations.

Using data scraped from the-jounal.ie experiments were set up using machine learning models to predict the user engagement metrics. These models used feature sets containing the data the reader had access to before engaging with each metric. The data used was gathered from the website, and external libraries were used to extract more features for the final feature set. The models produced coefficients relating to each of the features that determined the relationship between those features and the user engagement metrics.

Identifying the importance of these features is the major contribution of this research. The research shows that the title and blurb were more important than the article as predictors of user engagement, likely meaning readers decide on interacting before viewing the article. This has implications for the use of features from the article as predictors of user engagement, as it appears that the decision to interact often occurs before the reader views the article. Another key finding is that figure-based information in the title and blurb is found to be negatively correlated to both the number of views and comments received. The more concise the information provided in these sections, the less need a reader will feel to view them and these articles also tend to be less controversial and therefore receive fewer comments. It is also found that time-related features have a significant impact on the viewing and sharing metrics for an article, while having a substantially smaller effect on commenting behaviour. Overall, this research finds strong correlations between certain features in the articles and the user engagement metrics analysed.

Contents

List of Figures

List of Tables

1 Introduction

Online news platforms are becoming increasingly relevant in the modern consumption of news, and unlike traditional media the internet provides a platform that facilitates the interaction between readers and the news organisation in real time (8). This means that readers can express their opinion to a large community of people through comment sections or share articles to one another using social media. This interactivity is drawing a large proportion of the population to use online news platforms as either their primary or a supplemental news source. As seen in Figure 1.1, the proportion of users in Great Britain has grown by 250% in the last 13 years.

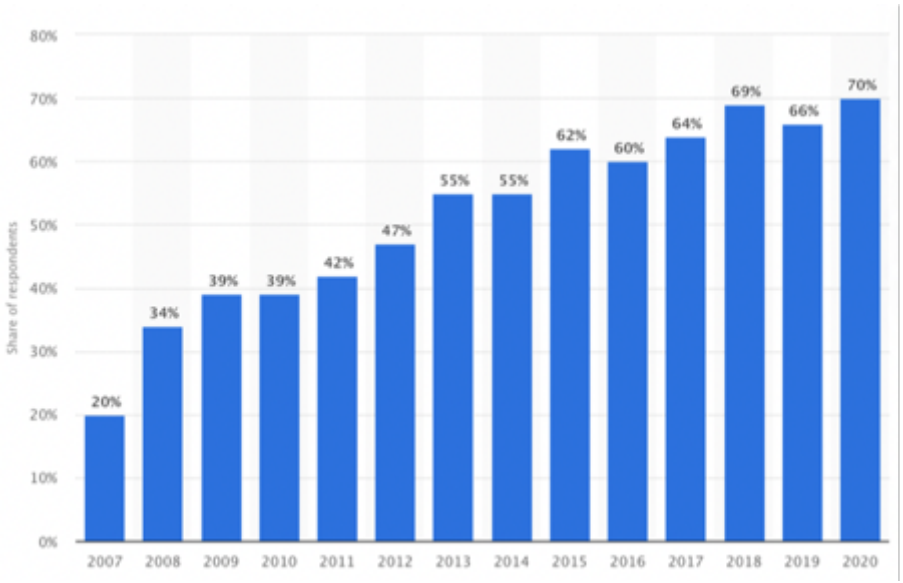

Figure 1.1: Share of individuals reading or downloading online news, newspapers or magazines in Great Britain from 2007 to 2020 (37)

Due to this increase in popularity, the influence that online news organisations hold is becoming more prominent. This comes with a large amount of responsibility due to the ability of news to control the way that stories are framed and presented to the reader. News organisations choose which information they feel most accurately portrays the details or narrative of a particular event and therefore have the ability to shape public opinion towards an event. The way in which they do this can have drastic repercussions for how events are perceived in the

public consciousness such as the effects of the media's coverage of the British royal family's 2019 trip to Africa (24). In this way, the media has the power to be both positive by educating us about world events and informing us about topics that greatly affect us, and negative by exacerbating the spread of disinformation. Although the sensational topics that often appear alongside disinformation or "fake news" tend to be good for the business in the short term as they create headlines that elicit reactions in the reader and are more likely to be interacted with than the mundane, they also decrease the societal trust in the news and this decrease in trust can have a negative net effect on mass news media as a whole. This is illustrated Figures 1.2 and 1.3 below which analyse Americans' trust in mass media.

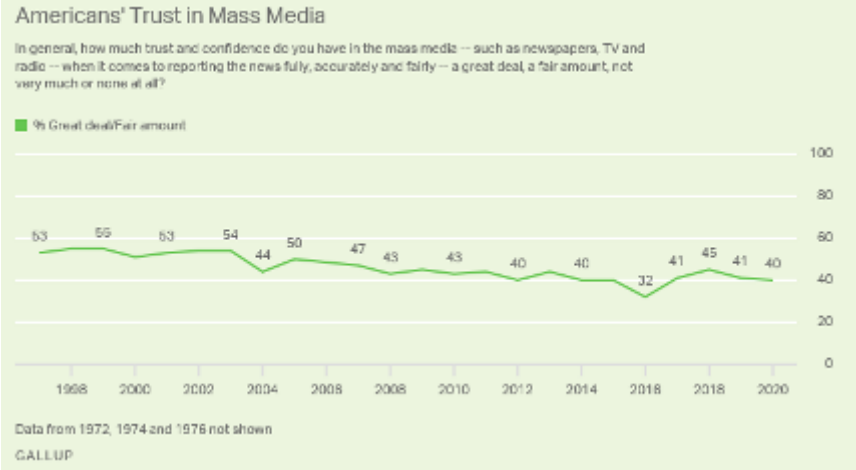

Figure 1.2: Americans' Trust in Mass Media (6)

The trend above is exacerbated when the graph is broken down by political party.

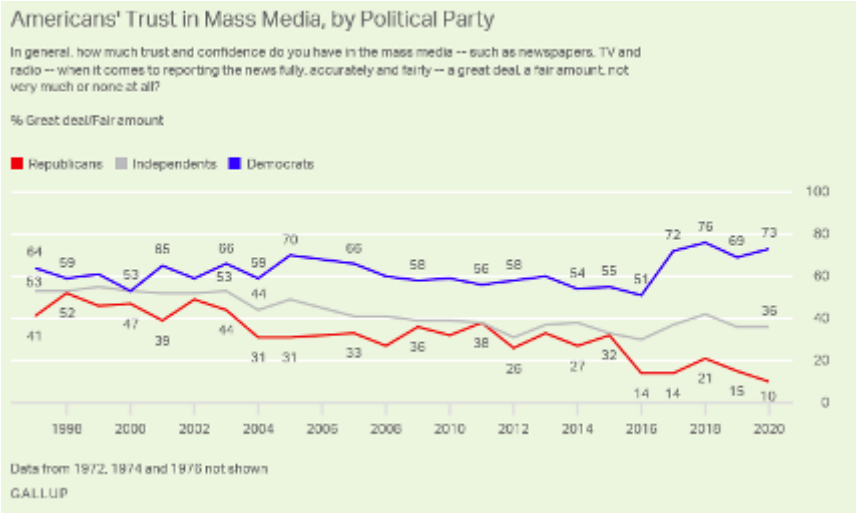

Figure 1.3: Americans' Trust in Mass Media, by Political Party (6)

Part of the reason for this increasing distrust of news media is due to the changing landscape of media caused by the internet. It is no longer a monopolistic market where a few large companies dictate what information the general population has access to. This was due to

the level of the barrier of entry into this market and the difficulty in finding like-minded people in a more closed off world. However, the internet has removed these barriers by making information sharing easier and faster than ever. This means that stories are more widely shared, increasing the ability of people to learn more about the issues that affect them and be introduced to stories from around the world that they would have otherwise had no way of accessing.

This access to more news content has also allowed for the creation of niche markets that are able to focus on individuals or groups with specific interests. This allows much more information to be shared and discovered in niche areas and allows groups of like-minded people to connect and discuss common interests. Although benefits include expanding public knowledge and allowing for information on all topics to be much more readily available, users are encourage by the amount of material online to gravitate towards content that confirm their biases, particularly through social media sharing and through biased online sources as described by Fletcher and Kleis Nielsen (7). News organisations are in a struggle to maximise their readership and user engagement, by either targeting the general population with content that incorporates the biases of the general psyche, targeting those in minority groups that have biases opposed to that of the general population, or by trying to provide unbiased information that provides users with facts of events and allows them to reach their own conclusions (this last form of news is hard to create and often not as enjoyable to consume, so tends to make up only a small percentage of the news available).

1.1 Problem Area

The way in which news media is consumed has changed rapidly since the invention of the internet, the focus of both the readers and the news organisations has moved to the online space. This has changed the way in which news organisations provide content and how they measure the success of their content. In the past, the goal had been to sell as many newspapers as possible in print media, this had meant writing stories with catchy headlines on the front page and including content that encouraged readers to choose their paper over another, such as a sports section etc. This is no longer how many of the online news media platforms operate. There are many different models that exist including subscription-based newspapers such as The Irish Times, and The New York Times, which have subscriber only sections and don't provide all of their content for free. This encourages them to provide informative articles that readers are interested in and cover a large range of topics. There is another model which has been created solely due to the invention of the internet and that is revenue through ads. Many recently established news organisations operate in this way including the-journal.ie, which is the focus of this thesis. As these websites are only paid based on the user engagement and, in particular, clicks (because this is what will show their

readers new ads), a difference exists in the way they tend to market their stories. Accordingly, they aim to optimise views and interaction on their content because the more interaction the users provide the more money the news organisation can make. This does not mean that the sole focus of these organisations is to get the highest possible number of views on an article at the expense of quality as that would not be sustainable over the long term, so a compromise needs to be made between attracting user engagement and providing quality content to their users.

Unlike traditional print media, online news media has a greater scope for user interaction. There is now the ability for users to provide feedback to the authors through the comments section which can be interacted with in real time, as opposed to letters to the editor which limited correspondence. Additionally, there are many more forms of user engagement now available such as liking and sharing on social media platforms. As it can be accurately tracked by computer software, this information is much more readily available with online news media as well as the readership numbers. Again, these forms of engagement are important as they allow news organisations to make money and also determine how readers feel about their product. The way this is done varies greatly between news organisations with some providing many more forms of interaction and others limiting it only to views, while not allowing any interaction between readers or between readers and the news organisation. These interaction types and how readers are using them are also changing rapidly much like the increase in the growth of online news. This is due to the changing preferences of readers and the insights that news organisations garner from other forms of media attempting to make the process as enjoyable and engaging as possible for the readers.

One of the main struggles that news organisations face is that they are looking to maximise user engagement in a space in which total user engagement is finite. All news organisations are competing against one another and against other forms of media for user's attention. This encourages news organisations to provide readers with sensational, attention grabbing headlines as seen in the research of Ransohoff and Ransohoff (25), which finds that "sensationalized reporting receives, by their very nature, a disproportionate amount of attention". News organisations are now using this as a tool to increase user engagement on their websites. This is a known feature of media that can lead to an increase in engagement but there are many other features that also contribute to a user's interactions. Although much work has been put into this area there is still a lack of understanding of the importance of a wide range of features and their effect on user engagement. Specifically the degree to which certain measurable features are important in determining the user engagement of news articles.

There has been similar research to this performed by different researchers, however the features have rarely been at the forefront of this research and the research has been focused more on optimising models in this area to create the best predictor of certain features such as the

number of comments (34). This looked at a number of different features but the focus was on the models, using a number of different models to determine which one optimised the problem. This is a trend that is seen in this research where the models are optimised without looking specifically into the effects of each of the features in determining the strength or weakness of each. The number of features is also limited in this research as they are looking to optimise models, this means that there is little research in the area which examines the impact of a wide array of features on the user engagement metrics.

1.2 Research Objectives

As most of the research in this area is focused on optimising the models that are used, it fails to address the importance of the individual features and the implications which this can have outside the scope of improving the model. Determining the magnitude of these features has the effect of allowing both researchers and news organisations to determine what exactly is causing readers to perform certain actions and understand the behaviour driving the user engagement. This is the space in which this thesis aims to work, producing features that can be analysed to get some bearing of their effect on user engagement and to determine the extent of this effect relative to the other features. An important note on this research is that the models will be unoptimized due to it not being the primary focus of the research, this will affect the model metrics but not to any degree which will skew the findings in this thesis. Removing and adapting certain features in the models would increase the success of prediction but would provide a less detailed look into the effects of the different features.

As there will be a range of features used it is important to determine where the features and metrics come from. Some of these will be explicitly stated in the gathered data from thejournal.ie website and others will be inferred from the text provided and the way in which the readers interact with the articles. All of the features used were determined to have some potential to affect the user engagement based on past research in the area focusing on social media and news articles. Defining these specific features is important in helping to understand the results of the analysis. As there are a large number of features and multiple metrics which they are being experimented against, this research aims to provide a detailed analysis on the effects of these features on each metric. There will be some features that are important for multiple metrics and others that only have a strong correlation to a single one and this will help in showing that user engagement can't be judged as a single variable. Each metric has its own important features and the individuality and overlap of this importance for each of these will be assessed and analysed to produce the most statistically significant to each individual metric as well as to user engagement as a whole.

The main metrics into which this research will be looking is the number of views, comments, Facebook shares, and e-mail shares that an article receives as well as determining the polarity

in the comments based on the same feature set. It also looks at the comments, based on a limited number of features, to determine the quantity of replies and likes they receive. Each of these metrics have different research and industry benefits and, depending on the goals of the news organisation, different ones may be optimised or a balance between each may be found. This research aims to provide an analysis of why the relationship of each of the features and metrics is the way it is and of the similarities and differences between these relationships across different features and metrics.

1.3 Contributions of Research

As this research is focused more on the features used for the model than the model itself, it aims to provide a list of these features and their relative importance to the user engagement of an article. This will benefit future research in this area in a few different ways. Initially, it assists researchers in the area of model creation by determining the optimal features to begin with depending on the type of user engagement that is being analysed. This can save time for users performing these experiments as well as improve models by suggesting new features that had not been considered for a specific user engagement type. It also provides an area for further research into the features selected, although this research uses a large number of metrics that were selected based on ease of availability and having an inherent link to the outputs (meaning not using article features in determining the views as this information would not be known at the time). The features used were not an exhaustive list so there is scope for further research into this area focusing on different feature sets to determine other features that could be important to the user engagement of articles.

It has contributions to the field of news article analysis specifically due to its links between the features and user engagement but also to a smaller extent to any online content with user interactions. Many of the user engagement metrics and features that were used were not inherently specific to news articles and are available on a wide variety of online content so it is likely that some of the research will be transferable and will be able to be applied to these other areas such as blogging (which has many similarities with a main portion of text followed by user interaction). There is an area for further research to corroborate this fact and see if the research is applicable to these other areas or even just if this research is applicable to other online news articles. As this was done in Ireland on a subset of news articles that are not subscription-based, there are many other news organisation types to which these features could be applied to see if there are similar results. It is possible that this research has implications on all forms of news media to some extent and could be used as a template in many of these areas to provide background knowledge on the features and their impact on the user engagement in many of these cases.

As will be seen later in this thesis, it also corroborates past research that has already been

performed in this area. This includes a number of papers in which similar experimentation was performed, and whose claims are further supported by the findings in this paper. In many of the cases it also builds on this data with new information related to the topic.

Outside of the area of research it has implications for both news organisations and the wider community as a whole. News organisations can use this information to optimise their articles by using the features provided. This can help them attain a larger share of that finite user engagement that is available by focusing their stories and their writing styles on topics and methods that readers are more likely to engage with. This could have an impact on their overall ability to produce content and make their articles more engaging to their readership and depending on the ability of the information to transfer across platforms it could also be used by other organisations that provide content in a similar way to optimise their content. As for the general population this information is useful for them in determining the reasoning behind some of their actions and providing information on what makes them engage the way they do. This information obviously will not apply perfectly to all individuals but it is likely that the majority of the population engage in a way dictated by the features and their influence demonstrated in this thesis.

2 Literature Review

2.1 Online News Media

Online news media is becoming a more important part of the general population's news consumption. Between 2007 and 2020, the proportion of the population using online news sources as either a primary or a supplemental news source has increased by 250% as seen in Figure 1.1 (37). This research was performed in the UK which has historically strong social ties to Ireland meaning many of the findings are likely highly transferable. The study also found that both television and the internet were both more likely to be a source of information over newspapers further reinforcing the change in the modern consumption of news. This has many benefits for society including the increased availability of up-to-date news due to the speed and ease at which the internet can be used to upload and distribute content, and the ability to provide users with avenues to interact and engage in an increased capacity. There are commenting sections that provide readers with a platform to open a discourse, share ideas and debate on the current news topic with both the author and other readers, while links to social media allowing users to share their stories on Facebook and Twitter bringing the conversation and debate from the article to other platforms and in doing so also opening up the article to a wider audience.

The new forms of user engagement being added to news sites are changing the way that readers consume and perceive news. The comment section has added an extra layer of information that readers experience when reading an article. Lee (15) describes the way a readers opinion changes about both the news article and their beliefs on public opinion based on the content evident in the comment section. Comments can mislead readers into misattributing the opinions in the comment with the thoughts of the author in the article. This shows the power that has been placed in the hands of the readers with the ability to actively engage in this manner. Readers are also more likely to "extrapolate public opinion from impersonal others' comments than from one's conversations with friends and family". This is the danger around this area and it can often lead to confirmation bias among readers with a lack of diversity in their reading habits. Those who read articles from news organisations with similar biases to their own could be unconsciously conditioned into thinking that their own

beliefs align with popular opinion based on the comments from other readers that they are seeing. This is compounded by social media which shows content based on user preferences and so will push content that will get interaction from the user, which in many cases tends to be content that aligns with the user's worldview. Fletcher and Kleis Nielsen (7) further examined this and how people on social media are more likely to be incidentally exposed to news, further helping news stories spread which is the desired effect of many of the large news media organisations by providing options for users to share their stories. The downside to this new ecosystem that is created is that many users tend to gravitate towards news stories and news organisations that align with their current world view which compounds their bias. Fletcher and Kleis Nielsen (7) also found that this exposure did not affect all users equally as incidental exposure was increased for those who use multiple online sources or were categorised as young people, potentially due to their increased computer literacy. This increased interaction with multiple news sources allows an increased consumption of news meaning they are less likely to experience bias. This is supported by Knobloch-Westerwick and Kleinman (11), who also concluded that this confirmation bias was strongest in those with infrequent online news consumption and those whose ideas most align with public opinion, (in the research scenario, those whose political party is likely to win the election). This is not the case for those who frequently use online news as they tend to not exhibit confirmation bias to the same extent.

Almgren and Olsson (2) analysed the use of the sharing feature and its effect on the dissemination of articles through different forms of social media. They were able to find a difference in the user base between different social media such as Facebook and Twitter as platforms for sharing news. Whereas Twitter was more likely to be used in cities for topics on a national scale, Facebook was more localised and shared news from local agencies, and stories more relevant to one's area and connections. This represents an ecosystem where users are much more aware of not only current affairs in their locality, but global news from around the world, in large part due to social media. Much like the differences outlined by Almgren and Olsson (2) in the types of news on different social medias, Szabo and Huberman (30) and Lerman and Ghosh (17) noted a difference in the dissemination of news in similar research. Although sites like Digg had a fast spread of news in local circles, the spread of news stories slowed significantly when promoted to a large number of unconnected users. News stories on Twitter, conversely, start spreading slower but continue at this rate and are more likely to penetrate deeper into the network due to the high interconnectivity of the users. The way in which news stories spread across different social media platforms is important to developing an understanding of the different types of articles that news organisations produce and the reasoning behind these differences.

Although most news organisations have a specific target demographic in mind for their news stories, like local news agencies creating stories specifically for a local audience, there is a

driving force both within and between news organisations to produce content that garners a large proportion of the attention. This is not a new phenomenon and has been occurring in the news media almost since its creation and results in sensationalism in the media. Newspapers such as the Sun in the UK exemplify this by using headlines meant to grab attention even if it misrepresents the contents. Zuckerman and Litle (38) found that individuals tended to be what they described as "sensation seeking", this showed a correlation between the morbid or sexual nature in media and sports events, and the curiosity aroused in participants. This helps explain why sensational topics tend to get more views due to this same arousal of curiosity. As with many forms of media, some articles tend to become more prominent than others and receive a large proportion of attention and the total engagement for the news site, as noted by Szabo and Huberman (30). Ransohoff and Ransohoff (25), believe that subtle incentives provided to journalists in part lead to sensationalism, in many cases this being a story that will perform better with some minor changes that improve the significance or excitement provided by the story. This however misleads public opinion, leading to a less educated population, specifically in the case of politics as mentioned by Ransohoff and Ransohoff (25) as this is generally considered a highly controversial topic with much room to sway public opinion. Sensationalism provides an avenue for journalists and news organisations to receive a larger proportion of the finite user engagement available in the field of news media as sensationalism grabs disproportionate amounts of attention. It however makes up only a small proportion of the total ways in which a user's engagement can be swayed ranging from the day of publication to the topic being mentioned. There are other features that can be explored and ideally allow news organisations to focus their attention away from these sensational methods of getting user attention and focus more on other features that control user engagement without being inherently misleading and counterproductive to the main reason for news which is to educate the population.

2.2 Text Analysis in Online Media

Text analysis is an important field in online media. It allows us to understand the different features that go into writing in a more in-depth manner than was available before computers were able to be used for data processing. This has led to many discoveries in the way we write, why we write the way we do, and how this can affect the way the reader perceives the text. It also allows us to identify the specific features that contribute to the sentiment and polarity of the text, and due to this there has been a lot of research performed in this area in recent years. Melville, Gryc, and Lawrence (18) performed sentiment analysis on blogs to determine the optimal method for identifying sentiment. The research discusses how much of the previous work in the field uses models that build on previous knowledge by using known data polarity in text to determine the polarity in a new text document. They build on this by developing a new method that incorporates lexical knowledge into supervised learning for text classification

to allow more accurate sentiment analysis to be performed. The state-of-the-art models are further expanded by Mohammad, Kiritchenko, and Zhu (21) which examines sentiment analysis as it pertains to tweets. This will be useful as they contain a similar structure to that of the comment section in the-journal.ie, much like blog posts contain similar structures to that of an article. Although the state of the art models will not necessarily be implemented due to constraints in time and resources, as well as the effects being quite small due to the nature of the work as it does not pertain to optimising models, the findings are still useful in understanding the reasoning behind their implementation. Mohammad, Kiritchenko, and Zhu (21) determine that sentiment lexicon features, both manually and automatically created, lead to an increase in performance of their models, meaning there is merit in both. For this research paper, it is the manually generated features that we are more interested in, as this allows analysis to be performed on them and provides reasons for these features' importance. The research also looked into the increased performance with the use of n-grams for both word and character n-grams as a means to lead to increased performance in the models. This use of n-grams was also identified to hold importance in identifying hate speech by Nobata et al (23). The n-grams have the ability to cluster connected words that alone would not be seen as important and can be missed by approaches that don't take into account these subtleties.

One of the most important features in identifying the user engagement in an article is in the topic that is being discussed. There are a number of different ways in which this can be performed all of which have different levels of success. Mishra and Vishwakarma (20) used a number of different models to determine the best for isolating important terms in the document and determined that TF-IDF (term frequency-inverse document frequency) was the best model out of all of its variants and was able to produce the highest precision score. For this, model pre-processing was performed to achieve a sense of consistency in the documents before the algorithm was performed. The model identified in this paper will be implemented as the initial step of topic extraction in this work. Once the key words in the documents have been identified work will need to be done to cluster these key words and determine common topics throughout the data set. Again, there has been a lot of work in this area with many different algorithms being used but one of the most consistently used for clustering in a research capacity is k-means clustering. Wartenna and Brussee (36) determined that topic detection by clustering a set of keywords and using a k-means algorithm works well and is relatively inexpensive compared to other models. It is best used when the number of topics can be determined by some other metric as it needs to be initialised beforehand. For this, there have been a few different metrics that were deemed to be good fits in this space. They are the cost function as determined by the k-means algorithm and silhouette clustering - both of which are useful to some extent in determining the number of clusters that should be used. Rosseeuw (27) sets out the framework for silhouette clustering and how it is used. The

silhouette outlines how similar an object is within its own cluster relative to that of another cluster. This is helpful in determining how different two clusters are and the average silhouette score can be used as an overall metric for the fit of a clustering algorithm with k number of clusters. Further research into this algorithm was performed by Wang et al. (35) in which a simplified version of silhouette clustering was used in combination with k-means clustering to determine the effectiveness relative to other metrics such as the k-means cost function. It was determined that they were both able to identify similarities in topic clustering with the k-means cost function being the most optimal and efficient method. It appears an approach that took into consideration both of these methods would be useful for this experimentation as both methods provide valuable information.

There is similar work to identifying topics based on text performed in areas such as Twitter in which information about the topics that interest a user can be determined from their likes and the posts they write. Michelson and Macskassy (19) looked into the determination of user interest topics based solely on their generated content to see if the common topics could be extracted and related to the users interest topics. They did this in a similar way to the manner described above in the other research by discovering the entities within the posts, and then determining the high-level categories defined by these entities. In doing so they were able to analyse the resulting categories and generate a topic profile for different users, which resulted in relatively accurate accounts. These results are encouraging for the method selected in this thesis for topic detection as it demonstrates that it has the ability to work in a similar online field. Ahuja et al. (1) also performed similar analysis in determining the best feature extraction technique and classification method for determining sentiment analysis. In this paper, they compared n-grams and TF-IDF side by side and concluded that TF-IDF was the better method for sentiment analysis performing on average 3-4% better. This was one of the main reasons TF-IDF was used in this project over n-grams. It also measured the use of TF-IDF against six different classification models and determined that logistic regression produced the best results for all metrics (accuracy, precision, f1-score, and recall) for TF-IDF. Although this research was performed on Tweets, the level of detail and comparison between different model types provided good direction for the selection of the optimal model and feature extraction techniques in this space.

2.3 Text and feature analysis in articles

There are numerous ways in which features can be extracted from articles and these range from surface level features that are explicitly defined such as the length of the article or the number of comments it receives to implicit features such as those in the area of sentiment analysis, entity detection, and polarity. Some of this content is generated by the news organisations and the rest is provided through user engagement - often through mediums set up by the

news organisation on their site. One feature that is particularly useful for readers is the comment section. Many comments found on news articles are similar in style to content found in other forms of online media, including Twitter and Facebook. In most cases there is an initial post, Tweet, or article followed by many comments and replies. The analysis of these interactions is integral to understanding how users engage with one another in online media. These interactions can be reactions to the initial post, or users can create a discourse in the section below the post and talk about the topic of the post with one another, and this can lead to topics of discussion not intended by the initial author. This is an important area in user interaction as all information, including tone and sentiment, must be displayed through words alone as body language or other physical indicators are not available (unlike in person interaction). Discussions can often span over days and include multiple people, but most online media sources have implemented systems for keeping track of these discussions to some degree. These tracking feature were not always necessary to be implemented by the news organisation as noted by Scuth, Marx, and de Rijke (28), users will look to create a discourse even when it is not set up in an easy way by mentioning the name of the commenter they are responding to or including an @ symbol in front of the username to get their attention. This shows the importance of this interaction to the readers as they will tend towards creating a discourse even when the tools are not set up for them to do so. This form of discussion analysis will not be necessary on the data set for the-journal.ie as the comments are structured in such a way as to account for these discussions.

User engagement is encouraged by news organisations in order to give readers a sense of power over the content and features such as likes and comments are enabled to allow this. The way in which news websites do this varies depending on the organisation and Lehmann et al. (16) determined there were a number of different models that news organisations could be broken up into. The models were grouped as either general, user-based, or time-based; each of these were made up by a number of different models meaning news websites could share common features across general and user based models but have a different time based model. This provided different but complementary insights into the user engagement and its diversity on these news sites. One key finding from this research is that sites of the same type (e.g. mainstream media) do not necessarily follow similar engagement models meaning that there is some reason outside of their audience causing sites to create their user engagement models in certain ways. This area was looked into by Stroud, Scacco, and Curry (29), who found that the features can vary between online news platforms and these differences tend to arise from the sites target audience or whether the news site is showing content originally for a newspaper or television news. This is interesting as it implies that many news organisations who have transitioned partially or in full to online news have brought engagement types from their original platforms and continue to view that form of engagement as more important than other news organisations do. The study also found that the more broadly targeted outlets

were more likely to have interactive features and see higher levels of use. In the case of the-journal.ie the main functionality provided to the user is commenting, liking, viewing, and sharing the content. As it is a solely web based news platform it focuses to a larger extent on the user engagement in this online context. Due to it beginning on the internet it is more likely to utilise the potential of the internet as shown by its ability to link to other forms of social media through the internet and having a large set of features the users can engage with. This is seen in newer media as it is a way to easily increase user engagement by giving readers more options on how they want to interact.

A key metric in the context of online content is determining what counts as a view and what does not. This topic is explored by Lagun and Lalmas (14), in which the different types of views are defined and a brief investigation of each is provided. The four types of views are determined to be bounces (quick views of less than 10 seconds), low engagement (less than 50% of the article read), deep engagement (over 50% of the article read), and complete engagement (when the article is interacted with by either sharing or commenting). This represents a much deeper level of insight into the time of exploration than it is usually given and the metrics that influence these different outcomes is interesting research. Features, such as the presence of media elements, encouraged the readers to dwell longer on a page but had no effect on the overall determination of if the reader read the entire article. The same was determined about a feature, such as the length where it was determined that the effect was negligible and that higher levels of reading were more dependent on the content and the way in which the story was presented. These advanced metrics, in terms of identifying the readers dwell length, require a large amount of experimental data that is not readily available on most news sites as this information requires a lot of tracking and analysing of users experience which is not allowed outside of consensual experimentation. This means that it will not be available for the purpose of this thesis, as using a metric like this would severely inhibit the number of other features that can be used and measured and the overall scale of the thesis. This information does however give us context in which to analyse the view metric determined by the-journal.ie, which requires an engagement level which would at least be considered a low engagement by Lagun and Lalmas (14).

2.4 Article features determining user engagement

The main area of research in which this thesis will be focused on determining what features are important in predicting user engagement. There has been a lot of research in this general area but a relatively small amount of it has focused on the specifics that relate the importance of different features in determining the user engagement and the correlation between these parameters that affect the control the news platform have over the popularity of their articles. Some of the features of the articles are outside the control of the news

organisations and depend on the events currently occurring in the world such as election cycles - as discussed by Boczkowski and Mitchelstein (5), who analysed the effects on key areas of user engagement such as most clicked, most emailed, and most commented on stories during periods of heightened or routine political activity. Although this is a feature (event) outside of the control of the news organisation, it still produced interesting findings around the change in user behaviour over a period of time. In this case, the beginning of a heightened election cycle tended to produce a large number of comments when compared to the number of clicks and emails that an article received. This changed over the course of the election cycle as later in the period the latter two features became more prominent. This research outlines a flaw in the work done in this thesis, as the number of features that would have been needed to account for popularity over the course of every month of every year would have overwhelmed the data. Instead, this was broken down by month and year (disregarding the importance of any particular month and year combination). This information is evidently important but requires a large amount of analysis into the events that occurred over the time period as well as a much larger focus on these features. This information further helps me define the scope of the project so that the area analysed produces some unique findings that are relevant and useful to the field.

One interesting feature that was not addressed in the paper by Boczkowski and Mitchelstein (5) was the relative difference between the user engagement types based on the topic as it solely focused on political news and related content. To address this, rather than looking at the types of user engagement relative to one another for the same topic Tenenboim and Cohen (33) analyse the effects of different topics in the types of user engagement that they incur. They noticed that although many items had high levels of user interaction, the way in which the interaction occurred varied depending on these topics. Sensational topics and those with curiosity-arousing elements tended to feature more prominently in the heavily clicked articles, whereas those with political/ social topics and controversial elements tended to be among the group of heavily commented items. The overlap found between these two items was between 40-59% which is a large disparity for items that could all be considered popular based on user engagement metrics. This is a key area of research and one of the driving factors behind my research as it demonstrates clearly that all types of popularity in articles are not the same and are determined by different features. In terms of Tenenboim and Cohen (33) the features that were analysed were the topics and the difference between sensational topics and controversial ones. There is reasoning to believe that if this kind of disparity is true for these features it would also be evident in other features relevant to determining the user engagement in articles. The fact that sensational news garners a large proportion of the total views and user engagement is not a new idea and has been used by many news companies for hundreds of years in an attempt to sell more papers. This notion is described by Zuckerman and Litle (38) as they identify that individuals tended to be what they describe as "sensation

seeking". Their research showed a correlation between the morbid or sexual nature in media and sports events and the curiosity aroused in participants. This helps explain why sensational topics tend to get more views due to this same arousal of curiosity.

There is also research confirming the findings of Tenenboim and Cohen (33) on the importance of the controversiality of topics in determining the number of comments an article would receive, notably the work of Ksiazek (12) which corroborates the findings that controversial topics tend to lead to more comments. This research delved deeper into the types of comments left on these articles and they concluded that comments left on controversial topics tended to be more hostile, potentially due to the nature of the debate. This area of research on the civility of comments is substantial due to its implications on the type of discussion and users preference for discussion in the comment section. Ksiazek (12) further determined that items such as multimedia in an article led to more comments but also tended to lead to a more hostile discussion. Another key finding of this research was that the intrusion of the journalist into the comments which had often been considered a potential breach of the user space by other researchers was seen to improve the civility of discussion in the comments. This finding was consistent after changing many other features and was likely due to the users seeing the journalist interaction as a positive as it allowed them to actively address their concerns to a professional in the area. Another feature that was examined was the effect of the use of sources on the comments. In terms of quantity this did not seem to have a noticeable effect but it did increase the hostility in the comments. This finding on hostility actually contradicts previous work done in the area by Ksiazek, Peer, and Zivic (13) which found that sources led to less hostile comments. However, the latter study was performed on a larger data set over numerous publications so likely carries more weight in this determination. This article clearly outlines a few key features that are strongly correlated to the number of comments received, due to the constraints of this thesis some of the key article features were extracted and used in the model creation whereas other features such as the use of sources or journalist engagement in the comment section were not as they were less easily identifiable. This is an area of possible further exploration in this research as determining the strengths of these features relative to the rest of the features examined in this research could be important for news organisations in deciding the approach they will take in attempting to optimise user engagement.

Most of this research is based on collecting large amounts of data from online news sites and using this to determine not only user engagement but user sentiment through avenues such as the comment section. Arapakis et al. (4) decided to perform similar experiments but in person rather than gathering data from online. This used a combination of sentiment and polarity metrics derived from the content and offline behavioural measures. Measuring sentiment and polarity in the reader based on the content of the article is important in understanding their interest and engagement in the article and analysing the degree to which this occurs can

help us understand differences in reactions from users. As the sentiment in these discussions depends on the content of the article, it is important for there to be a way of identifying these sentiments and analysing to what degree they are occurring, as the variation in sentiments can lead to a difference in reaction from users. Arapakis et al. (4) determined that genre was one of the most important factors that affects this sentimentality and polarity and there is a high variation depending on the genre. This is similar to the differences found by Tenenboim and Cohen (33), as factors that inspire human curiosity and drive attention were the most important factors in determining user engagement. Another interesting discovery of this work was achieved through the ability to examine users as they read news. By using an eye tracker, they found that readers were likely to look at interesting titles for longer and more likely to return to these topics, possibly indicating that buzz words are likely to draw and catch users attention making them more likely to read an article. This article highlights the importance of sentiment and polarity as metrics due to their effect on the mood and external behaviour of a user outside of the online news sphere. This is one of the reasons why polarity of comments was chosen as a metric for the research as it allows an insight into not only the level of user engagement but also the type of user engagement and how this is also affected by features present in the article.

This was similar research to Lagun and Lalmas (14), who also performed in person experiments in determining important article features. This research was more focused on determining the different types of view metrics, but also performed limited research into determining some important features such as content and the way the story is presented – finding that these matter more to a user reading an article than features such as length. They also performed analysis into how users read, such as spending the majority of time in the upper sections of the article. These findings, although rather limited in the scope of the area, provide a good basis to build off in terms of feature analysis. Jiang et al. (10) provided a deeper analysis on features affecting the user engagement on a news article focusing on the number of views that an article received. For this, there were a number of features selected relating to the layout and presentation of the headline. It was found that elements such as headline position, format, length, use of punctuation marks (square brackets or round brackets), recency, and popularity, aroused significantly different numbers of clicks on the headlines, respectively. This is a good analysis on the surface level features that affect the user in selecting different forms of media and shows the limitations that are present in developing a model to determine overall user prediction due to the large number of features in different areas that would be required. Although this data is evidently important in the prediction of views, the research is rather limited and constrained by only focusing on a small number of features relative to the user engagement. Tsagkias, Weerkamp, and De Rijke (34) address this by analysing a much larger feature set. Their work is very closely related to the work performed in this thesis with two main differences. The first is that the work in this thesis uses more features

for prediction of user engagement, with their work only focusing on 24 features, and more modes of user engagement, whereas this research just focuses on the number of comments an article receives. The second difference between the research lies within the research objective. Tsagkias, Weerkamp, and De Rijke (34) want to optimise their models by experimenting with many different model types and determining what feature set is optimal (the features are broken down into groups such as surface level, semantic, contextual, and cumulative). Although the research does a good job at breaking these down and providing a breakdown of the optimal models, it provides no evidence of the importance of individual features or which combination of features outside of these groups is optimal. This is less useful to news organisations analysing this data as the specific features to be analysed are not evident and only an overview of feature type is provided. As such, this is an area in the research that needs to be addressed.

Ambroselli et al. (3) performed similar work, focusing on the number of comments received as a measure for user engagement. This research looked at a number of features individually and the level to which these features were able to predict the number of comments an article would receive. This research is valuable to this thesis as it is one of the only times in which the effects of specific features were measured as opposed to solely optimising a model. This only made up a small portion of the research with a limited number of features and the rest of the work was focused on the models. This research was intended as a way to notify news organisations of articles that were likely to cause a large number of comments at an early stage to avoid having to shut down comments sections due to the nature of the comments in a heated debates. The research was able to produce relatively good models with the selected features but again highlights the need for a more thorough examination of features in this area. The research also focused on how the number of comments in a short period of time after the article is published could be used as a prediction method for the total number of comments and found that there was a correlation that increased quickly over the first hour of publication. This area of determining the total number of comments by using comments received early on was also examined by Tater et al. (32), and Tater et al. (31). Tater et al. (32) focused not on predicting the total number of comments received but the relationship between these articles by using a popularity ranking. They looked at this relationship over a longer period of time than Ambroselli et al. (3) who focused on the first hour after publication and extended this to the first day. After only two hours the rank correlation was determined to be just under 75% and over the next 24 hours this increased to 95%. This is likely due to the fact that the majority of comments are on newer articles and after this period of time there is a severe drop off in the number of new comments. This research was further expanded and corroborated by Tater et al. (31) who determined which models were the best at predicting this feature. There were no new findings in terms of the correlation between the data and the results show the difference in the prediction models is small, but they still determined that

the optimal model for this type of prediction was a simple linear regression model.

The link between this initial number of comments and the final number means there is likely similar correlations between the other methods of user engagements such as liking and sharing, and this is an area that can be discussed in future research. There are many features that were identified above that can be looked into to predict levels of user engagement for a specific article, most of the research to date has been focused on producing optimal models for determining user engagement metrics or uses experiments that have been performed with a small feature set. This leaves a gap in the research in identifying the correlation between the user engagement metrics such as the number of likes, comments, and shares that an article receives and the user features such as the entities in the texts and the topics of the articles. The importance of each of these features relative to one another is also a gap in the current research and will be examined in this thesis. The research of the relationship of these features to the news habits in Ireland and more specifically the journal, is an area that is important for future research in the area and has scope to provide interesting findings.

3 Design

The research goal for the thesis is to determine which features are the most influential in predicting the user engagement metrics in news articles. The research performed in this area is often focused around optimising models for prediction as described in the literature review section. This means that a gap exists in the research for determining the specific article features that are important in predicting user engagement metrics and the extent to which they are influential. This section will discuss both the methodology of the thesis and the manner in which it is designed . It will define the data set required for the experimentation and discuss the process of data gathering that will be performed to acquire it. It will further explain the process by which the data is cleaned and the manner in which the relevant features are extracted. It will also include an in-depth discussion of the machine learning model chosen, comparing its parameters and the way they affect the analysis of the results. The reasons for the decisions made in each of these sections will be discussed at length.

3.1 Defining the Data set

In this section the choices surrounding the selection of data and the basic features required will be discussed. It will lay out the reasons behind these choices and the possible effects they could have on the experimentation.

3.1.1 Data set Selection

As this thesis looks into the importance of features in news articles, it is necessary to select a data set that is as general as possible. For this reason, a news source that was not in any niche area and had a wide variety of news stories available was required. It was also necessary to select a news source that did not have a subscription-based model as it was found that only 12% of Irish consumers pay for news (9) so a non-subscription-based model would be more representative of the news habits of the general population.

In selecting the data set, the ease of collection was also considered. There are many large databases with historic news articles available online, particularly through many libraries. These papers are often too old or don't provide the required information with regards to user

metrics. An Irish newspaper was also required due to the knowledge of the cultural background that could assist in the analysis of the results. For all of these reasons it was determined that using a web scraper on an online Irish news website would be the optimal way to obtain the necessary data set.

3.1.2 Required Features

There were a number of metrics required in the data set for the analysis to be performed adequately. The most important metrics were the number of views that the article had received and a method for determining the number of comments. These were the two most important experiments as they build directly on much of the research already performed in this area and would likely provide new insights into this area of research. Other metrics that were looked at, as they could also contribute to interesting work, were the number of Facebook and Twitter shares on each article. There also needed to be some way of extracting features to predict these metrics and so certain features, such as titles, blurbs, and some time-related features including the date, were necessary to provide enough features for the experiments.

3.2 Data Gathering

In the data gathering section, the focus is on the collection and storage of data. This section is important due to the type of data set that was selected. The planning around the collection of this data was necessary to ensure it ran smoothly and that all of the correct information was gathered and stored in the appropriate manner (preventing the need to re-scrape any of the data).

3.2.1 Database

A database was selected as the storage type for the data for a number of reasons. Compared to the other methods for storing the data such as a large excel file, databases are easier to manage. They allow for easy separation of data through different tables, which removes the need for duplication in the data set. They also allow a much larger amount of data to be stored when compared to large excel files, as single excel files would not have been able handle the amount of data required. These were the two main methods compared when deciding on storing the data retrieved from the web scraper.

It was also necessary to select a database that could be stored locally on the computer, so that it was not reliant on external servers to run. The database also needed to have the ability to be accessed through python as that was the language used in this thesis. Even though many different databases allow this it was necessary to consider to ensure that there were no

issues with the data after it had been gathered.

3.2.2 Web Scraper

A web scraper was used to retrieve all of the important information. Due to the libraries available through python, it was easy to design as there are many built-in features that do much of the heavy lifting. As the full feature set had not been fully defined at this point, it was necessary to create a web scraper that collected all sections of the data and that all of these sections were clearly identified before the scraping began, so that relevant code could be implemented in the scraper to retrieve them.

There were a few considerations that needed to be made in selecting an appropriate web scraper, these included it having the ability to handle the dynamic content found on most news websites and handle time based events to accommodate these. Speed was not an important factor due to the time allocated to this section of the research to ensure it is performed correctly.

3.3 Cleaning and Extraction

In this section, the data was prepared for creation of the model. This meant determining the important features from the extracted data and producing new features from the text. These features along with the important metrics were used for the final feature set and predicted values for the model respectively. All features extracted from the data were stored in excel files at this stage. This was in order to reduce the need for the duplication of work in extracting these features and to organise this data into relevant sections which was easier to do in excel than a database with these small sections.

3.3.1 Extracting features

Once the data is scraped off of the website a further level of cleaning and extraction takes place to ensure that all relevant features are extracted. The process is different for each of the features but involves a mixture of external libraries and data analysis to create. These are broken up into three different categories: those that can be retrieved through basic extraction, those that require external libraries, and those that require performance analysis, data manipulation, and the use of external libraries.

The most basic of the features to obtain are those that only require a basic extraction. These are the features such as the length of the article that can be extracted easily from the scraped data. The other features that fall into this category are all of the length based features, as the in-built python len() function handles this, and all time related features, as the only extraction necessary is relating the day of the week to a value of 0 through 6 or similar for the months

and years. These factors have been used in other analysis to determine similar metrics and as they were easy to retrieve were considered good initial features to extract.

Natural Language processing Models

This research looks into a few data types that have not been analysed to a large extent in other research. These are the polarity and entities in the text, as they provide strong links to much of the existing research in the area and are an area in which analysis is needed. These two features were both extracted using natural language processing.

The research requires both the creation of features, including the polarity and the entities in the title, blurb, and article, and the creation of user engagement metrics in terms of polarity in the comments. The features are necessary as they are some of the key areas this research aims to explore, by analysing these features in relation to the user engagement metrics. The polarity in the comments will also allow a greater insight into the sentiment of the readers towards the content of the articles.

The requirements for the extraction of these features are that the overall sentiment in the different sections of the articles can be retrieved and that a list of entity types can be extracted from each of the article sections. The performance of these models needs to be sufficient for the research but is not required to be flawless due to the nature of the work. Due to the scale of the research small inaccuracies in the models can be accepted.

Topic Extraction Models

The final feature extraction that was performed on the data was for the topics. This is an important feature set as it has been found to have a strong correlation to the user engagement metrics in other research. It is likely that the data set used will not divide the articles into different sections as this is not common among most news websites, which means that topic analysis and extraction must be performed to determine which topic each article falls into. The two main stages of this are determining the importance of words in an article, this was performed using TF-IDF vectorizer, and the clustering of the similar words, which was performed using the k-means algorithm.

TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) is used to determine the most important words in each document compared to the importance of those words in other documents. In theory, this provides a topical overview of the article as the words that are used most frequently compared to other articles in the corpus are likely to be the most important to it. The benefit of TF-IDF is that it provides a metric based on the most important words in an article that can be used to perform clustering on the articles. This cannot be performed

with just the article text, as models, such as k-means, need numerical data to work. The main drawback of TF-IDF is the fact that it is based on a bag-of-words approach and so is not able to capture position in text or semantics. Due to the size of the data set, this is unlikely to cause major problems as even minor errors in the case of individual articles should become less important over the entire data set. There has also been research into using TF-IDF in sentiment analysis confirming that it is suitable for the task (1).

K-Means

K-means is used to sort the articles into clusters based on their TF-IDF results. K-means is widely used in research and there has been research into using k-means as a clustering technique for topics that have determined that it is a suitable model for the task (36). K-means is highly flexible and allows for easy adjustment of the model if problems arise. It is also a model that produces easy-to-interpret results. The main drawbacks are that the number of clusters is not determined by the algorithm and must be decided before-hand, and that the algorithm randomly selects starting points so repeated use can produce different results. However, due to the way the data set is distributed it is unlikely that this would cause significant problems for this data set. The first of these problems is addressed by performing two sets of analysis on the data set using the k-means algorithm. The first is the computation of the silhouette score (which is a measure of how similar objects are to objects in their own cluster, relative to those in other clusters). The second metric used in this analysis was the SSE (sum of the squared errors), this is used to measure the variation within a cluster. Wang et al. (35) performed an analysis on k-means with similar metrics and found both were good at determining the optimal number of clusters with SSE being marginally better. Both will be used to ensure the optimal number of clusters is chosen in this thesis.

3.3.2 Data Sorting

After the relevant features were extracted they needed to be organised and stored for use in the models. This is important to avoid replication of work in extracting the features before each model is run and to ensure that it is easy to find each of the relevant features when they are needed.

The experiments will be broken up into three different sections for each stage of user interaction throughout the experiment. These are before the reader has clicked on an article, after the reader has opened the article, and when the reader is in the comments. Each of these different stages requires a different feature set and it is quicker and more consistent to create this feature set before any of the experiments are run than at the start of each experiment. The feature sets differ by the features available to the reader at each point. This means that the number of views are not related to any of the features seen in the article as the reader is unable to see them at the point they click to view an article. The features related to individual comments

are also reduced due to these experiments looking at this data in isolation, therefore other features such as those related to the title, article, and blurb are ignored.

In storing this data there were again two main options similar to those present at the start for storing all of the data: using excel or a database. In this scenario excel was selected as the optimal solution as it allowed for easier analysis of the data. As the volume of data stored after extraction was small compared to that initially gathered, for example no full article texts, excel was also better at handling the extracted features. It also only contained the relevant data for each specific group of experiments which also made it much easier to manipulate than the initial data set. The set-up required to create and maintain new database tables is also slightly more difficult than the ease through which new excel tables can be created and therefore excel was chosen as the format for data storage.

3.4 Machine Learning

In this section, the design and decisions made concerning the selection of the machine learning model used in the experimentation, will be discussed. A machine learning model is used to predict the user engagement metrics using the features, and in doing so produce the coefficients showing the relationship between them. For this reason, it was necessary for the model that was created to be explainable, providing the results of the solution in a format that can be understood by humans, as this would allow for the important features to be analysed properly.

3.4.1 Model Selection

Choosing the best model for the problem required an analysis of the problem and how it would be approached. As the research aimed to understand the relationships between the features and the user engagement metrics, it was necessary to select a model that allowed this analysis. The two main models identified were neural networks and regression models. The main difference between these two model types is based around the hidden layers, in the fact that neural networks contain them and regression models do not. This is important for our model as one of the key features that was discussed above is its ability to be explainable, which is not the case for most neural networks. Therefore it was decided to use a regression model in this analysis.

There are many types of regression models, the two most prominent of these are linear regression and logistic regression. Both produce coefficients in a similar way and are therefore both useful in this type of analysis. It was decided that the best way to find the correlation between the features and the user engagement metrics was to select a cut-off point and focus on the metrics with the highest popularity, for the most part. This meant selecting the model

that best dealt with binary output decisions which is logistic regression.

3.4.2 Logistic Regression

Logistic regression is a regression technique based on the logistic or sigmoid function. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. It is used by statisticians in many areas as a prediction tool.

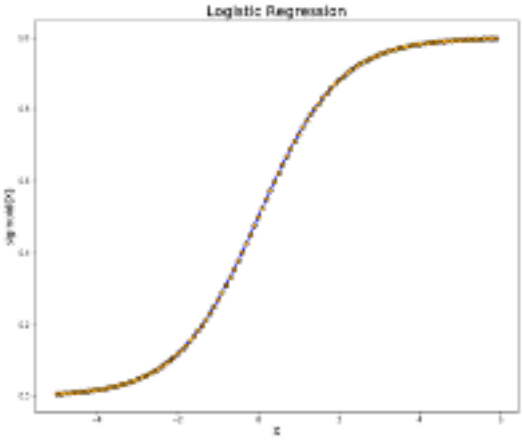

Figure 3.1: Logistic Curve (26)

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}$$

Figure 3.2: Logistic Formula

The formula in Figure 3.2 is a representation of how logistic regression works with the x values being the inputs, and the beta values being the coefficients. This predicts the probabilities of the given class being the default class, similar to the way linear regression predicts the value, and this probability prediction is transformed into the binary values.

The benefits of using logistic regression are its ease of use, both in terms of implementation and interpretation, as well as it being efficient to train. It also makes no assumptions about the distribution of the classes in the feature space making it good for this type of data set. It is also very fast at classifying due to the production of a model beforehand and it provides both a measure of how appropriate a predictor is and the direction of its association, both of which are needed for accurate analysis of the coefficients and their relationship to the user engagement metrics.

The major limitation of logistic regression is in its assumption of a linear boundary between the features and the metrics. This means that it is not as accurate at determining non-linear

problems. It constructs linear boundaries and determines which side of the boundary the inputs fall into. This separation means that it is important that there is at least some form of linearity between the inputs and the outputs to produce an accurate model.

The positives and negatives of logistic regression were taken into consideration but due to the objective of the research which in effect is to determine if there is some sort of linearity between the features and user engagement metrics, it makes sense to use a logistic regression model. There are a few decisions that had to be made once logistic regression was selected, such as the type of regularisation to be used.

Regularization

There were two main types of regularization looked at for use in the experiments. With logistic regression both an L1 penalty, Lasso, or an L2 penalty, Ridge, can be used. Each has its own benefits, with Lasso removing unused or unimportant features from the model and Ridge reducing them instead.

Ridge: Ridge regression is an extension of linear regression, where the loss function is modified to minimize the complexity of the model. This modification is done by adding a penalty parameter that is equivalent to the square of the magnitude of the coefficients. It reduces the size of unimportant features without removing them completely. Therefore, it will not make any definitive selections of optimal parameters for the model.

$$L_{ridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2$$

Figure 3.3: Ridge Regression Formula

Lasso: Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is also a modification of linear regression. In Lasso, the loss function is modified to minimise the complexity of the model by limiting the sum of the absolute values of the model coefficients. It works well with a small number of significant parameters and is able to set insignificant parameters to 0 removing them from the model without the need for feature selection.

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m}|\hat{\beta}_j|$$

Figure 3.4: Lasso Regression Formula

Both of these forms of regularization have positives and negatives for their use in these experiments. Ridge regression has been chosen because of the research objective of this

thesis. The objective is to discover what parameters affect the different user engagement metrics and to what extent they do so. This means looking at all of the features and not just those that are most influential. Lasso regression would only leave the most important features and would make it harder to analyse the effects of the less important features as all of the unimportant values would likely be set to 0. This does not mean that these features will be statistically significant and the results may conclude that some of the features appear to have no effect but to analyse them in this manner Ridge regression must be used. It is also important to note that in future work in this area, for the purposes of building a better model and only looking at the most important features, Lasso regression would likely be better a better choice.

3.4.3 Cross-Validation

Cross-validation is a common method that is used to test how well a model generalises to unseen data after training by using re-sampling. It is used to help identify if a model is prone to over-fitting of the training data. 5 fold cross-validation is used for these experiments and works by splitting the data set up into 5 smaller sub sets. Each of the models is trained five times, and each time four of the subsets are used to train the model and the other one is used to test the model. These results can then be averaged out to give a better estimation of how the model will react to unseen data by effectively allowing the model to be tested on unseen data multiple times for the same data set.

3.4.4 Model Metrics

Analysing the performance of the model created is an important factor in ensuring that the model and more importantly the coefficients received from the model are reliable. Good performance by the metrics means that there is a stronger correlation between the feature set and the outputs, and in this case means a stronger correlation between the data set and the user engagement metrics. The metrics used will be analysed, as will the reason for using them in determining the reliability of this model.

In this discussion positive does not correlate to good and negative does not correlate to bad, it is the way to differentiate between the two classes.

Accuracy

Accuracy is a measure of the number of correct predictions made out of the total number of predictions. This provides a good measure of the effectiveness of the model in data sets that contain even distributions between the classes.

$$Accuracy = \frac{TP\ (True\ Positive) + TN\ (True\ Negative)}{TP + TN + FP(False\ Positive) + FN(False\ Negative)}$$

Figure 3.5: Accuracy Formula

Precision

Precision is a measure of the correct positive predictions out of the total number of positive predictions made. The value determines how often a positive classification is actually positive.

$$Precision = \frac{TP}{TP + FP}$$

Figure 3.6: Precision Formula

Recall

Recall is a measure of the correct positive predictions out of the total number of positive classes. This value determines how often a positive classification is identified from a positive class.

$$Recall = \frac{TP}{TP + FN}$$

Figure 3.7: Recall Formula

F1-Score

F1-Score is a measure that takes into account both the precision and the recall. For a model like this, in which we want to optimise both metrics as neither is inherently better for the outcome, we want to also maximise F1-Score.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Figure 3.8: F1-Score Formula

ROC AUC

ROC AUC (Receiver Operating Characteristic Area Under Curve) is a measure used to determine the ability of a model to distinguish members of the positive class from members

of the negative class. A score of .5 is the score that a model that was randomly guessing would receive. This in effect determines the "skill" of the model and how effective it is at distinguishing the classes.

3.4.5 Experiment Layout

All of the experiments were laid out in a similar format. Initially, the relevant features for the experiments were selected and retrieved from the excel files. For all data sets where the data had to be broken up further to isolate the top ten percent of articles, this was then performed, whereas the data was not changed for the polarity in the comments. The data sets were made even, removing extra data points from the larger class so there were the same number of classifications for both. This made it possible to avoid situations as would arise in the top ten percent of data where the most popular class was selected and the model would always predict the most popular class as this resulted in the highest number of correct predictions.

The formatted data would then be used to determine the optimal cost value for the logistic regression model using cross-validation. The best performing model is then run again on 80% of the data set to create the feature coefficients. These coefficients are extracted from the model so that they can be analysed. The remaining 20% of the data set is used to test the model and produce performance metrics for it to determine the overall correlation of the feature set used and the user engagement metric predicted.

3.4.6 Result Analysis

As the primary objective of the research is to determine the link between the features selected and the user engagement metrics, it is important that the results are analysed in a sufficient manner. This means that a plan to analyse the results is put in place before they are collected. It is important to note, that due to the likely differences in the magnitudes of the coefficients collected from the logistic regression models and depending on the size of the cost value selected using cross-validation in logistic regression, the coefficients cannot be directly compared across models and as such their relative magnitude as well as the results of their metrics must be considered. The size of the coefficients can be used in determining the relative importance for each of the individual engagement metrics, but in determining the importance overall the relative importance to predicting the user engagement metric must instead be used which will require a further analysis than just using the coefficients.

By using the coefficients and the relative feature importance to the user engagement metrics, the most important features will be determined and an analysis of the reason behind the important features will be performed. This will provide news organisations and

other researchers with information on the most important features overall and for each of the specific user engagement metrics and will allow further research into the specific areas to be performed.

4 Implementation

This section will discuss the implementation of the thesis. As the research goal for the thesis is to determine which features present in an article are most influential in predicting each of the user engagement metrics, the implementation must ensure that the correct data is gathered for the features and user engagement metrics, and that the experiments are performed in the manner described in the design to produce the optimal results. There were a number of issues identified over the course of the implementation that caused problems, but solutions were found to keep the work consistent with the design. The choices made will, where necessary, be described in-depth to demonstrate the reasoning behind any changes made and the impact those changes had on the work in the thesis. As in the design the data set selection and the gathering of data will be discussed along with any issues encountered in these sections. This section will also discuss the cleaning and extraction of this data and how this cleaned data is then used to perform machine learning. An overview of the method used to analyse the results will also be performed.

4.1 Data set Selection

The data set selected needed to contain all of the features and metrics specified in the design section. This was so the experimentation and analysis planned in the design section could be performed. The news website the-journal.ie contained all of the features and metrics required and even had some additional features that had not been initially considered. This data set had consistent formatting to run a web scraper on and fit all of the requirements specified in the data set selection section of the design. It has a wide range of topics and is the most popular online news site in Ireland (22), which makes it the ideal website for this analysis.

4.2 Data Gathering

In this section the data collection and storage methods used in the thesis will be discussed. There were a few issues in the data gathering section related both to the scraper and to its interaction with the database. These led to a much slower data collection process than

initially planned. As the process began started early in the life cycle of the thesis, it allowed time for this data collection to take place without having a significant impact on the timeline of the thesis. The collection process resulted in a large data set, with just under 100,000 articles and nearly 4,000,000 comments recovered. The issues arising from scraping such a large data set and the solutions reached to solve them will be discussed in this section.

4.2.1 Database

A mySQL database was selected to manage the data as it was easy to use and provided all of the features mentioned in the design, such as allowing for the separation of data and allowing for a much larger amount of data to be stored than an excel file. MySQL is an easy database to set up and use, and also has a relatively low cost om computer memory and CPU usage compared to other databases, which is why it was chosen for this thesis.

The database was set up to deal with the data being scraped from the-journal.ie website. One of the main advantages of using a database was that it allowed for the data to be saved individually by row, meaning that the only data needed before a commit was made to the database was the name of the table the data needed to be inserted into and the data to be inserted. This ended up being another advantage the database had over using an excel file or similar data storage models, as these would have required the whole data set to be saved each time a new row needed to be added. This has little effect on a small data set, where the python program used can easily handle storing all of the data locally. However, due to the time it took to run the web scraper, the size of the data set that needed to be stored would have significantly slowed down the python program and any crashes or internet freezes would have resulted in all of the data being lost. This is discussed more in a later section.

The database contained four tables with sections for all of the data that was needed from the website. The tables were connected through keys and allowed for clear storage of the data, removing the need for duplication in the data set. The tables stored in the database are described below.

ID int PK	name varchar(255)	authorhref varchar(255)	fbshares int	twitterShares int	emailShares int	views int

Table 4.1: Authors mySQL table

ID int PK	name varchar(255)	accounthref varchar(255)	numberComments int	likes int

Table 4.2: Commenters mySQL table

ID int PK	authorID int FK	name varchar(255)	link varchar(255)	content int	blurb varchar(1000)
fbShares int	twitterShares int	emailShares int	views int	date varchar(255)	

Table 4.3: Articles mySQL table

ID int PK	commenterID int FK	articleID int FK	comment text	likes int	commentID int	replyID int	date varchar(255)

Table 4.4: Comments mySQL table

Each table stores an initial ID assigned to it from the database and the rest of the cells are features scraped from the website. Foreign keys were used for identifying connected tables, such as the commenterID in Table 4.4 which links to the primary key in Table 4.4. All of the lengths were assigned to be longer than necessary as the size of the data storage was not a concern.

4.2.2 Web Scraper

The web scraper was used to collect all of the relevant information from the web page using selenium as the tool for scraping. Selenium was selected as the scraper due to its compatibility with python and google chrome, but due to the similarities between many of the web scrapers all of which perform the same function, any could have been adapted for use in this work. Selenium allows for the manipulation of web pages and is often used as a testing tool in industry. The chrome-driver function allowed selenium to directly interact with the chrome browser and control all aspects of it. This provided some important features such as the ability to handle dynamic content and perform time based actions linked to other sections of the code as discussed in the design. This feature would allow the scraper to perform actions that are not available through other scrapers.

The drawbacks to using selenium were the size of the program, as it becomes quite large due to the external libraries used, and the speed at which it gathered data. It is slow compared to other web scraping libraries however, the data was gathered at a very early stage in the project so these drawbacks did not affect the project to a large extent.

On the-journal.ie website, selenium was necessary to isolate the individual sections of data in the HTML, using chrome developer tools, so that they could be identified by the selenium scraper and extracted. This meant finding unique markers in the HTML either through class names or tags present in the HTML. Some of the data was difficult to extract due to the

manner in which it was stored on the page but through manipulation of the selenium web driver it was possible to retrieve all of the relevant data.

The scraper used the URL for the-journal.ie to open up the selected page with a list of stories on it. The URLs for each of the stories on the page were scraped and stored in an array. The scraper then moved through each of the URLs collecting all of the relevant data from the pages it opened. Even though only some of the features were displayed on the initial page with all of the stories, all of the data including the title and blurb were displayed in each individual story.

Figure 4.1: Typical article layout

In Figure 4.1 some of the key features of the article have been identified. These are only the features that appear at the top of the article, but other features such as the comments and the author appear at the bottom of the article. Item 1 is the title of the article, and all of the features related to it, such as the entities and polarity, are extracted using this text. Most of the sections here were easy to find as they were clearly identified in the HTML. Item 2 is the blurb, item 3 is the date, item 4 is the number of views, item 5 is the number of comments, item 6 is the Facebook shares, item 7 is the e-mail shares, and item 8 is the article text.

Initially, it was planned to retrieve the Twitter shares from the article too as this was another interesting user engagement metric that the journal provided, however there was a problem in doing so using the scraper. This was caused either by the way the scraper was written or a problem with the-journal.ie not allowing for Twitter shares to register correctly, meaning that none of the Twitter shares for the articles were recorded. As such, Twitter shares had to be excluded from the final user engagement metrics and an analysis on the relationship between them and the article features was not able to be performed.

The first problem encountered with the running of the script occurred when the script was initially run. A cookie pop up appeared, which occurs before the article is opened the first time the journal is accessed using a new web browser. Due to the manner in which the web driver was running, it caused this screen to appear most of the time, but not all of the time.

Another difficulty was that the pop up did not appear immediately and so required the web driver to wait and then check if it was there. Once the problem was identified it required some manipulation of the web driver to ensure that it always performed the wait correctly, then checked if the pop up appeared, before selecting the right option so that it could continue scraping. The pop-up that appeared is shown in the diagram below.

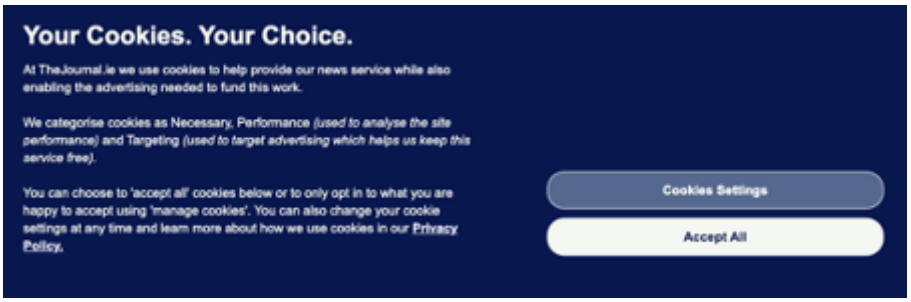

Figure 4.2: Journal cookie pop-up

As this slowed down the process, along with the time it took for the scraper to run by itself, the scraper ran slower than expected. At this pace it would have taken the scraper over two months to retrieve all of the data. This was if there were no interruptions, such as the internet dropping which would cause the code to fail and force the process to restart from the same point when it was noticed. To speed up the process, it was determined that using multiple python threads concurrently was the best method to shorten the time it would need to run to collect all of the relevant features.

Setting up multiple threads with python is easy as it only requires starting the code from different pages on the journal website and running the same script in different terminals. This put some pressure on the CPU but this was often run at night when the computer would have been inactive, negating any major ramifications to using the laptop. This was possible due to chrome-driver: the specific web driver used in selenium, which allowed for scraping to be performed in the background and meant that each scraper did not try to take up the screen whenever a new web page was opened. This led to some issues with concurrency in the database as multiple transactions could not be performed at the same time. However, these were solved by checking if the database was currently being accessed. If it was, the scraper would wait a random amount of time and check again. This could have also been accomplished using keys instead, however due to the simplicity of implementing the wait system and the overhead of implementing a key system from scratch, the system with waits was selected.

Using the database for storage caused a number of problems due to the manner in which the data was being collected. However, compared to using other file systems such as a CSV or excel sheet, it appears to have been the optimal method. Problems, such as the internet dropping, which would have caused data loss with these other methods, only caused a single

instance of the scraper to fail and this could be easily restarted from the specific place of failure. It also meant that issues such as concurrency of data were much easier to work around due to the structure of the database. There were a few implementation changes that would have shortened the overall time needed for data collection. However, due to the time available these were not necessary, as the time was better spent elsewhere on the thesis and these methods would have resulted in no change to the data collected.

4.3 Cleaning and Extraction

This section analyses how the data was processed before being used in the models. Due to the nature of the data set, considerations needed to be made about how the data was sorted into groups so that they could be used for their respective experiments. There were also some considerations needed with regards to the performance of the feature extraction methods, to ensure the methods used the optimal parameters for extraction. The scraper collected a large amount of data all of which was not necessary for the experiments, this meant determining the correct data, extracting it from the current data set, and storing it in an easily accessible format.

An initial analysis of the data set led to some interesting findings about the distribution of the data. The plan had been to use a data set from the end of 2020 stretching back to the middle of 2010 as this would provide a wide variation in the data, due to the cultural change over this time-span, and would also consist of a large data set. After the data had been gathered however a problem was noticed in its distribution due to the problems mentioned in the scraper. As can be seen in Figure 4.3 there are problems with the data where there are seemingly large chunks of data and sometimes complete chunks missing over time-spans. To combat this the area of research was focused on the most complete section of the data set as shown in the graph below. This was gathered between August 2013 and April 2016, and roughly 45,000 articles were gathered during this time span.

After the time-span was decided, the feature extraction on the data set was performed by first determining the features that would be present at each of the stages of the article viewing process. These are the pre-article view (before the article has been opened when only the title and blurb are present), the post-article view (when the reader clicks on and can view the entire article), and the comment view (this disregards the other features and looks solely at the individual comments). These different stages of viewing allow for the user engagement metrics to be broken up into different sections. This is beneficial in the organisation of the experiments and in allowing the data to be processed only a single time for each of the main stages of viewing.

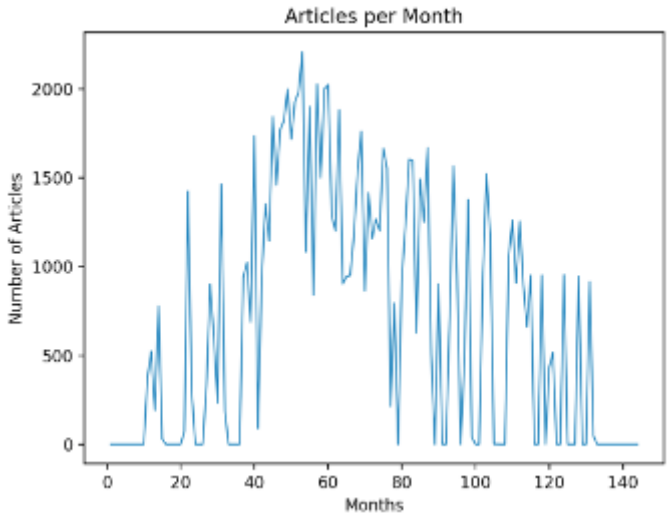

Figure 4.3: Articles per month

4.3.1 Topic Clustering

Topic clustering was performed on all of the articles to determine what the key topic of each article was and how it related to the other articles. This provided a powerful feature as it relates to much of the previous research performed in this area. This feature was extracted from the article text by using both TF-IDF and k-means as described in the design section and using the scikit learn library discussed in the machine learning section below. There were a number of considerations made for both the TF-IDF and k-means models to ensure that they performed optimally.

The TF-IDF vectorizer works by comparing the number of times a word is used in a specific document to the number of times it appears in the corpus of work. This means that it is trained on the article text, therefore considerations need to be made about the cleaning and formatting of the text before it is trained using the TF-IDF vectorizer. The data was pre-processed by removing the stop words from the texts using an in-built function in the TF-IDF vectorizer. Tests were also done using manual lemmatization and removal of stop words, however despite the extra pre-processing the manual lemmatization returned lower silhouette scores when the pre-processed data was used for clustering, which is one of the main performance metrics we are using to analyse the model parameters in this scenario.

After the TF-IDF vectorizer was run on the data, it allowed the analysis to be performed to determine the optimal number of clusters for the k-means models using the silhouette scores and SSE as metrics. As can be seen in Table 4.5 and Figure 4.4 around 9 clusters is the optimal number for this data. The main objective of this is to produce distinct topics that could be used as features. A further analysis was performed on the k-means model into the different clusters to determine the most important words in each so the topic could be discerned, the results of which are provided in Figure ??.

Number of Clusters	Silhouette Score
2	0.0192976
3	0.0189166
4	0.0188185
5	0.0166508
6	0.0192489
7	0.0192976
8	0.0191476
9	0.0193768
10	0.0193075
11	0.0193590
12	0.0188540
13	0.0193187
14	0.0190229
15	0.0189175
16	0.0189286
17	0.0190510
18	0.0192796
19	0.0190015
20	0.0192293

Table 4.5: Silhouette scores for each cluster size

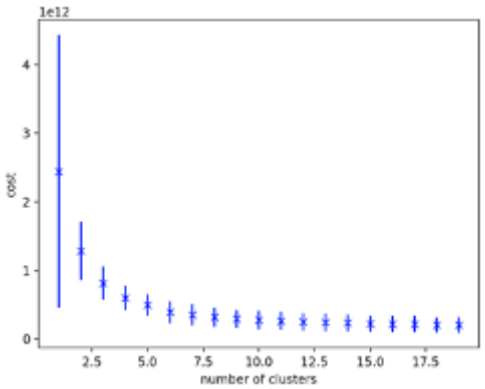

Figure 4.4: Cost for each cluster size

In Table 8.1 we can see the clear differences in topics between most of the clusters. Some of the topics are less clear than others but most contain an easily identifiable theme which is important for the analysis as this information was not available on the journal website. Cluster 0 appears to be related to crime, cluster 1 is an odd case, (as it is the largest cluster and likely is a result of all of the articles that did not fit cleanly into any of the other topics), cluster 2 is related to politics, cluster 3 is related to the internet in Ireland, cluster 4 is related to finance, cluster 5 is related to accidents/crime, cluster 6 is related to lifestyle and social media, cluster 7 is related to sports, and cluster 8 is related to journalism. Apart from niche topics that appear in news, this topic clustering appears to cover all of the main topics usually

discussed in news analysis. It is likely that there are more topics, particularly those in cluster 1 that are made up of smaller sub-topics, but for this experiment all of the main topics seem to be identified.

4.3.2 Natural Language processing Models

All of the other features would require external libraries to extract. The two largest of these sets is the polarity and the entities present in each of the texts. As there are many libraries available for each it was important to choose one that fulfilled the needs of the project as outlined in the design section. The two libraries chosen were NLTK [1] for sentiment analysis and SpaCy [2] for entity detection as they satisfied these needs. Although SpaCy can be used for both, NLTK is designed for research purposes which seemed more fitting due to the nature of the work performed.

NLTK is widely used by researchers as a tool to perform sentiment analysis and similar natural language processing. There are some positives and negatives to using NLTK as there are with all libraries, but for the purposes of the project NLTK appeared to be the best solution. It is a good performer at analysing public opinion, and performing a competitive analysis and is efficient at analysing large data sets such as the one gathered in this thesis. It does lack the ability to deal with spelling and grammatical errors, while also struggling with sarcasm and irony. However, due to the nature of its use in this work it should be suitable.

SpaCy is more widely used in industry, with many applications running SpaCy in their code base. It is being used to perform entity detection due to its high performance and ease of implementation. There is scope to edit the entity detection software and train SpaCy on a different data set but the standard one was used as the entities provided appear to cover all of the main needs in this research. The drawback to the entity detection was the time it took to run which is the reason it was not implemented on the much larger comment data set.

The external libraries worked well and provided no problems in extracting the data needed. The entities extracted by SpaCy were easily able to be stored in arrays for each of the articles which could be passed into the machine learning models. SpaCy was run on the titles, blurbs, and articles' texts to extract the relevant entities so that the differences could be seen in the importance of these areas and the overall effect entities had in these areas. NLTK was used in a similar way to extract the polarity in each of these sections. The only difference was that absolute polarity was used as an extra feature by turning all of the negative numbers to positive numbers. This had the benefit of allowing the magnitude of the polarity values to be identified. This allowed further analysis of the polarity by determining the importance of its values, large values for polarity with low absolute polarities mean they are less effective than

[1] The NLTK library can be found at *https://www.nltk.org/*
[2] The SpaCy library can be found at *https://spacy.io/*

they appear. The opposite is true for small polarity values that are in fact very polarizing with strong positive and negative sentiments that cancel each other out.

The performance of these libraries was not tested on the data set due to the time constraints of the work. Due to the way in which this data is being used and to such a large scale, small errors in either the selection of entities or the polarity of text is unlikely to make a significant difference to the overall trends and results. Therefore, it was also not deemed an important aspect for the overall analysis.

4.4 Machine Learning

There were a number of aspects that need to be considered while performing the machine learning to ensure that the optimal approach was taken and that the results would allow the feature coefficients to be analysed to a sufficient extent. The main areas concerned were in the machine learning library chosen, the data preparation, and the analysis of the models.

There are a number of different libraries that were considered for the machine learning aspect of the project. However, due to the nature of these libraries, many of them have the same models and features as each other. The main concern for the research was that the library would be able to run the logistic regression model that was being used for classification. All of the libraries analysed had this capability and the only difference between most was interface-based. Scikit learn [3] was identified as an easily implementable library that had advantages in areas such as implementing clustering, including k-means, over other models. It also provided a method for extracting the coefficients of the logistic regression models and had the TF-IDF vectorizer required. As the scikit learn library had all of the features necessary to perform the machine learning needed in this thesis, it was the library chosen.

Due to the different magnitudes in the data after it was extracted and cleaned, there needed to be a further step of data preparation performed to ensure the magnitude of the coefficients were of similar sizes relative to their strengths and could be compared. To compensate for this, the data set was normalised. This meant that all of the data was extrapolated between 0 and 1, and -1 and 1 in the case of the polarity (but that is a unique circumstance). This meant that the coefficients would be able to be compared accurately and provide insight into the relationships of the features to the user engagement metrics. An array of the topics was created using the same method as was used for the days of the week. In this, the relevant topic was set to 1 and the rest of the array was set to 0 so that they could be compared in the analysis with the same magnitude as the other features.

Once all of the data had been normalised, it was put into a single array so that it could be used as an input in the machine learning model. Running one large experiment for each of the

[3]The Scikit learn library can be found at *https://scikit-learn.org/stable/*

user engagement metrics as opposed to separate experiments for each of the smaller feature sets allowed for a complete comparison of the features. Although the exact effect of each of the smaller feature sets cannot be analysed this way it provides a better comparison between features in different feature sets for each of the metrics.

Due to the nature of most of the experiments which look to analyse the most popular articles for different user engagement metrics, the top 10% of articles was isolated. Using the top 10% as the positive class and the rest of the data as the negative class would result in the model predicting the most popular class all of the time as this produces the most correct classifications. However, as this data is not useful in determining the most important features, the negative class needed to be separated into nine different smaller data sets so that one could be selected to create an even number of positive and negative classes. This was performed just before the data was used for the model and significantly decreased the size of the data set. Despite this, the new data set still contained just under 10,000 data points due to the size of the initial data set.

4.4.1 Cross-Validation

As there were just under 10,000 data points used in these experiments, 5 fold cross-validation was used for each of the experiments to ensure that the results were statistically significant and to select the optimal parameters for the model. The size of the error in these models would demonstrate the accuracy of the provided metric results. The cross-validation also allowed for the experiments used to decide the optimal C value to be used for producing the optimal model. The C value is the inverse of lambda which represents the amount of regularisation that occurs in the model. This means that choosing a value that produces a good model while not over-fitting or under-fitting the model to the training data is important. As can be seen in Figure 4.5 below, the accuracy and precision values for each of the C values were plotted on a graph and the highest value was selected, $C = 1$ in this case. This was performed on each of the different models to determine the best value, as it varies based on the features and user engagement metric being predicted.

4.5 Result Analysis

The results were extracted from the models by using the coefficients. These represent the magnitude and direction of the correlation for each of the features on predicting the specific user engagement metric. The comparison of the coefficients was performed separately for each of the different user engagement metrics due to the change in the magnitude of these values in the different models. This was in part due to the parameters chosen, such as the C value, but also relied on the relative strengths of the features in the model.

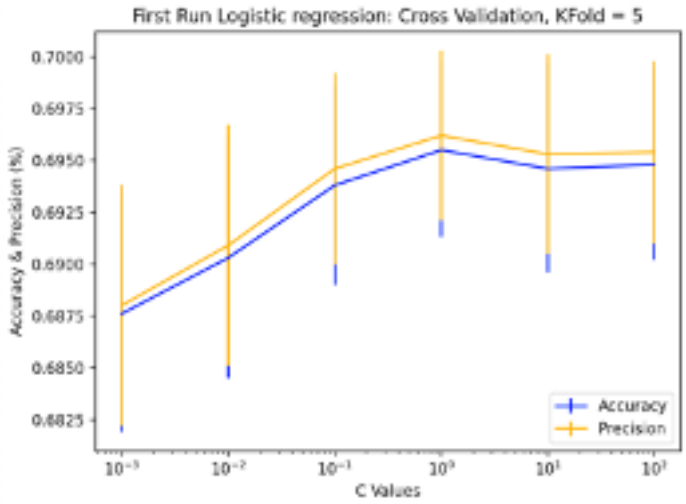

Figure 4.5: Example of cross-validation

Once the initial analysis was completed, a comparison between models could be performed. This is not based on the strength of the coefficients, but based on the relative relationship each of the features had compared to one another and the overall performance of the model in terms of classification metrics. The better performing models obviously have stronger correlations between the features and the user engagement metrics than the weaker ones. This means that considerations need to be taken into account based on the models metrics when determining the overall effect of the features on prediction of the user engagement metric.

As the research objective of this thesis is to determine the relationship of each of the features in predicting the user engagement metrics it is important to consider how these relationships were determined. This means taking into account the performance metrics of the models and the manner in which the features and metrics were gathered and extracted, as all of this plays an effect on the analysis of the results. Due to the methods used in this thesis, these results should accurately portray the relationship between the features in the data set and the user engagement metrics at least to the extent of the model performance metrics. This varies between experiments but will be analysed in greater detail in the analysis and discussion section.

5 Results

This section focuses on the results of the experimentation performed. There were seven main experiments performed on each of the main user engagement metrics. Four of these relate to the article and the features present in it and the final two relate to the comments which are user generated. The first experiment looks into the number of views that an article receives, based on a list of 74 features. This helps determine what makes a user more likely to click on a title and read it based on the information provided to the user at the time (both in the title and in the blurb). The second, third, fourth, and fifth experiments all use the same 95 features and they look at the number of comments, Facebook shares, and email shares that an article receives, as well as the polarity in the comments under a given article. All of these user engagement metrics use the same features as these are present to the user at the same time after the article has been opened (this does not necessarily mean the article has been read). These include features from the title and blurb, as well as introducing more features that are present in the article which are only present to the reader after it has been opened. The final two experiments focus on the comments and the number of likes and replies that can be determined from these. This uses a much smaller feature set with only three features, so the results from this area will be much more restricted than those for the other user engagement metrics.

5.1 Pre-Article View

In the pre-article view, which is before the reader has clicked on a specific article, there are two main features apparent to the reader: the title and the blurb (which is a short portion of text describing the article). These two main features, along with some other key information about the article, were broken down into 74 pre-article view features which were used in predicting the number of views for the article. These features were gathered from a few different areas as described below and were all normalised based on the highest value for each feature:

Entities in title and blurb: There were 18 total entities identified in both the title and the blurb which were used in the prediction of views. They are person, NORP (nationalities, religious and political groups), FAC (buildings, airports etc.), organisations, GPE (countries,

cities etc.), LOC (mountain ranges, water bodies etc.), products, events (event names), works of art (books, song titles), law (legal document titles), languages (named languages), dates, times, percents, money, quantities, ordinals, and cardinals.

Polarity in title and blurb: This represents the level of polarity on a scale of -1 to +1 and is unique for both the title and blurb.

Time-related features: This is related to features such as the day of the week, month of the year, and the year of publication. These are stored as an array with the key sections identified as a 1 and the rest of the array as a 0. So a Monday would be [1,0,0,0,0,0,0].

Lengths of title and blurb: This is a value that represents the number of characters in the text of each.

Absolute polarity: This is a measure similar to that of polarity without taking into account negative values. Any values which appear as negative are changed to a positive value and this is to determine if strong polarity is important as this could be lost through the other method of polarity as many strong positive and negative values could cancel each other out.

Topics: The last feature is for the topics. There were nine topics chosen from analysis of the data set and these features were used in the same way as the time-related features, with an array where the relevant topic was set to 1 while the rest were set to 0.

5.1.1 Predicting Number of Article Views

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be 1 for the model. There is not a large variance for the total accuracies and precisions for different C values but the graph in Figure 5.1 clearly shows that a C value of 1 is optimal for predicting the number of views based on the features.

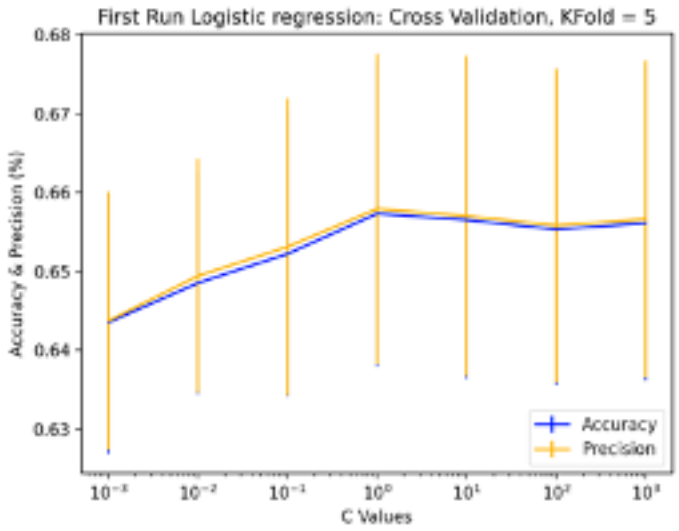

Figure 5.1: Cross-validation for number of views

Features and Coefficients

Top Features	Coefficients	Bottom Features	Coefficients
Topic 3	1.3386232	Blurb Length	-1.5311413
2016	0.63466812	Date in Title	-0.8390401
Time in Title	0.38059506	Topic 4	-0.5557585
Sunday	0.34526781	2013	-0.5388894
Topic 5	0.3382647	Topic 2	-0.4621744
Time in Blurb	0.31058689	Topic 7	-0.3927899
Title Length	0.27598487	Org in Blurb	-0.3525258
NORP in Title	0.25196512	Org in Title	-0.3514351
Topic 6	0.23070019	Date in Blurb	-0.3452428
Saturday	0.2086245	GPE in Blurb	-0.3210024
November	0.16639795	Cardinal in Blurb	-0.310002
2015	0.16430373	Topic 8	-0.2983579
March	0.1290057	NORP in Blurb	-0.2832817
Monday	0.11142547	LOC in Title	-0.2647599
February	0.10563174	2014	-0.2588635
Quantity in Title	0.10168378	Tuesday	-0.2136955
FAC in Title	0.09353574	Topic 1	-0.1711594
Person in Title	0.08685557	Thursday	-0.1638242
Work of Art in Title	0.0801685	Title Polarity	-0.1637612
Person in Blurb	0.07489815	December	-0.1536576
FAC in Blurb	0.06965355	Wednesday	-0.1448895
GPE in Title	0.06307493	Friday	-0.1416895

Product in Blurb	0.06097876	September	-0.1414991
August	0.05218271	LOC in Blurb	-0.1291681
May	0.04701999	Ordinal in Blurb	-0.1180923
July	0.04262345	Blurb Polarity	-0.1176634
Event in Title	0.02086923	Quantity in Blurb	-0.105442
Work of Art in Blurb	0.019221	October	-0.1033425
Money in Blurb	0.01664163	Money in Title	-0.1022958
Language in Title	0.00857869	Percent in Blurb	-0.0828412
Absolute Polarity in Title	0.00643314	April	-0.0824667
January	0.00605501	June	-0.0667317
Product in Title	0.00133229	Absolute Polarity in Blurb	-0.0599904
Language in Blurb	0.00130795	Event in Blurb	-0.0487459
Ordinal in Title	-0.0127175	Cardinal in Title	-0.0473241
Topic 0	-0.026129	Law in Title	-0.0381633
Law in Blurb	-0.0285791	Percent in Title	-0.0316395

Table 5.1: Correlation of features to number of views

We can see from Table 5.1 that there is no specific area of the features that are overwhelming in determining the views an article receives, although as expected the topics are amongst the most important.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.662	0.631	0.664	0.647	0.662

Table 5.2: Model performance metrics for the number of views

The model performed well, considering there are a large number of external factors that were not accounted for. This shows that the features gathered and used in the analysis were good features and have a strong correlation to the number of views.

5.2 Post Article View

In the post article view, which is after the reader has clicked on a specific article, the reader has all of the information from the article including the title, blurb, and any information that is contained in the article itself. These main features, along with some other key information about the article, were broken down into 95 post view features which were used in predicting

the number of comments, Facebook shares, and e-mail shares that an article receives. These features were also used to predict the polarity in the comments. These features were gathered from a few different areas as described below and were all normalised based on the highest value for each feature:

Entities in title, blurb, and article: There were 18 total entities identified in both the title and the blurb which were used in the prediction of post view metrics. They are person, NORP (nationalities, religious and political groups), FAC (buildings, airports etc.), organisations, GPE (countries, cities etc.), LOC (mountain ranges, water bodies etc.), products, events (event names), works of art (books, song titles), law (legal document titles), languages (named languages), dates, times, percents, money, quantities, ordinals and cardinals.

Polarity in title, blurb, and article: This represents the level of polarity on a scale of -1 to +1 and it is unique for both the title and blurb.

Time related features: This is related to features such as the day of the week, month of the year, and the year of publication. These are stored as an array with the key sections identified as a 1 and the rest of the array as a 0. So a Monday would be [1,0,0,0,0,0,0].

Lengths of title, blurb, and article: This is a value that represents the number of characters in the text of each.

Absolute polarity in title, blurb, and article: This is a measure similar to that of polarity without taking into account negative values. Any values which appear as negative are changed to a positive value and this is to determine if strong polarity is important as this could be lost through the other method of polarity as many strong positive and negative values could cancel each other out.

Topics: The last feature is for the topics. There were nine features chosen and this was performed in the same way as the time-related features, with an array where the relevant topic was set to 1 while the rest were set to 0.

5.2.1 Predicting Number of Comments

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be 1 for the model. Looking at the graph in Figure 5.2 we see that there is a local maximum when C is 1. This decreases before increasing again above this point. The reason the value

of 1 is chosen is to reduce over-fitting. The larger the C value the more likely over-fitting is to occur so this value is chosen to reduce the chances of this occurring.

Figure 5.2: Cross-validation for number of comments

Features and Coefficients

Top Features	Coefficients	Bottom Features	Coefficients
Person in Title	1.42077621	Date in Title	-2.2697667
Blurb Length	1.27573755	Cardinal in Title	-1.4898836
NORP in Blurb	1.22576173	Event in Title	-1.2943291
GPE in Title	1.12541519	Cardinal in Blurb	-0.8949406
Title Length	1.09244606	Quantity in Blurb	-0.8692321
Time in Blurb	1.07296627	Money in Title	-0.8680063
NORP in Article	0.96261896	Topic 7	-0.8212369
NORP in Title	0.95560584	Date in Article	-0.8188868
Topic 2	0.88496497	Topic 0	-0.7086338
Org in Article	0.74951973	LOC in Title	-0.7010999
Percent in Title	0.65707246	Date in Blurb	-0.6988392
Law in Blurb	0.65612888	Topic 5	-0.6893412
Article Length	0.55763031	Time in Title	-0.6625248
Person in Article	0.408593	Event in Blur	-0.6491634
May	0.40503393	Topic 8	-0.5690436
Time in Article	0.36609629	LOC in Blurb	-0.5497807
Percent in Article	0.29767009	Ordinal in Blurb	-0.4436399
Org in Title	0.29745583	June	-0.4264863
Money in Article	0.20985539	Org in Blurb	-0.3902289

February	0.20215546	Product in Blurb	-0.3482586
FAC in Blurb	0.17136056	Absolute Blurb Polarity	-0.2960884
Sunday	0.15396957	GPE in Article	-0.2811784
Topic 3	0.14532934	Cardinal in Article	-0.2740874
Topic 1	0.12553108	GPE in Blurb	-0.2537842
Friday	0.11501346	Language in Title	-0.2236921
Person in Blurb	0.09861135	Money in Blurb	-0.2166879
2015	0.09236374	Ordinal in Article	-0.2115965
Quantity in Title	0.08524563	Topic 6	-0.2057335
Law in Article	0.0844537	Article Polarity	-0.1922481
2016	0.07820776	Monday	-0.1735885
Ordinal in Title	0.07337612	Event in Article	-0.1683302
Work of Art in Blurb	0.06850338	December	-0.1678139
August	0.06065726	Topic 4	-0.165835
Percent in Blurb	0.03433623	Absolute Title Polarity	-0.1623421
November	0.03200246	Quantity in Article	-0.1614964
LOC in Article	0.02924345	2014	-0.1606885
September	0.02403289	Blurb Polarity	-0.112843
October	0.01412834	Title Polarity	-0.1025748
April	0.00461641	Work of Art in Article	-0.0979966
Law in Title	0.00408869	Absolute Article Polarity	-0.0973601
Thursday	-0.0108632	March	-0.0929508
Tuesday	-0.0109277	Language in Blurb	-0.0784506
Saturday	-0.0237963	Wednesday	-0.0768938
Product in Title	-0.0249426	Work of Art in Title	-0.0738916
Language in Article	-0.0301376	FAC in Title	-0.0712648
2013	-0.0369695	FAC in Article	-0.0641299
Product in Article	-0.0377506	July	-0.0413193
January	-0.0411429		

Table 5.3: Correlation of features to number of comments

We can see from the top features that most are related to entities in the title and blurb, as opposed to being focused on specific topics or other features. This likely means there are specific entities that people find comment worthy, more than likely these entities are controversial (33). It also seems to imply that the decision to comment is made before the article is viewed due to all of these features being found in the title and blurb.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.6	0.593	0.614	0.603	0.6

Table 5.4: Model performance metrics for the number of comments

The model performed worse than before, but this is expected as the number of comments has a larger number of external factors that were not accounted for. Despite these external features which were not tracked, the model still performed to a level to which good features can be extracted and related to the number of comments.

5.2.2 Predicting the Number of Facebook shares

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be .001 for the model. Again, there is not a large variance for the total accuracies and precisions for different C values, but there is a significant dip after this value, but the graph in Figure 5.3 clearly shows that a C value of .001 is optimal for predicting the number of Facebook shares based on the features. It is important to note that the correlation is very weak so the feature coefficients are likely not as reliable. The error is also quite large for the chosen C value, but even with this taken into account it looks like the optimal value.

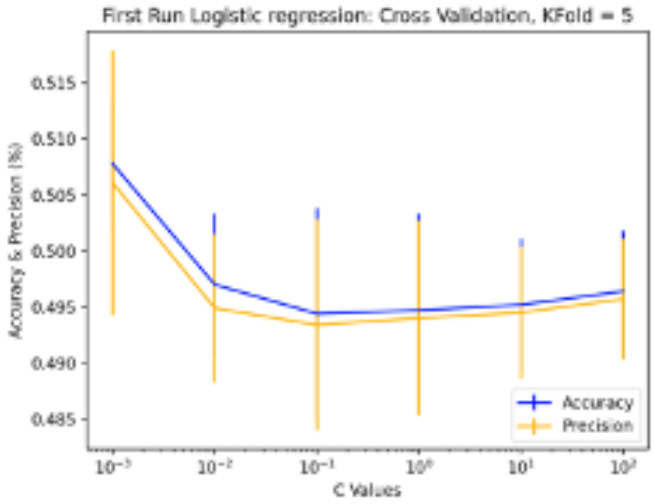

Figure 5.3: Cross-validation for number of Facebook shares

Features and Coefficients

Top Features	Coefficients	Bottom Features	Coefficients
2015	0.0266119	December	-0.0191908
March	0.0170088	2013	-0.0159609
May	0.0156461	Saturday	-0.0154305
Topic 2	0.0110011	April	-0.0145766
February	0.0108522	2014	-0.0121511
Wednesday	0.0092519	Topic 6	-0.0116799
Topic 4	0.0086248	Topic 8	-0.0100730
Tuesday	0.0077885	September	-0.0098887
Blurb Polarity	0.0069410	Friday	-0.0096179
Topic 0	0.0065228	Article Polarity	-0.0069524
July	0.0055010	June	-0.0059881
Topic 1	0.0053180	August	-0.0034782
October	0.0051953	Org in Title	-0.0030144
Monday	0.0043316	January	-0.0029618
Absolute Title Polarity	0.0038322	GPE in Title	-0.0028484
Absolute Article Polarity	0.0038011	Topic 5	-0.0027708
Person in Title	0.0030362	Topic 3	-0.0026455
Thursday	0.0030146	Topic 7	-0.0025602
Cardinal in Blurb	0.0028981	Date in Title	-0.0020453
NORP in Blurb	0.0027674	Quantity in Title	-0.0020089
Absolute Blurb Polarity	0.0026158	Person in Blurb	-0.0018094
November	0.0018804	Article Length	-0.0016972
Time in Blurb	0.0015956	Ordinal in Title	-0.0016052
2016	0.0014997	FAC in Blurb	-0.0014282
Product in Title	0.0013959	FAC in Title	-0.0013752
Time in Title	0.0013762	Person in Article	-0.0011995
Money in Title	0.0013126	Quantity in Blurb	-0.0009371
Money in Blurb	0.0011404	NORP in Title	-0.0006501
Cardinal in Title	0.0010345	Title Polarity	-0.0005884
Title Length	0.0008773	LOC in Title	-0.0005792
Product in Blurb	0.0008144	Org in Article	-0.0004850
Blurb Length	0.0007020	Work of Art in Title	-0.0003687
Sunday	0.0006614	Org in Blurb	-0.0003525
Percent in Title	0.0004124	Work of Art in Blurb	-0.0002505
Date in Blurb	0.0004102	Language in Blurb	-0.0002483

Language in Title	0.0004030	GPE in Blurb	-0.0001997
Event in Blurb	0.0003826	Date in Article	-0.0001850
LOC in Blurb	0.0003383	Percent in Article	-0.0001377
Money in Article	0.0001115	Ordinal in Blurb	-0.0001054
Event in Title	0.0001024	Ordinal in Article	-0.0001003
NORP in Article	0.0000856	Law in Title	-0.0000982
Law in Blurb	0.0000696	GPE in Article	-0.0000826
Percent in Blurb	0.0000530	Cardinal in Article	-0.0000606
LOC in Article	0.0000245	Time in Article	-0.0000503
Language in Article	0.0000008	FAC in Article	-0.0000420
Quantity in Article	-0.0000021	Event in Article	-0.0000276
Law in Article	-0.0000152	Product in Article	-0.0000236
Work of Art in Article	-0.0000183		

Table 5.5: Correlation of features to number of Facebook shares

In this case, all of the features are weakly correlated, most of them being weaker than many of the other cases. This is likely due to the low C value which tends to make the coefficients smaller. When looking at this data, it is also worth considering that the performance is not great for the model, so these features likely only weakly relate to the number of Facebook shares received. In this case, the entities seem to only play a small role and the topics and time relative features seem much more important.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.538	0.514	0.544	0.528	0.539

Table 5.6: Model performance metrics for the number of Facebook shares

The model performed poorly in this scenario, indicating that the selected features likely only had a small effect on the overall Facebook shares that an article receives. It is likely that the fact that this feature is not used as often as others on the-journal.ie plays a role in the model performing worse in this scenario.

5.2.3 Predicting the Number of E-mail shares

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be .1 for the model. The variance for the total accuracies and precisions for different C values is

just over .01 so does not have a large effect, but the graph in Figure 5.4 clearly shows that a C value of .1 is optimal for predicting the number of e-mail shares based on the features. The correlation is stronger than that of the Facebook shares but is still not a strong prediction. There will still be means to determine some important features from this model.

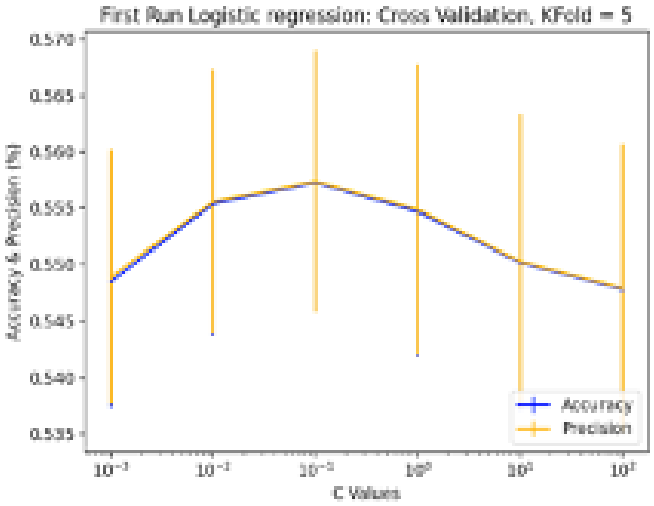

Figure 5.4: Cross-validation for number of e-mail shares

Features and Coefficients

Top Features	Coefficients	Bottom Features	Coefficients
April	0.3606331	November	-0.5097429
May	0.3249388	December	-0.2638396
October	0.3067781	September	-0.2636803
July	0.3059964	2015	-0.2084763
Money in Blurb	0.2292558	June	-0.2003193
Title Length	0.2043928	Money in Title	-0.1713139
2016	0.1881303	2014	-0.1628982
2013	0.1817951	Ordinal in Title	-0.1564731
Absolute Blurb Polarity	0.1725353	Saturday	-0.1445531
Cardinal in Blurb	0.1722915	Product in Title	-0.1355864
Absolute Article Polarity	0.1709186	Topic 0	-0.1300642
Date in Blurb	0.1454451	Topic 2	-0.1299745
Sunday	0.1382569	February	-0.1242729
Person in Blurb	0.1239201	Time in Blurb	-0.1173208
Cardinal in Title	0.1006320	Ordinal in Blurb	-0.0945235
NORP in Title	0.0991563	Product in Blurb	-0.0914238
March	0.0880196	NORP in Blurb	-0.0911764
Org in Title	0.0873038	Topic 6	-0.0910671

Quantity in Blurb	0.0838267	Topic 4	-0.0848094
Monday	0.0738565	Law in Title	-0.0837040
Percent	0.0712165	Person in Title	-0.0811569
Org in Blurb	0.0664327	Date in Title	-0.0802904
Topic 1	0.0471393	Topic 3	-0.0712255
FAC in Blurb	0.0465843	Topic 7	-0.0562520
LOC in Title	0.0460976	Time in Title	-0.0531467
Topic 5	0.0366755	Cardinal in Article	-0.0460591
Language in Blurb	0.0312670	Wednesday	-0.0457753
Blurb Length	0.0262431	Event in Title	-0.0444819
Tuesday	0.0237885	August	-0.0428045
Absolute Title Polarity	0.0235398	GPE in Article	-0.0377787
Language in Title	0.0230211	GPE in Blurb	-0.0376654
Topic 8	0.0207173	Percent in Blurb	-0.0367527
FAC in Title	0.0189290	Event in Blurb	-0.0309348
Title Polarity	0.0170461	Person in Article	-0.0298607
January	0.0168443	Thursday	-0.0298129
Date in Article	0.0081285	GPE in Title	-0.0268771
Blurb Polarity	0.0081256	LOC in Blurb	-0.0243005
Work of Art in Blurb	0.0078480	Article Polarity	-0.0238633
Article Length	0.0051866	Percent in Article	-0.0183228
FAC in Article	0.0050375	Friday	-0.0172098
Quantity in Article	0.0049897	Quantity in Title	-0.0157907
Work of Art in Article	0.0040718	Law in Blurb	-0.0101384
LOC in Article	0.0036054	Work of Art in Title	-0.0081824
Language in Article	0.0029951	Ordinal in Article	-0.0057324
Time in Article	0.0020278	Event in Article	-0.0018577
Product in Article	0.0016556	Law in Article	-0.0008852
Money in Article	0.0001065	NORP in Article	-0.0001919
Org in Article	-0.0001384		

Table 5.7: Correlation of features to number of e-mail shares

As t top and bottom features seem to relate to the monthly features, this seems to suggest that the time of the year has a large effect on this. The year and some polarity features also appear particularly strongly correlated to this form of user engagement.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.562	0.58	0.581	0.581	0.562

Table 5.8: Model performance metrics for the number of e-mail shares

The model was weaker than many of the other models. It is apparent that there is a correlation in some specific areas, but like Facebook shares, this is not an overly utilised feature and so it makes it hard to produce a good model based on this engagement type on this data set.

5.2.4 Predicting Polarity in Comments

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be 1 for the model. There is not a large variance for the total accuracies and precisions for different C values, with each value differing by no more than .01, but the graph in Figure 5.4 clearly shows that a C value of 1 is optimal for predicting the polarity based on the features.

Figure 5.5: Cross-validation for comment polarity

Features and Coefficients

Top Features	Coefficients	Bottom Features	Coefficients
Person in Article	1.99577349	NORP in Article	-0.9087508
Topic 7	1.78043172	Blurb Length	-0.7255311
Topic 6	1.04791251	Org in Article	-0.6841485
Article Polarity	1.03984305	GPE in Article	-0.6046785

Topic 4	0.96124652	Title Length	-0.5020621
LOC in Blurb	0.91675047	FAC in Title	-0.3790794
FAC in Blurb	0.89991504	GPE in Title	-0.3683371
Work of Art in Title	0.82442945	Time in Title	-0.32268
Topic 3	0.79576595	NORP in Title	-0.3195793
Title Polarity	0.77539599	Time in Blurb	-0.278807
Event in Title	0.72927318	Date in Blurb	-0.2582661
LOC in Title	0.71640208	NORP in Blurb	-0.2160134
Work of Art in Blurb	0.63167335	Article Length	-0.1902045
Ordinal in Article	0.62362489	Money in Blurb	-0.1852627
Money in Title	0.62241759	Law in Title	-0.1287948
Topic 2	0.61745066	July	-0.113
Cardinal in Article	0.56278714	Absolute Blurb Polarity	-0.1093024
Language in Title	0.54343889	November	-0.1069962
Person in Blurb	0.53461482	2015	-0.1017909
Quantity in Title	0.49617941	Wednesday	-0.0682484
Person in Title	0.49347506	Absolute Title Polarity	-0.0618357
Ordinal in Title	0.48362216	Tuesday	-0.0523488
Topic 8	0.47298867	October	-0.0474832
Topic 1	0.36559017	Law in Article	-0.0459847
Event in Blurb	0.35147314	2016	-0.0369743
Blurb Polarity	0.34955859	September	-0.0357098
Ordinal in Blurb	0.33601999	Quantity in Blurb	-0.0217776
GPE in Blurb	0.32805347	2014	-0.0172636
Date in Title	0.30783231	February	-0.0114776
Product in Blurb	0.27675422	August	-0.00657
Org in Title	0.26520316	May	0.00180
Quantity in Article	0.23824699	Friday	0.0190421
Percent in Article	0.22711742	Thursday	0.02122183
Org in Blurb	0.21949009	Saturday	0.02126941
Language in Blurb	0.18958606	Product in Article	0.02206167
Percent in Blurb	0.18898966	April	0.0281
FAC in Article	0.18195016	March	0.0309
Event in Article	0.18116727	Monday	0.03518957
Percent in Title	0.17797977	Sunday	0.03853361
Work of Art in Article	0.17704756	Product in Title	0.04188374
2013	0.17068825	June	0.0540

Cardinal in Blurb	0.16920883	Language in Article	0.05667979
Topic 0	0.16232542	December	0.06257921
January	0.1587247	Law in Blurb	0.08948388
Time in Article	0.15840874	Absolute Article Polarity	0.09776125
Cardinal in Title	0.13968765	Topic 5	0.10427447
Date in Article	0.12637539	Money in Article	0.10831741
LOC in Article	0.11164936		

Table 5.9: Correlation of features to number of polarity in comments

The top features in predicting a positive polarity tend to be topic related with three of the top five features being related to them. It seems to be more entity-dependant on the negative side, meaning that it is likely that people don't seem to react negatively to certain topics but more so entities in those topics.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.694	0.705	0.694	0.7	0.694

Table 5.10: Model performance metrics for the polarity in the comments

The model performed well, considering there are other factors that likely play a part in determining a readers polarity such as the current polarity of the comments. Despite this, it is the best performing model in the post view section.

5.3 Comment View

The comment view is a much smaller area of research in this thesis than the others and requires more research than is contained in this experiment in order to achieve a full understanding. This research will give an idea of how the features of length and polarity contribute to the replies and likes that a comment receives. This is ignoring all of the article data and looking specifically at the comment text. There is scope to increase this by looking at the entities in each of the comments. However, as the look at these comments has a smaller effect on the overall user engagement relative to the other user engagement metrics, a smaller analysis was performed.

Length of comment: This is a value representing the number of characters in the comment text.

Polarity in comment: This represents the level of polarity on a scale of -1 to +1 for the comment.

Absolute polarity in comment: This is a measure similar to that of polarity without taking into account negative values. Any values which appear as negative are changed to a positive value and this is to determine if strong polarity is important as this could be lost through the other method of polarity as many strong positive and negative values could cancel each other out.

5.3.1 Predicting Comment Likes

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be .001 for the model. There is not a large variance for the total accuracies and precisions for different C values, with each value differing by no more than .005, but the graph in Figure 5.6 clearly shows that a C value of .001 is optimal due to the decrease after this for all other values.

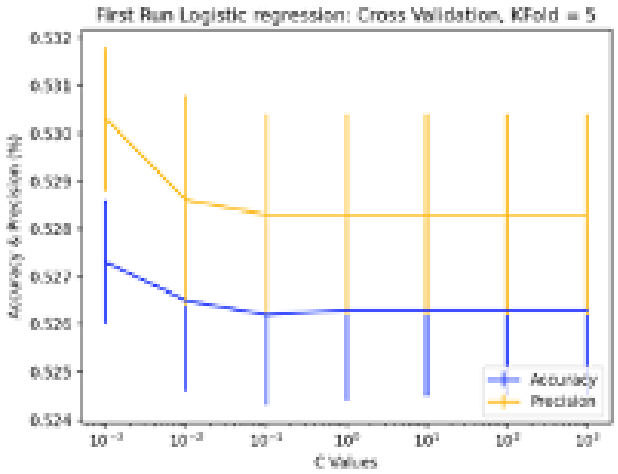

Figure 5.6: Cross-validation for comment likes

Features and Coefficients

Top Features	Coefficients
Absolute Polarity	0.2154323
Polarity	0.00555073
Length	-0.00093176

Table 5.11: Correlation of features to number of likes on comments

We can see that the absolute polarity is the most correlated feature, this is in reference to comments with strong polarity leading to more likes. We can see by the drastic change between this and polarity, meaning that it does not matter if this polarity is positive or negative - just that there is polarity. The length, although being slightly negatively correlated, seems to have no large effect.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.53	0.709	0.525	0.603	0.534

Table 5.12: Model performance metrics for the number of likes on comments

The model performed well in some metrics and poorly in others. The accuracy is only .53, meaning that there is not a strong correlation overall; but still it is quite precise with a value of over .7. The model is not strong but this is largely due to the limited number of features relative to those that influence this metric.

5.3.2 Predicting Comment Replies

Cross-Validation

Using 5 fold cross-validation as described in 3.4.3, the optimal C value was determined to be .1 for the model. This result in Figure 5.7 very different to the others as there seems to be minimal change between most of the values. The only difference is the increase seen in precision up to a C value of .1.

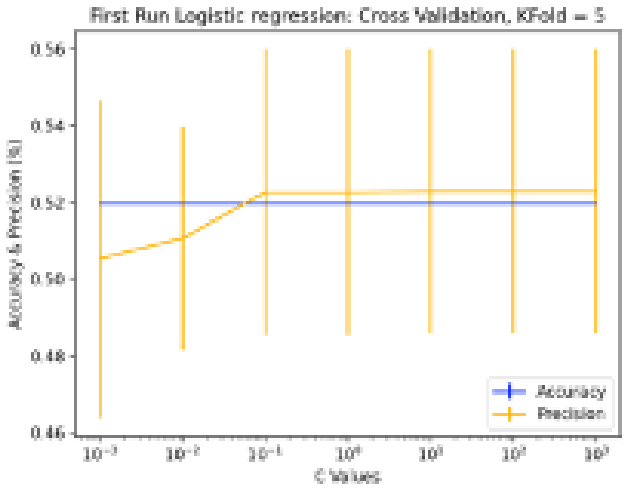

Figure 5.7: Cross-validation for comment replies

Features and Coefficients

Top Features	Coefficients
Polarity	0.05955
Length	0.000006
Absolute Polarity	-0.10447

Table 5.13: Correlation of features to number of replies on comments

We can see that there is little to no correlation between these features and the number of replies that a comment receives. All correlations are either around or below .1.

Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.52	0.511	0.439	0.346	0.48

Table 5.14: Model performance metrics for the number of replies on comments

The model performed poorly for all metrics and had some of the worst metrics out of all of the experiments. This means that the features used had very little to no correlation with the number of replies that a comment receives.

6 Analysis and Discussion

In this section the findings and the results of the experiments will be discussed. This involves determining the most influential features in predicting user engagement, both positively and negatively, and determining the extent to which these are important. The effects of the models metrics will also be discussed in relation to the results. An initial discussion is provided for each of the user engagement metrics, outlying its use and importance in the area of research. This is followed by a brief analysis of the models performance and then a breakdown of each feature sets effects on the specific user engagement metric. The different feature sets, such as the topics and entities in each of the sections, will be analysed separately so the most important features in each of these sets is determined and analysed, providing a structured and beneficial analysis for each user engagement metric.

6.1 Number of Views

Due to the importance of views and clicks to many online news organizations, particularly those that get much of their revenue through ads, this is possibly the most important metric to them. The number of views determines how many individual readers the news organization was able to convince to look at their article, and this depends on a large number of different features. This experiment gathered many of those related to the text and content, as well as some time-based features to produce the model. The model in this scenario will be limited by features outside of those mentioned, such as layout, other news websites, and even stochastic reader behaviour. As such the optimal results for the model are obviously quite a bit lower than 1 for all of the metrics. So although these metrics are important to an extent, considering the model is performing to an adequate level, these will not have a major impact on the analysis.

The features used in this experiment were chosen based on their availability to the reader before this point. They are all features that are either time based and outside the control of the news organization, or related to the title and blurb, as these are the only two items that the user can see before selecting an article (apart from a small picture which is not considered in this analysis but is another external factor). These features are broken down into further

sections: entities in title and blurb, polarity in title and blurb, time related features, lengths in title and blurb, absolute polarity in title and blurb, and topic covered. Each of these was gathered directly from the title and blurb apart from the topic which was recovered from the article text using feature analysis and clustering as in described in section 3.3.

The objective was to determine which features led to the most popular news stories which is why the top 10% of views was the barrier that was determined instead of just over 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets, but are representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in Table 5.1 was narrowly affected by less than .02 for both accuracy and precision over the range. The value of 1 was chosen based on the graph and this is small enough that it is not likely to result in over-fitting of the data, even though the data set used was quite large.

6.1.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.662	0.631	0.664	0.647	0.662

Table 6.1: Model performance metrics for the number of views

The model ended up performing well despite the external features not analysed in these experiments that likely contribute to the prediction of the results and due to the fact that the model was not optimised. Optimal features were not chosen and inter-model analysis was not performed which contributes to this model not performing as well as it possibly could have. However, despite this, the results garnered can still be useful as it is their magnitude and relationship to one another that is important. With the results seen in Table 5.1 it is clear that there is still a strong correlation between the chosen features and the number of views, with all metrics being statistically significant.

6.1.2 Topics

The topics were quite influential in determining the number of views that an article received. This makes sense, as they are what most of the users are interested in and many readers tend to be interested in the same topics, especially if they are reading the same newspapers. If all

of the other features were removed from the model, the topics would still likely be able to predict the overall number of views at least to some extent.

Topic 3 is the most positively influential feature with a correlation of 1.339. This is over twice as influential as the next most important positively influential feature. Topic 3 is concerned with the internet in Ireland and is likely so popular due to the importance of internet related stories to those in Ireland for those that are using the internet in Ireland, as this is the key demographic of the-journal.ie. It is likely that this feature would not be as popular if a similar experiment was performed on a data set of hard copy newspapers, but in this scenario all readers share some common interests, most basic of which is they are on the internet and likely in Ireland as it is an Irish website.

After this we have Topic 5 (0.338), and Topic 6 (0.231) which are accidents/crime, and lifestyle and social media respectively. These are less popular but are still strongly positively correlated to the number of views. Topic 5 seems to focus more on road accidents and related crimes, which seems to be more interesting to readers compared to Topic 0 (which also looks into crime but focuses more on attacks). Topic 0 has almost no correlation and the difference seems to suggest that stories containing road related crimes and accidents are much more interesting than those discussing other types of crimes. Topic 6 is not as clear a topic but it seems to be related to the lifestyle and social media sphere. This being popular would seem to link to celebrity news which is considered an interesting topic by many and even has tabloids devoted to it, which explains its popularity here.

Most of the negative topics tended to be quite specific and focus on overall less popular topics. These include Topic 4 (Finance, -0.556) which was the least popular topic, likely because it is a niche area with a small number of very interested readers and an apathetic general population. This is similar to the next two most negatively correlated topics, which are politics (-0.462) and sport (-0.393). However, both of these are interesting because, in general, people have opinions about both of these either about the current government or about their favourite team, the difference is that these are both quite broad subjects. In sport, for example, if you support a football team and watch every game it does not mean you want to know about rugby, or even to an extent other football teams and so most articles in this area won't be interesting. Politics lands somewhere in between, more people are interested in it than finance but it is also fractured with a broad number of topics, so individuals might not be interested in one specific area but not want to read about any of the others.

Topic 8 (Journalism, -0.298) and Topic 1 (Other, -0.171) were less negatively correlated than the other topics but were still negative. For Journalism this is likely due to many of these topics being related to the-journal.ie and contributions to their efforts in providing news. As this is looking for donations and usually does not provide a large amount of information on current affairs, it is often over-looked. Topic 1 is interesting as it was the largest topic in

the data set and seemed to comprise all of the articles that did not fit into any of the other topics. It seems to contain a lot of words that would be used generally in many topics such as Ireland, Irish, today, children, new etc., but it is hard to derive reasons for this from this limited view of the topics concerned.

6.1.3 Entities in the Title

For this model, we start with the most important features and then discuss any other interesting features that arise. The most important entity in the title is a date. This is negatively correlated with a coefficient of -0.839. This implies that having a date in the title dissuades readers from viewing the article. This could be due to a number of reasons but is likely due to the topics that are also bottom features such as sports which would tend to have dates in them. It is also possible that titles with dates in them tend to give a lot of information away and therefore don't tend to be clicked on or that dates in the past are considered no longer relevant to readers after a certain period of time.

Organisations being named in the title also tend to be negatively correlated with the number of views received with a coefficient of -0.351. This is likely due to their correlation to the topics in this area. Topics such as sports, politics, and finance all contain a large proportion of organisations due to the topics and stories discussed. As these are the bottom features it is easy to see why organisations in the title would also produce negative correlation.

LOC (mountain ranges, water bodies etc.) in the title, -0.265, are likely uninteresting topics to the general population. They also likely do not appear often and when they do they appeal to a niche section of the reader base that is interested in outdoors and adventure topics.

Time and NORP (nationalities, religious and political groups) are both positively correlated with coefficients of 0.381 and 0.252 respectively. Time is an interesting feature but likely has correlation to the-journal.ie news format which produces headlines such as the 5 at 5 which are recaps of important stories into shorter more digestible chunks and give readers a chance to get information quickly. It is also correlated to the topic of crime/ accidents as time is often important in this topic to demonstrate when the event occurred. NORP is interesting as there seems to be no link as obvious as between organisations and their topics as the most positively correlated topics are the internet in Ireland and accidents/ crime. Interestingly, entities that include political groups are popular even though politics as a topic is not. This likely means that the other aspects of this likely result in most of the positive correlation meaning the nationalities and religious groups are more important than this feature would suggest.

6.1.4 Entities in the Blurb

The entities in the blurb are for the most part similar to those in the title with some notable exceptions. For the blurb, the most important entity is an organisation. It is still negative with a coefficient of -0.353. This is likely due to the reasons discussed above in its link to unpopular topics.

Other notable negative entities in the blurb are the date, GPE (countries, cities etc), and cardinal numbers. Dates and cardinal numbers likely carry similar reasoning in that these tend to give away a lot of factual information in a short amount of time, after seeing this it is possible that many readers think they have learned all they need about the article and don't proceed to read it (for example think of a headline that reads "4 dead in hospital fire last Thursday in Dublin" compared to "Many deaths in hospital fire"). Although both convey the same story, one requires further reading due to the lack of clarity in the title, and the other is much more likely to get skipped over. It is also important to note that this is becoming more prominent in online news stories with titles meant for "click bait", this can lead to readers getting frustrated at the news source. The example above also showed a possible link to GPE in its use of further describing a situation, but it is more likely that these entities are often in reference to sports teams and as stated before sports is also a negatively correlated topic, meaning GPE would be too.

The only entity that is significantly positively correlated is time. This is likely due to the same reason as with both the recap of stories including times, as well as its link to crime/ accidents.

The most interesting entity in the blurb is the NORP as unlike in the title where it was positively correlated it is now negatively correlated. The most likely reason for this is due to how it is used in both cases. As mentioned before, due to politics being negatively correlated NORP in the title, likely did not contain many political party names and when it was used it was likely in reference to nationalities and religious groups, in the blurb it is more likely that the political party names will appear as it is a more in-depth look. Whereas the title would usually contain the names of specific ministers, the blurb is where their party and other information would be contained. This means that its prevalence in the blurb is likely due to its link to politics as opposed to the other two features. This correlation cannot be determined for definite due to the manner in which the data was gathered but explains the reason for this difference and provides scope for further research in this area.

6.1.5 Polarity in the Title

Title polarity is negatively correlated, -0.164, meaning that stories with negative titles are more likely to get views. This is in line with much of the research around the fascination with

morbid events by Zuckerman and Litle (38).

6.1.6 Polarity in the Blurb

Blurb polarity is also negatively correlated, -0.118 , with a weaker coefficient than the title. This means that sentiment here is less important, specifically if it is negative. This is likely for the same reason surrounding the work of Zuckerman and Litle (38).

6.1.7 Absolute Polarity in the Title

The absolute polarity in the title is only 0.006 meaning that the correlation almost does not exist. This must mean that although the negative polarity has a strong correlation the positive side does not. This is interesting, as it means readers are more likely to read negatively polarised articles, at least in terms of titles, than positive ones and that the relationship in the positive sense has no correlation to a reader's decision to view an article.

6.1.8 Absolute Polarity in the Blurb

This is similar to the title with a correlation of -0.06. This is quite a bit stronger than the title but is overall only a small factor in relation to views. This means that the polarity in both likely does not contribute to a large extent to a reader viewing an article with even a slight dissuasion of readers from sentiment content in the blurb.

6.1.9 Time Related Features

This section was broken down further into three more sections including years, month of the year, and days of the week. The most important of which in this case appears to be the years. This makes sense in terms of the journals popularity as any trends in this data would likely signify how the number of overall readers changed over this time. The second most positively correlated feature is the article being written in 2016 while the fourth most negatively correlated was the article being written in 2013, with both other years falling in between, with 2015 positive and 2014 negative. This makes sense as the journal has become increasingly popular over this time as has online media as a whole. The values were 2016 (0.635), 2015 (0.164), 2014 (-0.259), and 2013 (-0.539). This seems to suggest that the largest increase in viewers came between 2015 and 2016.

After this, the next most important time-based feature was days of the week. This is for the most part as expected, with Sunday being the most popular day of the week to consume online content with a correlation of 0.345. This is followed by Saturday and Monday which are also both positively correlated. Interestingly Monday is more popular than Friday by a large extent, with Monday being positive, 0.111, and Friday being negative, -0.142. Tuesday is the

least popular day with a correlation of -0.214, followed by Thursday then Wednesday. This seems to coincide with how the days of the week are traditionally portrayed with a special edition posted on Sundays due to readers having more time over the weekend. During the middle of the week, when people are traditionally at work and have less time the views dip which is also expected.

The months of the year are not as correlated as the other time based features. The main trend seems to be that months later in the year tend to be more negatively correlated and those at the start of the year tend to be more positively correlated. There is a major exception to this and it is November that is the most positively correlated month (0.166). This is followed by March (0.129) and February (0.105). The most negatively correlated is December (-0.154) followed by September (-0.141) and then October (-0.103). The reason for this trend is not immediately clear and is possibly due to these being traditionally busy times of the year as they are near major holidays, such as Christmas, or other important times in the year such as back to school for kids and the summer where more time is spent outside. This seems to explain most of the trends but may not be the only reason.

6.1.10 Title Length

The title length is positively correlated with a correlation coefficient of 0.276. This is quite positive and seems to show that longer titles lead to more views. This could be due to the fact that the longer titles have more room to grab the readers' interest and are also more likely to stand out on a page when a reader is scrolling. We already determined that titles containing figure-based data, cardinal numbers, and dates, were less read and it is possible that these types of concrete data lead to shorter titles and so longer titles are better at enticing the user to view the article as they contain less of these features. It is also possible that it is related to the topics that perform well such as the internet in Ireland and accidents/ crime.

6.1.11 Blurb Length

Unlike the title length, this is negatively correlated to quite a large extent, -1.53, and is the most influential feature in determining article views. The reasons for this seem more clear than the title length, as it is likely that there are some articles with very long blurbs that provide a short synopsis of the article. This may seem beneficial for the reader to let them know a complete overview of the article before they read it, but many readers may see this and feel they know what the story is about, without feeling the need to view the article. It is also possible longer blurbs are attributed to specific topics or authors that tend to garner more views.

6.1.12 Views Discussion

This experiment gives an overview of the most influential features in the article for determining views. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence. This does not mean all of the reasoning will definitely be accurate in all situations, but provides a platform for further research to be continued in the area. From the analysis, the topics appear to be the overall most important features with most being significant to a large extent. The lengths also play an important role and so too does the year the article was published. After this the results are mixed with entities like the time and date in the title being important and an ordinal in the title being unimportant. These unimportant features are interesting as they provide insight into what features are not correlated to views, but they do not provide as much information as the influential features for assisting in future work.

6.2 Number of Views

The number of comments that an article receives is a good metric for determining user engagement as it is one of the most commonly utilized forms of engagement outside of viewing an article. It allows users to provide feedback for the article and allows news organisations to interact with their users. It also helps retain users as those that interact and go through the process of setting up an account are more likely to have a preference for that particular news organisation. This experiment gathered all of the features that were available to the reader related to the text and content, as well as some time based features to produce the model. The model will be limited by features outside of those mentioned, such as pictures, layout, other news websites, and even stochastic reader behaviour. As such the optimal results for the model are obviously quite a bit lower than 1 for all of the metrics. So although these metrics are important to an extent, considering the model is performing to an adequate level these will not have a major impact on the analysis.

The features used in this experiment were chosen based on their availability to the reader at this point (after the reader has opened the article). They are all features that are either time-based and outside of the control of the news organization, or are related to the title, blurb, or article. This is because these are the main components that go into making up the overall presented article. There are likely many features that haven't been analysed here that also have an effect on the outcome such as the number of comments currently on the article and the discussion that is occurring, as well as countless other features including the layout and any related media in the article. The features chosen were broken down into further sections: entities in title, blurb, and article, polarity in title, blurb, and article, time related features, lengths in title, blurb, and article, absolute polarity in title, blurb, and article, and topic covered. Each of these was gathered directly from the title, blurb, and article either

using text analysis or by scraping displayed features.

The objective was to determine which features led to the most popular news stories which is why the top 10% of the number of comments was the barrier that was determined instead of just over 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets but is representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in the GRAPH was larger than most of the other experiments changing by up to .07 for both accuracy and precision over the range. The value of 1 was chosen based on the graph as this was a local maximum. Although the graph trends upwards later after this point, the model is likely to result in over-fitting which is why the smaller C value of 1 was selected.

6.2.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.6	0.593	0.614	0.603	0.6

Table 6.2: Model performance metrics for the number of comments

The model performed worse than the model predicting the number of views and this is likely due to the fact that there are more external and non-measured features that go into determining the number of comments than the number of views. Like the model for the views, this model was not optimised as the optimal features were not selected and inter-model analysis was not performed to select the best model for this experiment, meaning that the results are worse than they could have been if these measures were accounted for. However, despite this, the results garnered can still be useful as it is their magnitude and relationship to one another that is important. With the results seen above in Table 6.2 it is clear that there is still a strong correlation between the chosen features and the number of comments, despite many features not being accounted for.

6.2.2 Topics

The topics are influential in determining the number of comments, with the most influential topic being positively correlated and the next six being negatively correlated. This seems to suggest that there are a large number of topics that don't seem to have many comments and

relatively few that spark a lot of comments; and that other features in the text appear to have a strong effect on the number of comments received.

The most influential topic is Topic 2 (Politics, 0.885) which is an inherently controversial topic due to the fact that by its nature it divides the population. The controversial nature of politics seems to confirm much of the research in this area, which indicates that controversial topics tend to receive more comments (33).

The other two positively correlated topics were Topic 3 (The internet in Ireland, 0.145) and Topic 1 (Other, 0.126). The internet in Ireland is likely here due to its popularity being the most highly viewed topic, therefore it is likely to receive more comments due to the number of readers, as there seems to be no other inherent reason for this topic to receive so many comments. Topic 1 is harder to interpret due to the topic being quite vague. It has some similar themes to Topic 3 in terms of Ireland being a prominent feature but it is mainly made up of the left over articles that did not fit into the other topics. It is hard to interpret any correlations to it for this reason.

Topic 7 (Sports, -0.821) is the least commented topic followed by Topic 0 (Crime, -0.701), Topic 5 (Accidents/ Crime, -0.689), and Topic 8 (Journalism, -0.569). Apart from sports it is clear why the other three topics don't elicit many comments due to the nature of the topics. Two of the topics are based around crime, which is an uncontroversial topic as most people are against crime. These topics are also quite data focused explaining what happened and where, providing the reader with a lot of factual information. This also tends to make it quite uncontroversial and, in a similar way to controversial topics receiving comments, uncontroversial topics don't. This is similar for journalism, as this is also an uncontroversial topic and as we saw in relation to the number of views it is also not very popular, both of these together result in a topic with a negative correlation to a number of comments received. Sports is probably the least commented for a similar reason, as although it can be quite a controversial subject, those who are not interested in a specific sport or club won't tend to view an article on them and by extension won't comment on them.

Topic 6 (Lifestyle and Social Media, -0.206) and Topic 4 (Finance, -0.166) are also both negatively correlated. Finance likely follows on from being one of the less popular article topics to view. The relation and the fact it is less negatively correlated likely means that a higher proportion of the readers visiting are commenting but this still accounts for negative correlation overall. In much the opposite way Topic 6 was popular in views with a positive correlation and is negatively correlated now likely due to it not being a controversial topic most of the time.

6.2.3 Entities in the Title

The entities mentioned in the title are the most influential features in predicting the number of comments an article receives. This seems to suggest two things. The first is that readers seem to be influenced by the title, and to a lesser extent the blurb, more than the actual article in terms of commenting. This would mean that many readers are not reading the article before commenting and are doing so based purely off of the information in the title and blurb. This also means that readers often don't have the full story when commenting and are doing so based off a reactionary response to the title and blurb. The second is that entities rather than topics are more controversial. Most research in this area seems to have focused on how different topics affect the number of comments it receives, but looking at the data gathered it seems that entities play a much larger role.

The two most important features negatively influence the number of comments an article receives are entities in the title. These are a date in the title (-2.27) and a cardinal in the title (-1.499). The reason for this seems to follow along from previous research in this area in that both of these features tend to be uncontroversial as they usually appear around factual information, and factual information by its nature is uncontroversial.

The entities in the titles with the most positive coefficients are a person in the title (1.421), GPE (countries, cities etc.) in the title (1.125), NORP (nationalities, religious and political groups) in the title (0.956), and percent in the title (0.657). From the first three of these we can see an interesting trend, all of these entities are proper nouns, likely meaning readers tend to have strong opinions on specific people, places, and groups. It seems to follow that these are the most controversial elements for readers and they provoke the reader to comment. Percent is an interesting entity as it would appear that it would be similar to date and cardinal in providing factual information to the reader which are generally uncontroversial topics. However, in this case, it likely derives from opinion polls run that have their results in terms of percentages of the population, and controversy could arrive from debate between the two sides over the result of the poll. This is just speculation, due to the lack of a thorough analysis into the specific articles that contain percentages in the title but is an interesting area for future work and in particular how it relates to other figure-based information.

The other interesting features are all negatively correlated and are event in the title (-1.294), money in the title (-0.868), LOC (mountain ranges, water bodies etc.) in the title (-0.701), and time in the title (-0.662). Events likely don't receive many comments for the same reason as sport as they would predominantly be the events mentioned, even those that are controversial likely don't counteract the number of sporting events that are written about due to their regularity. Time likely appears here because of the recap on the journal as mentioned before they have a "5 at 5" which is 5 important stories at 5 o'clock most weekdays, these stories are likely to get viewed but not commented on as they contain too much information

about different stories and readers will go to the individual story if they want to engage in conversation. Money is less clear but appears to be related to topics that are uncontroversial such as crime and finance. LOC is a topic that has very little interest, as indicated by its negative correlation to views. This is compounded by the fact that it is generally not considered controversial which is why it has a negative correlation to comments.

6.2.4 Entities in the Blurb

The entities in the blurb are less important than those in the title, but still more important than many in the article. This seems to mean that often people are commenting after very little reading with some commenting after just reading the title, then more commenting after reading the blurb, and the rest after reading the article. This could also be due to the entities in the title and blurb accurately reflecting the content of the article, but likely has some relation to readers commenting before reading the article.

NORP in the blurb (1.226) is the strongest of these features and likely has much the same reason as in the title by representing groups that would be controversial. This is clear by its link to political groups which is the most controversial topic determined in this work.

Time and law are the next most important positively correlated features with coefficients of 1.073 and 0.656 respectively. Time is interesting, as when it is in the title, it receives much fewer comments, so there must be an inherent difference between the use of time in the title and the use of time in the blurb. There does not seem to be a clear difference in the disparity, maybe it is due to the fact that, as described above, the "5 at 5" receives few comments but the reason that time is a prominent feature in the positively influenced features is harder to decipher. It does not seem to have a link to politics in any meaningful way which is the most positively correlated topic, so it is likely a good area for future research as this discrepancy does not seem to have an obvious logical reason. Law is quite heavily linked to politics and is likely why it is positively correlated in predicting the number of comments.

The negatively correlated entities are quite similar to those in the titles with cardinal, quantity, date, event, and LOC in the blurb, with coefficients of -0.895, -0.869, -0.699, -0.649, and -0.55 respectively. These reasons for these are likely negatively correlated for similar reasons to those described in the title. Date and cardinal in the blurb tend to convey factual uncontroversial information, events in the blurb because of their relationship to sports and LOC in the blurb because it has little interest and is also not controversial. The only new entity is quantity in the blurb but this is likely negatively correlated for the same reason as date and cardinal in that it is likely to represent factual uncontroversial information.

6.2.5 Entities in the Article

As mentioned before, the entities in the article seemed to have a much weaker correlation when compared to those in the title and the blurb, likely due to readers commenting without reading the full story. There are also much fewer important entities in this regard for this reason.

The positively correlated entities in the article are NORP (0.963), Organisations (0.75), and people (0.409). This seems to suggest that once in the article the use of entities for the most part is positively correlated to the number of comments received. NORP and person have both appeared in the title and blurb and the reasoning is the same - that readers tend to have strong opinions towards these specific people and groups. They are also both related to politics, which is the most positively influential topic. Organisations likely follow in a similar way with readers having strong opinions towards them and also due to the fact that they have a relationship with politics.

The only strongly negative correlation found was for the date in the article. This is likely similar to earlier - where the date is linked to sports - and repeated use of dates likely talks about multiple games and when they are happening.

6.2.6 Polarity in the Title

Title polarity is negatively correlated, -0.103. This means that negative stories, with negative polarity, are more likely to receive comments, although this factor is only weakly correlated relative to many of the other features. This can only show if there is a preference to positive or negative titles and not the extent of this.

6.2.7 Polarity in the Blurb

Blurb polarity is similarly negatively correlated, -0.113. This is slightly more relevant compared to the title; but both are still very weak with little effect prediction relative to other features.

6.2.8 Polarity in the Article

Article polarity, -0.192, is more relevant than the other two polarities, meaning that a negative polarity in the article is more likely to incite a higher proportion of comments than a positive one when compared to the other two polarities.

6.2.9 Absolute Polarity in the Title

The absolute title polarity is larger than the title polarity, -0.162. This means that the magnitude of polarity in this title is more relevant than either it being positive or negative and that titles with polarity receive fewer comments.

6.2.10 Absolute Polarity in the Blurb

In the blurb the absolute polarity is -0.296 - which is quite significant. This means that polarity in the blurb is likely to get a reader to comment, more so than in any other section. Again, it is beneficial if it is negative as shown by the polarity above. It is hard to make a judgement on positively polarised blurbs but there could be more research done into the differences that arise here.

6.2.11 Absolute Polarity in the Article

The absolute polarity in the article is weak with only a -0.097 coefficient. As this is lower than the polarity in the article, it means that the amount of negative polarity has a greater effect than the magnitude of the polarity.

6.2.12 Time-Related Features

This section was broken down further into three more sections including years, month of the year, and days of the week. None of which seem to have a significant impact on the model. There are some individual features that are important, but overall there is no time related feature that seems to have a big effect on the number of comments an article receives.

The years are interesting as it would follow from the views that the later years would get more comments as there are more people viewing the articles. However, this is not the case. 2015 receives the most comments with a positive correlation of 0.092, followed by 2016 with 0.078. It is not significantly smaller but it is interesting that users seem to have commented less in a later year. The opposite is seen between 2014 and 2013 with coefficients of -0.161 and -0.037 respectively. This is interesting for the opposite reason and the data shows that although there is an overall trend, there is not a definitive correlation between views and number of comments - at least in their correlation over time.

The months of the year had the next most important features, but for the most part these correlations were not strong. May is the most positively correlated, with a coefficient of 0.405. This is followed by February with a correlation of 0.202. Initially, there seems to be no reason these two months would be the most commented but research into events over this time shows local and European elections in May 2014, and a general election in February 2016. These were the only major elections that appear over this time span and are a good indicator as to

why these months were popularly commented on. June was the most negatively correlated month with a coefficient of -0.426, there does not seem to be any clear reason for this apart from it being after May in which there were elections held. This link is based on evidence solely related to elections seeing as these are the main political events that occur but it is also possible that there are other reasons for these such as June being a month that many people take off work and so would see a decrease in commenters due to readers doing other activities, which is seen in the decrease in views during this month.

The days of the week have some interesting changes from the important features in relation to the views. Sunday is the most important still with a coefficient of 0.154. However, after this is Friday which is negatively correlated in terms of views, but has a positive coefficient of 0.115 in terms of comments. Saturday has become less correlated with a coefficient of -0.025 with no real effect on the number of comments and Monday has gone from a positive predictor of views to a negative in terms of comments with a coefficient of -0.174. These are some interesting changes and most of them are likely due to a result of the difference between the topics that are traditionally published on these days, as well as their relation to comments and the time and energy readers have. This is another area that would benefit from more research as the data needs to be analysed in a way that retains the link between topic and the days/ other time features in the article.

6.2.13 Title Length

The title length is strongly positively correlated to the number of comments received, 1.092. This is another reason which explains why the hypothesis that readers comment before reading the article is true as longer titles are likely to portray more information and with this information; the reader feels they know enough about the article to comment.

6.2.14 Blurb Length

This further confirms the theory that readers don't always read the article as the blurb length is the second most positively correlated feature, 1.276. This likely means that readers are reading the blurb and if it is long enough they gather information from this and make their assessment on the whole article without reading the content of the article first. There is the possibility that the long blurb is related to the topic, for example political stories tend to have longer blurbs, but the information seems to support the first theory.

6.2.15 Article Length

The article length is also positively correlated to an extent, 0.558. This is interesting as it demonstrates that it seems to be a matter of having a lot of information in determining if a reader is going to comment. The other possible reason is that the length is linked to the

topic or even the authors that seem to get the most comments. Authorship is an interesting area of analysis that could also be performed to determine both their links to the number of comments and views and also to the length of the articles they produce.

6.2.16 Comments Discussion

This experiment gives an overview of the most influential features for determining the number of comments on an article. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence. This does not mean all of the reasoning will definitely be accurate in all situations, but provides a platform for further research to be continued in the area. From the analysis the entities in the title seem to be the most important features, followed by those in the blurb. The topics also play an important role and so too does the lengths of the texts. The time related features and the polarity of the texts seem to be less important in determining the number of comments, but the reasoning and extent of this is also interesting.

6.3 Number of Facebook Shares

The number of Facebook shares that an article receives is a metric for determining user engagement outside of features available on the article page. It allows news organisations to see how readers are sharing their stories outside of their platform and it is important as it is a possible way for news organisations to receive new readers, this is why it is made easy to use by the-journal.ie. This experiment gathered all of the features related to the text and content that were available to the reader, as well as some time-based features to produce the model. The model will be limited by features outside of those mentioned, such as pictures, layout, other news websites, and even stochastic reader behaviour. As such, the optimal results for the model are obviously quite a bit lower than 1 for all of the metrics. So although these metrics are important to an extent in determining how effective the model is, there is still room to analyse the features used in determining their relationship to the number of Facebook shares.

The features used in this experiment were chosen based on their availability to the reader at this point, after the reader has opened the article. They are all features that are either time-based and outside the control of the news organisation, or are related to the title, blurb, or article, as these are the main components that go into making up the overall presented article. There are likely many features that haven't been analysed here that also have an effect on the outcome, such as the number of current Facebook shares as well as countless other features including the layout and any related media in the article. The features chosen were broken down into further sections: entities in the title, blurb, and article, polarity in the title, blurb, and article, time-related features, lengths in the title, blurb, and article, absolute polarity in

the title, blurb, and article, and topic covered. Each of these was gathered directly from the title, blurb, and article either using text analysis or by scraping displayed features.

As before, the objective was to determine which features led to the most popular news stories which is why the top 10% of the number of Facebook shares was the barrier that was determined instead of just over 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets but is representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in Figure 5.3 was small with a change of only 0.015 over the range. The value of 0.001 was chosen due to the graph as it maximises both precision and accuracy. The graph dips after this, before slowly rising again - but at no other point has better metrics. This small value of C risks under-fitting to an extent but it seems best for this model.

6.3.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.538	0.514	0.544	0.528	0.539

Table 6.3: Model performance metrics for the number of Facebook shares

The model performed worse than both of the previous two models and there are two main reasons for this. The first is similar to the previous two models in that there are other external and non-measured features that go into determining the number of Facebook shares that an article receives. Like the previous models, this model was not optimised due to the optimal features not being selected, and inter-model analysis not being performed to select the best model for this experiment, meaning the results are worse than they could have been if these measures were accounted for. This model also had the drawback of the Facebook share feature being utilised to a less significant extent than the other models. This means that it is predicting based on a much lower output and there is therefore a much smaller range for it to be predicting, and this means any randomness in the data will be amplified. Despite this, the results will still be useful as they will provide some insight into the relationship between the features in the analysis and the number of Facebook shares received. With the results seen in 6.3, it is clear that the correlation is not as strong as in the other models, this is also partly due to the C value chosen, but there is still a relationship between some of these features

and the Facebook shares. Due to the weak correlation of most of these features, only the strongest features will be analysed.

6.3.2 Topics

Topics play an important role in determining the number of Facebook shares with a few of them having large significance. Topic 2 (Politics, 0.011) and Topic 4 (Finance, 0.009) appear to be the most positively correlated topics. Topic 6 (Lifestyle and Social media, -0.012) and Topic 8 (Journalism, -0.01) are the most negatively correlated. There is no clear correlation between these topics. This might have to do with reader age as politics and finance are usually of higher interest to those in an older age demographic especially when compared to lifestyle and social media. Interestingly though articles on social media don't tend to be shared on Facebook.

6.3.3 Entities in the Title

The entities in the title appear to only have a small correlation relative to the other features. The top three features of note are person, organisation, and GPE in the title with coefficients of 0.00303, -0.00301, and -0.00285 respectively. This suggests that news which focuses on individuals is more likely to be shared compared to less personal news that focuses on organisations or countries. This confirms the work done by Almgren and Olsson (2) which describes how Facebook is used for more local and personal stories compared to other social media platforms.

6.3.4 Entities in the Blurb

Similar to entities in the title, there appears to only be a small correlation between these features and the number of Facebook shares. The most significant of which are both positively correlated - Cardinal and NORP, 0.0029 and 0.0028 respectively. NORP suggests a link to political groups as we know these are important from the topics and cardinal numbers suggest that the information tends to be specific, and could be due to its relationship to finance as another important topic.

6.3.5 Entities in the Article

There is almost no correlation between any of the entities in the article with many more entities in both the title and the blurb being more influential. The most important is person in the article which is negatively correlated, -0.0012, but this begins to get quite insignificant to the overall model prediction. This is possibly due to its link to sports which involves the names of a lot of players (but again this is not particularly significant).

6.3.6 Polarity in the Title

Title polarity is negatively correlated, -0.0006. This is relatively insignificant so it is likely that polarity, either positive or negative, is insignificant in the title for predicting the number of Facebook shares.

6.3.7 Polarity in the Blurb

Blurb polarity is positively correlated, 0.0069. This is relatively significant so it is likely that positive polarity in the blurb is more likely to lead to Facebook shares.

6.3.8 Polarity in the Article

Article polarity is negatively correlated, -0.007. This is relatively significant so it is likely that negative polarity in the blurb is more likely to lead to Facebook shares. This is interesting as different polarities in each of the different sections of the article leads to higher Facebook shares. This means that polarities must be inconsistent throughout the different article sections and that some article types are more likely to have strong polarities in different sections.

6.3.9 Absolute Polarity in the Title

The absolute title polarity is much larger than the title polarity, 0.0038, this means that the magnitude of polarity in this title is more relevant than either it being positive or negative.

6.3.10 Absolute Polarity in the Blurb

In the blurb the absolute polarity is 0.0038. This is the same value as in the title and means polarity in each is beneficial for the article. For the blurb, the polarity being positive is much more important than in the title.

6.3.11 Absolute Polarity in the Article

The absolute polarity in the article is lower than the other two with a value of 0.0026. This means that polarity in the article is still positively correlated to a higher number of Facebook shares, especially if it is negative, but the effect is lower than the other two absolute polarity features.

6.3.12 Time-Related Features

This section was broken down further into three more sections including years, month of the year, and days of the week. These sections seem to have the largest effect on the model with the three most positively correlated features being time-related and the five most negatively correlated features. This means that the use of the Facebook sharing feature is heavily reliant on the dates a reader sees an article as opposed to any of the other article features.

The years are interesting, as they begin to follow a pattern with 2013 being the most negatively correlated year, -0.015, and the second most negatively correlated feature. This is followed by 2014, which is the fifth most negatively correlated feature overall, -0.012. The interesting fact here is that 2015 is the most positively correlated feature overall, 0.027, meaning that there was a large increase in the use of this feature between 2014 and 2015. For 2016, the coefficient decreases again to 0.0015, this means that there was another sharp decline after 2015; meaning that there was a spike in the Facebook sharing feature usage in 2015. There is no clear reason for this, but likely is due to an external factor, such as a change in the layout to make the feature more prominent or an increase in Facebook users overall during this time period.

The months of the year appear to be the next most important feature with nearly all of the months having strong relations to the number of Facebook shares. These don't seem to be split into any specific time of the year with March (0.017) and May (0.015) being two of the most positively correlated features, and April (-0.014) being one of the most negatively correlated. December (-0.019) is the most negatively correlated overall. It appears that it is linked similarly to the months in the comments in that there appears to be more activity around months that have elections, such as May and February (0.011), and not as much activity when readers are likely busy with other things in life such as December and September (-0.01).

The least important of these features is the days of the week, although these are still quite important with Saturday being the third most negatively correlated, -0.015. The trend in the days of the week seems to suggest the opposite of both of the previous models, as the weekend appears to be the least popular time to share on Facebook with the coefficient for Friday also being negative, -0.01. Sunday, however, is relatively unimportant with a coefficient of 0.0007. Wednesday (0.009), Tuesday (0.008), Monday (0.004), and Thursday (0.003) are all positively correlated to some extent showing that during the week seems to be one of the more important features in determining if an article will get shared.

6.3.13 Title Length

The length of the title is weakly correlated, 0.0009, meaning it probably has no effect on an article being shared on Facebook.

6.3.14 Blurb Length

The blurb length is similar to the title length, 0.0007, meaning it also likely has no effect on the article being shared on Facebook.

6.3.15 Article Length

The article length coefficient appears stronger than that of the other two length features, -0.0017. This is likely due to readers not wanting to share an article with if it would take too long to read, or if it was too long to read themselves.

6.3.16 Facebook Shares Discussion

This experiment gives an overview of the most influential features for determining the number of Facebook shares on an article. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence. This does not mean that all of the reasoning will be accurate in all situations, but provides a platform for further research to be continued in the area. From the analysis, the time-related features seem to be the most important features, with the years being the most important, then the months, followed by the days of the week. These features are followed by the topics as they also play an important role. The entities in the texts seem to be the least important in determining the number of Facebook shares and the difference between this model and the others is interesting.

6.4 Number of E-mail Shares

The number of e-mail shares that an article receives is a metric for determining user engagement outside of features available on the article page. It allows news organisations to see how readers are sharing their stories outside of their platform and it is important as it is a possible way for news organisations to receive new readers, but it is less effective than Facebook sharing at reaching a large proportion of people. This experiment gathered all of the features that were available to the reader related to the text and content, as well as some time-based features to produce the model. The model will be limited by features outside of those mentioned, such as pictures, layout, other news websites, and even stochastic reader behaviour. As such, the optimal results for the model are obviously quite a bit lower than 1 for all of the metrics. So although these metrics are important to an extent in determining how effective the model is, there is still room to analyse the features used in determining their relationship to the number of e-mail shares.

The features used in this experiment were chosen based on their availability to the reader at this point, after the reader has opened the article. They are all features that are either

time-based and outside the control of the news organization, or are related to the title, blurb, or article, as these are the main components that go into making up the overall presented article. There are likely many features that haven't been analysed here that also have an effect on the outcome such as the number of current e-mail shares as well as countless other features including the layout and any related media in the article. The features chosen were broken down into further sections: entities in title, blurb, and article, polarity in title, blurb, and article, time related features, lengths in title, blurb, and article, absolute polarity in title, blurb, and article, and topic covered. Each of these was gathered directly from the title, blurb, and article either using text analysis or by scraping displayed features.

As before, the objective was to determine which features led to the most popular news stories, which is why the top 10% of the number of e-mail shares was the barrier that was determined instead of just over 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets but is representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in the GRAPH was small with a change of only 0.015 over the range. A C value of 0.1 was chosen due to the graph as it maximises both precision and accuracy.

6.4.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.562	0.58	0.581	0.581	0.562

Table 6.4: Model performance metrics for the number of e-mail shares

The model performed better than the model predicting Facebook shares in all metrics, but worse than the first two models. There seems to be a stronger correlation than there was with Facebook shares, but like Facebook shares there are other external and non-measured features that go into determining the number of e-mail shares that an article receives compared to the number of comments and views. Like the previous models, this model was not optimised due to the optimal features not being selected and inter-model analysis not being performed to select the best model for this experiment, meaning that the results are worse than they could have been if these measures were accounted for. This model also had the drawback of the e-mail share feature being utilised to a less significant extent than the first two models. This means that it is predicting based on a much lower output and there is therefore a much smaller range for it to be predicting, and this means any randomness in the data will be

amplified. Despite this, the results will still be useful as they will provide some insight into the relationship between the features in the analysis and the number of e-mail shares received. The coefficients, as seen in 5.7, are larger than Facebook shares due in part to the larger C value and provides a good platform for analysis of the features.

6.4.2 Topics

The topics appear relatively unimportant compared to other features in determining the number of e-mail shares received by an article. The negatively correlated topics are also much more important than the positively correlated ones. Topic 0 (Crime, -0.13) and Topic 2 (Politics, -0.13) are the most negatively correlated. This means that there is a distinct difference in their usage compared to Facebook shares, where politics was the most positively correlated topic. In this experiment Topic 1 (Other, 0.047) was the most positively correlated, which only tells us that there is no specific topic users prefer to share through e-mails. This exemplifies the fact that topics seem to be relatively unimportant in determining the number of e-mail shares received.

6.4.3 Entities in the Title

The entities in the title are a more important feature, but are similarly skewed with a higher number of large negatively correlated features. Money, ordinal, and product in the title are all negatively skewed with coefficients of -0.17, -0.16, -0.14 respectively. These features are usually unimportant in determining other user engagement metrics, but it makes sense in this scenario as these are often stories that have to be read to get an understanding, particularly articles about products or money, compared to the most positively correlated features which are cardinal (0.1) and NORP (0.1) in the title which can give away much more information in the title.

6.4.4 Entities in the Blurb

Interestingly, as opposed to entities in the title, in the blurb money is the most positively correlated entity, 0.229. This likely means that when it is used here there is more room to explain its significance than when it is used in the title, meaning that readers are interested in the reason for the value rather than the value itself. Cardinal, date, and person in the blurb are the next most positively influential features, 0.172, 0.145, 0.124 respectively. Cardinal and date are interesting as they appear to be the inverse of how they appear in the number of views and comments where they are both very negatively correlated, this confirms that the e-mail feature is based more off of the reader only really wanting to share the title and the blurb and not being focused on other forms of interaction. Person is likely positively correlated for the same reason as in the other scenarios as people are generally considered influential in

the other user engagement metrics.

Time is the most negatively correlated feature in the blurb, -0.117, likely again due to it being the opposite of the number of views and comments both of which have time as an important feature. Readers are not likely to share a roundup of articles such as the "5 at 5" and are more likely to send specific interesting articles.

6.4.5 Entities in the Article

Similar to the Facebook shares, there is almost no correlation between any of the entities in the article, with many more entities in both the title and the blurb being more influential. The most important is cardinal which is negatively correlated, -0.046, but this is quite insignificant to the overall model prediction. The reason for this is possibly due to its link to crime which includes numbers and statistics but it is relatively unimportant overall as are all of the entities in the article.

6.4.6 Polarity in the Title

Title polarity is positively correlated, 0.017, but only to a small extent. This means there is likely no effect on the number of e-mail shares based on the title being positive or negative.

6.4.7 Polarity in the Blurb

Blurb polarity is also positively correlated, 0.008. This is also only to a small extent and means there is likely no effect on the number of e-mail shares based on the blurb being positive or negative.

6.4.8 Polarity in the Article

Contrary to the other two polarities, article polarity is negatively correlated, -0.024. This however is also quite small meaning it has a little effect on the number of e-mail shares based on the article being positive or negative.

6.4.9 Absolute Polarity in the Title

The absolute title polarity is also quite small, being slightly positively skewed, 0.024, meaning that more polarity in the titles, the more likely they are to receive more e-mail shares but not to a significant extent.

6.4.10 Absolute Polarity in the Blurb

In the blurb, the absolute polarity is 0.173. This is much more significant than the absolute polarity in the title, meaning that polarity in the blurb positively contributes to the article being shared by e-mail. This is interesting as it is again contrary to both the number of comments where the absolute polarities are negatively correlated and the views where they do not appear to be important. This again shows that the articles that are shared through e-mail are very different to those that are popular using the other user engagement metrics.

6.4.11 Absolute Polarity in the Article

The absolute polarity in the article is also important with a positive correlation of 0.171. This confirms that in general more polarising articles are shared by e-mail relative to those viewed or commented upon. This seems to be the only feature in which the content of the article is important and this is likely due to the fact that these articles share polarity between the blurb and the article.

6.4.12 Time Related Features

This section was further broken down into three more sections including years, month of the year, and days of the week. As with the number of Facebook shares, the time related features appear to be the most influential in determining the number of e-mail shares received by an article, accounting for the four most positively correlated features and the top five most negatively correlated features.

In this case, the years don't appear to follow a trend. 2016 (0.188) and 2013 (0.182) are the most positively correlated and 2015 (-0.208) and 2014 (-0.163) are the most negatively correlated. This appears to suggest that readers used the feature in the past, 2013, then stopped using it as much in the subsequent years followed by an uptake again in 2016. It seems that the rise in Facebook sharing through this time likely played a role in readers preferred method of sharing, as it peaked in 2015 when e-mail shares were at their lowest.

The months of the year appear to be the most significant predictor in the number of e-mail shares with nearly all months having strong relationships to them. The months are split similarly to before with many of the less busy periods of time being positively correlated April (0.361), May (0.325), October (0.307) and July (0.306) and the negatively correlated being busier times of the year December (-0.264), October (-0.264), or times when people traditionally holiday abroad June (-0.2). The outlier is the most negatively correlated month which is November (-0.51). This has often been correlated with the less busy times of the year in the other experiments and is likely here due to the differences in the types of articles that are shared through e-mails and those that garner other forms of user engagement, as

discussed.

The days of the week are also important, but not to the same extent as the other time related features. Interestingly, the end of the weekend appears to be positively correlated Sunday (0.138) and Monday (0.074) compared to the start of the weekend, Saturday (-0.144). This is interesting and does not have a clear reason for the drastic change but could be related to the topics often discussed on these days. The other days of the week are relatively unimportant.

6.4.13 Title Length

Title length is strongly positively correlated with a coefficient of 0.204, this means that a long title is likely to get more e-mail shares. This is likely due to the reasons discussed that a reader appears to use this feature primarily based off of the title and article, therefore a longer title can provide the user with more information.

6.4.14 Blurb Length

The blurb length is likely similarly positively correlated because of this reason, 0.026. It is a much weaker correlation so has less of an effect compared to the title but it still has some influence.

6.4.15 Article Length

The article length is also positively correlated, 0.005, which is much weaker than either of the other two lengths meaning there is almost no correlation. This is due to the fact that the article plays a much smaller role in determining the number of e-mail shares compared to the title and blurb.

6.4.16 E-mail Shares Discussion

This experiment gives an overview of the most influential features for determining the number of e-mail shares on an article. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence. This does not mean all of the reasoning will definitely be accurate in all situations but provides a platform for further research to be continued in the area. From the analysis the time related features seem to be the most important features, with the months being the most important, then the years, followed by the days of the week. These features are followed by the absolute polarities, and entities in the title and blurb as they also play an important role. The entities in the article seem to be the least important features in determining the number of e-mail shares and the contrast between this model and the others is interesting.

6.5 Polarity in Comments

The overall polarity in the comments is important in determining if the readers are reacting positively or negatively to the article. This is not as beneficial to news organisations, but should provide an insight into the minds of the readers and which features elicit different responses. This experiment gathered all of the features that were available to the reader related to the text and content, as well as some time based features to produce the model. The model will be limited by features outside of those mentioned, such as pictures, layout, other news websites, and even stochastic reader behaviour. As such, the optimal results for the model are obviously quite a bit lower than 1 for all of the metrics. So although these metrics are important to an extent in determining how effective the model is, there is still room to analyse the features used in determining their relationship to the comment polarity.

The features used in this experiment were chosen based on their availability to the reader at this point, after the reader has opened the article. They are all features that are either time-based and outside the control of the news organization, or are related to the title, blurb, or article, as these are the main components that go into making up the overall presented article. There are likely many features that haven't been analysed here that also have an effect on the outcome such as the current polarity in the comments when a new comment is left, as well as countless other features including the layout and any related media in the article. The features chosen were broken down into further sections: entities in title, blurb, and article, polarity in title, blurb, and article, time related features, lengths in title, blurb, and article, absolute polarity in title, blurb, and article, and topic covered. Each of these was gathered directly from the title, blurb, and article either using text analysis or by scraping displayed features.

Due to the nature of the user engagement metric, the objective was to determine if the overall polarity was positive or negative. As opposed to the other experiments, this meant that the entire data set could be used as the classification was if the overall comment polarity was positive or negative.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in 5.5 was small with a change of only 0.01 over the range. A C value of 1 was chosen due to the graph as it maximises both precision and accuracy. It also reduces the risk of either under-fitting or over-fitting.

6.5.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.694	0.705	0.694	0.7	0.694

Table 6.5: Model performance metrics for the polarity in the comments

The model performed the best out of all of the models predicting the right polarity about 70% of the time. The correlation is likely to be the strongest in this model compared to that of the other models which shows that the features selected are good for this task. However, like the previous models, this model was not optimised due to the optimal features not being selected and inter-model analysis not being performed to select the best model for this experiment, meaning the results are worse than they could have been if these measures were accounted for. There were also likely other features not gathered but despite this the model still performed well and the coefficients should give a clear example of the relationships each of the features has to the polarity in the comments.

The polarity in the comments is different to all of the other metrics as there is not optimal output based on the features. In all of the other experiments the features that optimised it were identified. The objective of the news organisation is to optimise these different features to improve the user engagement metrics, however this is not the case in terms of polarity in comments as this is an analysis into reader behaviour as opposed to determining optimal features for engagement. This analysis will provide details on features readers respond positively to as well as an overall trend in the data. An important thing to note is that this does not look at absolute polarity as is examined as a feature. This means that if a feature is highly polarising with strong opinions both positively and negatively towards it, it will not appear in the analysis. This is an area for future research extending from this thesis.

6.5.2 Topics

The topics are important in determining the polarity of the comment section. This is likely due to users having an overall similar sentiment to the same topics. All topics are also positively correlated which seems to suggest that the overall polarity in the comment section of journal articles are also positive.

Topic 7 (Sport, 1.78) is the most positively influential feature. This means that the overall response to sports articles is very positive. There are likely a few reasons for this, such as sport playing a big role in society with many people strongly attached to local, national, and international teams. Topic 6 (Lifestyle and social media, 1.05) and Topic 4 (Finance, 0.96) are the next most important two topics, the lifestyle stories are often feel good stories which are designed to produce a positive response and the finance stories are likely positive due to the

stable economy over this time, resulting in generally positive stories in this space. Comparing the polarity in the comments to the polarity in specific topics would be another interesting topic for future work.

The two least positively correlated topics are Topic 5 (Accidents/Crime, 0.104) and Topic 0 (Crime, 0.162). These topics both discuss dark subject matter so it is obvious to see why this does not result in overall positive comment polarity. As mentioned above there were no negatively correlated topics meaning these were the lowest values of correlation. It is interesting that these are not negative even though they are the lowest performing showing that overall the comments of the journal are predominantly positive.

6.5.3 Entities in the Title

As with most of the other features the entities in the title are skewed to positive correlation, due to the overall positive nature of the comments. An interesting feature of these is that there are important features that are not relevant to any other user engagement metrics such as a work of art and FAC in the title.

Work of art in the title (0.824) is the most positively correlated feature. This is likely due to its rarity in this section and the stories which contain works of art tend to be quite positive. Event, LOC (mountain ranges, water bodies etc.), and money are also very important features in the title with coefficients of 0.73, 0.716, and 0.622 respectively. LOC is likely due to the same reasons as works of art as they occupy a similar place in the public consciousness as stories in this area are also rare but generally positive. Events are likely important due to their relation to sport and money due to its relation to finance both of which are positive topics.

FAC (buildings, airports etc.), GPE (countries, cities etc.), Time, and NORP (nationalities, religious and political groups) are the most negatively correlated with coefficients of -0.379, -0.368, -0.323, and -0.32 respectively. FAC is likely negative due to it being a rare topic and the main reason for it to be mentioned in an article is due to a problem in this area. This is similar to GPE and NORP where their main usage in a title usually relates to a problem that has occurred with them rather than a positive reflection. Time is an interesting feature and based on its usage in the journal as a feature in the title likely relates to the articles that provide an overview of stories for the day meaning there must be a reason for these overviews to elicit more negative reactions than the normal stories themselves.

6.5.4 Entities in the Blurb

The important entities in the blurb appear similar to those in the title with the exception of FAC. Again it is interesting that the majority of entities in the blurb are more positive than

negative, leading to the same conclusion about the overall trend in the data.

The LOC, FAC, and work of art in the blurb are the most positively correlated with coefficients of 0.917, 0.9, and 0.631 respectively. LOC and work of art are likely due to similar reasons as before due to the fact that these stories are rare and when they feature they tend to be positive therefore eliciting a positive reaction from the readers. FAC is interesting due to it being negative in the title. The reason for this appears to imply a difference in usage between these two. In the title the FAC is likely the focus of the article and this only appears when there is a problem in one as described above, in the blurb however, it is much more likely to be a stadium mentioned in a report about a game or other stories in which it is not the main focus. Stories that relate them to sport which is overall positive is likely the reason for this difference.

Time, date, and NORP in the blurb are negatively correlated features, this is likely due to the similar reasons as in the title for NORP and time. NORP due to its mention usually being around a negative story surrounding the group as positive stories are less likely to be reported on in this case. Time due to its link to the overview of stories in the journal context as mentioned before. There does not seem to be an obvious reason for why a date in a blurb would produce negative sentiment but it must be due to the types of stories in which they appear. This further shows how there is further work to be done into this area to determine the reasons for some of these relationships.

6.5.5 Entities in the Article

The entities in the article are some of the most important features with both the highest and lowest correlation being entities in the article. This is also seemingly the only group of features in which the positive and negatives seem even over the data set.

Person, ordinal, and cardinal in the article are the most important positive features with coefficients of 1.996, 0.624, and 0.563. Person is likely due to its relationship to sports and sports stories tend to include a large number of names. Ordinal and cardinal are also likely related to sports to an extent but also likely deal with some of the other positive topics such as finance.

The negatively correlated entities in the article are NORP, Organisations, and GPE with correlation coefficients of -0.909, -0.684, -0.605 respectively. These have been explained above but likely relate to the fact that these entities are discussed more frequently in stories which they are presented in a negative light and therefore receive more negative comments.

6.5.6 Polarity in the Title

The polarity in the title is 0.775, meaning that a positive title is strongly correlated to a positive polarity in the comments and the same is true for negative comments. This analysis is limited based on the feature used as a measure for the total positiveness and total negativeness in polarity for these features would have allowed a more in depth analysis. All this shows is that positive titles are likely to receive positive comments and negative titles receive negative comments but not the specific magnitudes of each.

6.5.7 Polarity in the Blurb

The same is true for the blurb but to a lesser extent, it's polarity is 0.35. Positive blurbs lead to positive polarity in the comments and negative blurbs lead to negative polarity in the comments.

6.5.8 Polarity in the Article

The same is true for the article but to a greater extent, it's polarity is 1.04 so it is very strongly correlated. Positive articles therefore lead to positive polarity in the comments and negative articles lead to negative polarity in the comments to a larger extent than in the other sections of text.

6.5.9 Absolute Polarity in the Title

The absolute title polarity is small and negatively correlated, -0.062. This is interesting as it means that stronger polarities in the title are more likely to result in negative polarity in the comments, likely meaning the relationship between negative polarity in the title and negative comments is stronger than the relationship of positive polarity in the title to positive polarity in the comments identified in the section above.

6.5.10 Absolute Polarity in the Blurb

The absolute blurb polarity is similar to the title but slightly larger, -0.109. This has a similar effect meaning that stronger polarities in the blurb are more likely to result in negative polarity in the comments. The same effect is true about the strengths of the polarities.

6.5.11 Absolute Polarity in the Article

The absolute article polarity is weakly positive, 0.098. A similar relationship exists except it is slightly the opposite meaning that strong sentiment in the article is more likely to results in positive polarity in the comments. It also means that the a positive polarity in the article

is more likely to lead to positive comments than a negative one is to lead to negative comments.

6.5.12 Time Related Features

This section was broken down further into three more sections including years, month of the year, and days of the week. The time related features appear unimportant in predicting the polarity in the comments. There therefore seems to be very little correlation between the two.

Looking at the overall trends in the years it appears that comment polarity has gotten more negative over time with 2013 having a most positive correlation of 0.171 and 2015 having the most negatively correlated of -0.102. These are both relatively insignificant overall and it seems to only be a small trend in the data hinting at a possible correlation between more readers and negative sentiment.

The months appear to also have very little effect with January being the most positively correlated, 0.159, and June and November being the most negatively correlated with coefficients of -0.113 and -0.107 respectively. There does not seem to be any clear reason for these results and it is likely due to there not really being any correlation between the month of the year and the polarity in the comments.

There seems to be even less correlation between the day of the week and the polarity in the comments with no coefficient with a magnitude larger than .07, meaning in effect there is no correlation.

6.5.13 Title Length

The title length is strongly negatively correlated, -0.502, meaning that the longer the title the more likely readers will respond negatively. This is possibly due to the types of articles that have long titles being related and a possible correlation between title polarity and length, although there is no further evidence for either.

6.5.14 Blurb Length

The blurb length is similarly strongly negatively correlated, -0.726, meaning that longer blurbs are more likely to receive negative comments. This is possibly due to the types of articles that have long blurbs being related and a possible correlation between blurb polarity and length, although there is no clear evidence for either. This is the second most negatively correlated feature meaning that there is a strong reason for longer blurbs being negative.

6.5.15 Article Length

Interestingly this is true for the article length as well, it is less negatively correlated but is still -0.19. This appears to suggest that longer text seems to leave readers with a negative sentiment regardless of where it occurs and that shorter text in all cases is beneficial for positive polarity in the comments.

6.5.16 Comment Polarity Discussion

This experiment gives an overview of the most influential features for determining the polarity of comments on an article. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence. This does not mean all of the reasoning will definitely be accurate in all situations but provides a platform for further research to be continued in the area. From the analysis it appears that the topic, lengths of the texts, and entities in the article are the most important features in predicting the polarity in the comments. This is followed by the entities in the title and blurb. The polarity is also important to an extent although further research could garner more insight into this relationship. Interestingly all of the time related variables seem to have very little effect on polarity in the comments.

6.6 Number of Likes on Comments

The number of likes that a comment receives is likely based on a large number of features, similar to the other metrics analysed in this thesis. The difference between the other metrics and those based on the comments is that these are related to user-generated features. This means that this information is less important to the news organisation and is more of an analysis of user behaviour and interaction in itself. This research is also very limited as this was not the main focus of the thesis, the number of features are therefore significantly reduced although the data set which it was performed on was quite large, having 3,945,580 comments to learn from. Similar to the other models, the feature set is limited based on time constraints and the data collected. As such, the results for this section perform significantly worse than the other models. However, this minor analysis is still interesting as it shows the effect of some basic features on user interaction.

There are only three features used in this model and they are polarity, absolute polarity, and comment length. These were the three easiest features to extract from the comments but a future look into this work could analyse this to a further extent, such as gathering more features and performing more in-depth text analysis.

The objective was to determine which features led to the most likes for each individual comment, specifically targeting those that receive the highest number which is why the top

10% was used instead of just setting the threshold at 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets but is representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in Figure 5.6 was small with no more than .005 between the highest precision and the lowest accuracy. The value of .001 was chosen as this produced the optimal results for the graph. Under-fitting is a bit of a concern because the value is so low but it seems to be the optimal value for the graph.

6.6.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.53	0.709	0.525	0.603	0.534

Table 6.6: Model performance metrics for the number of likes on comments

The model performed poorly in general, with a low accuracy score despite its high precision. This data gathered from this model appears to show just a general trend in the data without being able to make any definitive predictions in any particular case. This is due to the number of external features not used in the analysis which is too large in this case. The data definitely seems to be better than a standard 50% guess based on the size of the data set, but it is not significant so the analysis will be short.

6.6.2 Polarity

Polarity, 0.005, appears to have no effect on the overall result of the model, meaning that neither positive or negative polarity is a good indicator.

6.6.3 Absolute Polarity

This is the only feature that appears to have any significance, with a coefficient of 0.215 appearing to suggest that comments with high polarity receive more likes. Since there are no dislikes to compare against, this seems to make sense as people will like comments to make them more prominent, and polarising comments would see more people with similar thinking liking the comment.

6.6.4 Length

The coefficient for length is effectively 0, meaning that it has no correlation to the overall prediction.

6.6.5 Comment Likes Discussion

This experiment gives an overview of the most influential features for determining the number of likes a comment receives based on the features collected. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence, where applicable. This does not mean all of the reasoning will definitely be accurate in all situations but provides a platform for further research to be continued in the area. The main conclusion from this experiment is that there were too few features and that more are necessary to provide a better overview of the metric. There seems to be some correlation between high polarity comments and likes, but not to such a significant extent that definitive conclusions can be drawn.

6.7 Number of Replies on Comments

The number of replies that a comment receives is likely based on a large number of features, similar to the other metrics analysed in this thesis. The difference between the other metrics and those based on the comments is that these are related to user generated features. This means that this information is less important to the news organisation and is more of an analysis of user behaviour and interaction in itself. This research is also very limited as this was not the main focus of the thesis, the number of features are therefore significantly reduced although the data set which it was performed on was quite large having 3,945,580 comments to learn from. Similar to the other models, the feature set is limited based on time constraints and the data collected. As such the results for this section perform significantly worse than the other models, but this minor analysis is still interesting as it shows the effect of some basic features on user interaction.

There are only three features used in this model and they are polarity, absolute polarity, and comment length. These were the three easiest features to extract from the comments but a future look into this work could analyse this to a further extent, such as gathering more features and performing more in depth text analysis.

The objective was to determine which features led to the most replies for each individual comment, specifically targeting those that receive the highest number which is why the top 10% was used instead of just setting the threshold at 50%. This meant that the data set was skewed and only some of the data could be used. The data chosen to be used was randomly selected from the data set and the experiments were performed on many different sets of the

negative data (the data in the bottom 90%), all of which returned similar results. The results here are demonstrated by one of those nine data sets but is representative of the entire data set.

Cross-validation was then performed on this data to determine the optimal C value for the model. The range of values present in Figure 5.7 was small with only a .02 change in the precision and an accuracy . The value of .001 was chosen as this produced the optimal results for the graph. Under-fitting is a bit of a concern because the value is so low but it seems to be the optimal value for the graph.

6.7.1 Model Performance

Accuracy	Precision	Recall	F1-Score	AUC
0.52	0.511	0.439	0.346	0.48

Table 6.7: Model performance metrics for the number of replies on comments

The model performed poorly in general with the lowest accuracy score for all models. The data gathered from this model appears to show a general trend in the data without being able to make any definitive predictions in any particular case. This is due to the number of external features which is too large in this case. The data definitely seems to be better than a standard 50% guess based on the size of the data set but it is not significant so the analysis will be short.

6.7.2 Polarity

Polarity, 0.006, appears to have no effect on the overall result of the model meaning that neither positive or negative is a good indicator.

6.7.3 Absolute Polarity

This is the only feature that appears to have any significance with a coefficient of -0.105, appearing to suggest that comments with low polarity receive more replies, this is interesting as it seems to imply that readers ignore high polarity comments when deciding what to reply to. This is the opposite of the findings for the comments that receive likes.

6.7.4 Length

The coefficient for length is effectively 0 meaning that it has no correlation to the overall result.

6.7.5 Comment Replies Discussion

This experiment gives an overview of the most influential features for determining the number of replies a comment receives based on the features collected. The analysis provided was extracted from the magnitude of the coefficients and the likely reasons for their influence, where applicable. This does not mean all of the reasoning will be accurate in all situations but provides a platform for further research to be continued in the area. The main conclusion from this experiment is that there were too few features and that more are necessary to provide a better overview of the work. There seems to be some correlation between low polarity comments and replies but not to a significant extent that definitive conclusions can be drawn.

7 Conclusion

The objective of this research was to determine the features in the article that were most influential in predicting user engagement metrics. In this thesis the process taken to gather fitting features and use them to create models able to predict the chosen user engagement metrics was described. The models created provided coefficients that correlated to the individual features effect on predicting the desired metric. The effectiveness of these models varied depending on the metric being analysed but overall the models produced provide an insight into the importance of each of the different features on the specific metric. A review of each of the experiments will be followed by a comparison and discussion.

7.1 Conclusion of analysis

7.1.1 Number of Views

The number of views is an interesting topic as it is a measure of how many readers each article is getting, which is an important metric for news organisations. The model performed well in all measures considering the large number of external factors that likely go into predicting this metric. For the-journal.ie the number of views has increased over the last few years with the correlation to each year increasing as the years progress. In terms of other time related features the weekend appears to be when most readers use the website. It also appears that topics with relatable content or shock value perform better compared to topics that contain divisive topics such as politics or niche topics like sports. Entities that contain figure-based information about the article tend to be negatively correlated such as dates and numbers as they potentially give away the interesting information of the story. In contrast, entities with times tend to improve the number of views potentially due to their link to the popular accident/crime topic.

7.1.2 Number of Comments

The number of comments received allows the organisation to determine the amount of interaction they are seeing with their users. This is important in building a community and

encouraging readers to return. The model performed worse than the one predicting the number of views but not to a significant extent. There are also likely more features that go into determining the number of articles which explains the lower performance. The topics were less important in predicting the number of comments but the findings corroborate the research performed by Tenenboim and Cohen (33) and Ksiazek (12) which finds that controversial topics tend to produce more comments. This can be seen in the most commented-on topic, politics, which is by its nature controversial compared to uncontroversial topics like crime that receive fewer comments. It also appears that many readers comment without reading the article as the most influential features all relate to the title and blurb. Longer titles and blurbs also lead to more comments likely due to the fact that they contain more information. Again, articles with figure-based information in the title and blurb tend to perform worse in this area likely due to the fact they contain factual information which would be less controversial. Articles with people, groups, or locations in the title and blurb tend to receive more comments,likely due to these elements being more controversial.

7.1.3 Number of Facebook Shares

The number of Facebook shares an article receives is an important metric for news organisations as it allows them to see the extent to which their article is being shared, it also provides an avenue of free advertising for their articles. This model performed poorly and so there are obviously many more features that were not analysed which are important in predicting this metric. The time-based features appeared more important than in the previous models, all of which are strong predictors. It appears Facebook sharing peaked in 2015 despite the continued growth in popularity of the-journal.ie. Contrary to views, articles that receive more shares seem to be posted during the week with less occurring on the weekend and the number of shares also decreases at times of the year people are otherwise occupied, with back to school, Christmas etc. The topics are also important, but there seems no obvious link behind the motives for sharing based on them.

7.1.4 Number of E-Mail Shares

The number of e-mail shares serves a similar function to the number of Facebook shares by increasing the potential number of readers the news organisation can receive. This model performed better than the Facebook model, but worse than the model predicting the number of comments - again meaning there are likely many other features important in predicting this feature. The time-related features appear to be the best predictors, with a similar importance related to the months as in the number of Facebook shares. The popularity of the feature from 2013 appears to dip in 2014 and 2015 before returning to the same level in 2016. Days of the week and topics are less important features here, whereas features such as the longer

titles seem to be able to predict more shares. Interestingly, money in the blurb predicts more comments, whereas in the title it predicts fewer comments. This likely means that the reasoning, which is easier to explain in the blurb, is more interesting than the figure itself.

7.1.5 Comment Polarity

This model predicts whether a feature leads to positive or negative polarity in the comments so is not like the other model as depending on the story either might want to be maximised. It is also the best performing model in all metrics. The topics are important and generally predict positive comments with sports being the most positive. The article polarity is also a good predictor of the polarity in the comments, likely due to the correlation between what readers read in the article and their sentiment after. Other important features are the entities, particularly in the article. People, art, and nature all seem to predict more positive comments, whereas organisations, groups, and places all predict more negative comments. This is likely due the first set of entities containing more feel-good content, with the second set containing more controversial entities. Longer text in the title, blurb, and article all predict more negative comments. Although the model predicting the polarity in the comments is the best performing model, it is difficult to compare due to the differences in the metric measured. It is therefore referenced where necessary but not able to be giving the same type of comparison. This is still an interesting user engagement metric but is primarily beneficial to analyse in isolation.

7.1.6 Comment Likes

The number of likes a comment receives helps understand user behaviour on the article better. The feature set was limited in this analysis and the metrics show this with poor performances despite the larger data set. This means there are many more features, other than the ones used, that have an effect on the number of likes a comment receives. The only feature that is important is the absolute polarity, meaning strongly polarised comments get more likes. Again due to the overall performance this is only to a small extent.

7.1.7 Comment Replies

The number of replies a comment receives also helps understand user behaviour on the article better. The experiment performed was similar to determining the likes a comment received and so the feature set was also limited in this analysis. The metrics show this with poor performance. This means there are many more features, other than the ones used, that have an effect on the number of replies a comment receives. Interestingly, the same feature is important (the absolute polarity). However, the direction is the opposite meaning that strongly polarised comments are less likely to lead to replies, potentially due to the quality of

the discourse being superior in this case. Again, due to the overall performance, this is only to a small extent.

7.1.8 Model Comparisons

In this section the features that are influential in all of the models and the differences between them will be identified. As the models that predicted the number of views, comments, and the polarity in the comments performed the best, these will be focused on. There are also some interesting findings that relate to the features in the other models and these will also be analysed.

The topics were interesting as there appeared to be a wide variation in the effect they had on the different user engagement metrics. The internet in Ireland, which was positively correlated to both the number of comments and the number of views was negatively correlated to sharing. This was likely due to the fact that the topic is interesting to individuals but not deemed important enough to share with others. Politics is another topic that is either strongly positively correlated or strongly negatively correlated. This is due to the fact that although it is a controversial topic which leads to a large number of comments as demonstrated in previous work, all stories are not interesting to the general public, so unless there is a big story they tend to receive fewer views. This is similar in a sense to sport but to a more significant extent. Sport performed poorly in almost all of the metrics. Sport is considered a very popular topic which made this finding interesting. The likely reason for this is that although most people find sport interesting, their interest in sport is generally niche. People who follow sport generally follow a specific part of sport rather than sport as a whole, be it a club or even a sport but not all sport. This means that generally these topics received fewer views and comments as people tend to ignore sports stories they are not directly interested in. This in a sense makes sports stories niche topics as they by their nature focus on specific areas of sport. One of the main conclusions from this work is that there seems to be a difference in terms of topics that are influential in determining the number of comments and the number of views.

The time-related features also provide an interesting analysis, as there appears to be a large difference in their influence between models. In the cases of the number of comments and the polarity of comments they seem to have little effect, whereas in the remaining models they tend to be some of the most important features. Months of the year seem particularly important in these models with the traditional holiday months such as December and June and busy months such as back to school in September negatively affecting the user engagement metrics. This is consistent across all cases where the months of the year are considered important features. Another interesting time related feature is the days of the week. In contrast to the months of the year, where the features were similar, in terms of days of the

week, sharing is more likely to occur during the week, whereas viewing is more likely to occur on the weekend.

It appears from the data that the title and blurb are more important than the article itself in predicting most of the user engagement metrics apart from the polarity where the opposite is seen. This likely means that the decision to perform either sharing or commenting is decided upon before the article is read. It is interesting that this is consistent for all quantity-based user engagement metrics, apart from the number of views as they don't take into account article features.

Figure-based data is also negatively correlated to most user engagement metrics and in particular the number of views and the number of comments. The reasons behind this are slightly different in both scenarios, due to giving away too much information about the story in terms of views and being less controversial in terms of comments, but both lead to the same negative correlation in including this data in the title and blurb of an article.

The number of views and the number of comments both performed well and due to the size of the data set and the magnitude of their values are the most valuable models gathered from this work. The research gathered from these helps build on much of the research previously performed in this area and provides further insights into some of the important features in the respective areas. This research is valuable both to future researchers in the area and to news organisations looking to optimise their user engagement. This research highlights the importance of certain features in the analysis of user engagement metrics and the extent to which they are important. This area has generally been focused on optimising models but in doing so lacked a comprehensive analysis of the features used in determining these models. This research aims to help fill this gap and provide a platform to build on and gain a more in-depth knowledge of the reasoning behind each of the features effects on the respective user engagement metrics.

7.2 Future Work

This work provides a large scope for future developments in the area. As it builds on research already performed it is clear the area of research which it benefits. There are a few key areas which can be analysed further from the results of this work, as they require a more in-depth study to determine the specific reasoning behind the correlations. The features identified will also contribute to other research in the area focused on the production optimised models for any of the user engagement metrics analysed. The final area is in increasing the feature set used in the analysis as there was room to add more features in all of the models, as identified by their performance metrics. This area is interesting as there are both research and commercial benefits from the work and is a continually evolving area due to the rate of

change in human behaviour particularly in the online sphere.

As most of the work identified in this area focuses on the optimisation of models in predicting user engagement metrics this work is necessary to provide the research behind the importance of the different features used in these models. The features used in these models can be analysed and extracted for use in future model-building and the unimportant features can be avoided. This will help build stronger models as well as give an insight into the reasoning behind the features in determining the model's performance. As this is the main area of focus in this research these features could prove beneficial in contributing to our understanding of these models.

There is scope to increase the size of the feature set used and introduce more features to determine their relative relationships. Due to the performance of the models it is clear that there are more features that contribute to determining the user engagement metrics. This can range from other articles on the page to media in the articles. As this is a large area of study there are more features that can be analysed as well as a closer look at human behaviour while reading the article. These features could give a greater insight into the factors affecting the user engagement metrics and help further optimise models, improving both the research in this area and giving news organisations a greater understanding of their readers.

There are many areas in the research in which it was identified that there is scope for future work. These areas include the type of regularisation, which could play an important role as using lasso instead of ridge in future experiments could help to more clearly identify important features due to the manner in which it handles the coefficients. There are also interesting areas identified in relation to the polarity. These include using absolute polarity of comments as a user engagement metric to determine the effect the features have on it and its relationship to the polarity metric, and determining the relationship between polarity and specific topics as there appears to be a link based on the analysis of the data but the experiments were not set up in such a way as to confirm this.

The work was performed on Irish news media outlet the-journal.ie. This means that the results identified directly apply to this data. There are other types of news organisations, for example those that use a subscription-based model or those in other countries, where similar work could be performed to corroborate the results found or identify differences between these organisation types. Further insight could be obtained by analysing if this research is also applicable to other similar areas of media such as blogging which has a similar structure and format.

The research in this thesis was designed to analyse an area in news media research that had not been addressed sufficiently. This was looking at the relationship between features in the articles and the user engagement metrics. This research helps provide an analysis into why certain features have the relationships they do and compares these relationships. However,

there is still a large scope for research in this area particularly due to the fact that this is an evolving area due to the nature and speed at which the internet moves. The areas above highlight some of the ways in which future research can move and how further can be beneficial to both increasing the research understanding and helping industry in this area.

8 Appendix

8.1 Topic Clusters

Clusters	Key Words
Cluster 0	"police", "man", "old", "officers", "arrested", "year", "woman", "incident", "attack", "people", "killed", "area"
Cluster 1	"people", "irish", "ireland", "today", "year", "dublin", "children", "new", "minister", "water", "court", "source"
Cluster 2	"party", "gael", "election", "fine", "labour", "fianna", "fáil", "sinn", "government", "féin", "dáil", "kenny"
Cluster 3	"source", "twitter", "tweet", "youtube", "couldn", "just", "people", "like", "dublin", "irish", "ireland", "com"
Cluster 4	"cent", "million", "ireland", "company", "year", "irish", "bank", "new", "tax", "billion", "000", "jobs"
Cluster 5	"garda", "gardaí", "man", "station", "arrested", "car", "scene", "road", "incident", "hospital", "dublin", "contact"
Cluster 6	"people", "like", "source", "just", "time", "don", "life", "day", "know", "youtube", "new", "ve"
Cluster 7	"game", "ireland", "team", "inpho", "players", "cup", "league", "final", "munster", "football", "rugby", "win"
Cluster 8	"news", "support", "journalism", "journal", "contributions", "deliver", "stories", "bad", "open", "continue", "help", "important"

Table 8.1: Key words for each cluster

Bibliography

[1] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152:341–348, 2019.

[2] Susanne Almgren and Tobias Olsson. Commenting, sharing and tweeting news: Measuring online news participation. *Nordicom Review*, 37(2):67–81, 2016.

[3] Carl Ambroselli, Julian Risch, Ralf Krestel, and Andreas Loos. Prediction for the newsroom: Which articles will get the most comments? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 193–199, 2018.

[4] Ioannis Arapakis, Mounia Lalmas, B Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M Jose. User engagement in online n ews: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014.

[5] Pablo J Boczkowski and Eugenia Mitchelstein. How users take advantage of different forms of interactivity on online news sites: Clicking, e-mailing, and commenting. *Human communication research*, 38(1):1–22, 2012.

[6] Megan Brenan. Americans remain distrustful of mass media. `https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx`, 2020.

[7] Richard Fletcher and Rasmus Kleis Nielsen. Are people incidentally exposed to news on social media? a comparative analysis. *New media & society*, 20(7):2450–2468, 2018.

[8] Jurgen Habermas and Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991.

[9] Reuters Institute. Use of social media for news amongst irish consumers declines while understanding of how news appears in their social media feeds remains low, 2018.

https://www.bai.ie/en/use-of-social-media-for-news-amongst-irish-consumers-declir
understanding-of-how-news-appears-in-their-social-media-feeds-remains-low/.

[10] Tingting Jiang, Qian Guo, Shunchang Chen, and Jiaqi Yang. What prompts users to click on news headlines? evidence from unobtrusive data analysis. *Aslib Journal of Information Management*, 2019.

[11] Silvia Knobloch-Westerwick and Steven B Kleinman. Preelection selective exposure: Confirmation bias versus informational utility. *Communication research*, 39(2):170–193, 2012.

[12] Thomas B Ksiazek. Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism studies*, 19(5):650–673, 2018.

[13] Thomas B Ksiazek, Limor Peer, and Andrew Zivic. Discussing the news: Civility and hostility in user comments. *Digital journalism*, 3(6):850–870, 2015.

[14] Dmitry Lagun and Mounia Lalmas. Understanding user attention and engagement in online news reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 113–122, 2016.

[15] Eun-Ju Lee. That's not the way it is: How user-generated comments on the news affect perceived media bias. *Journal of Computer-Mediated Communication*, 18(1):32–45, 2012.

[16] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of user engagement. In *International conference on user modeling, adaptation, and personalization*, pages 164–175. Springer, 2012.

[17] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[18] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284, 2009.

[19] Matthew Michelson and Sofus A Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80, 2010.

[20] Apra Mishra and Santosh Vishwakarma. Analysis of tf-idf model and its variant for document retrieval. In *2015 international conference on computational intelligence and communication networks (cicn)*, pages 772–776. IEEE, 2015.

[21] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

[22] Jack Murray. Top 5 most popular online irish news sources. `https://mediahq.com/top-5-popular-online-irish-news-sources/`, 2019.

[23] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[24] Maja Ottosson and Amalia Nordås. The british media portrayal of an "american royal": A critical discourse analysis of the articles published by the british press covering the duchess of sussex during the royal africa tour 2019, 2019.

[25] David F Ransohoff and Richard M Ransohoff. Sensationalism in the media: when scientists and journalists may be complicit collaborators. *Effective clinical practice*, 4(4), 2001.

[26] `https://towardsdatascience.com/logistic-regression-explained-9ee73cede081`.

[27] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[28] Anne Schuth, Maarten Marx, and Maarten De Rijke. Extracting the discussion structure in comments on news-articles. In *Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 97–104, 2007.

[29] Natalie Jomini Stroud, Joshua M Scacco, and Alexander L Curry. The presence and use of interactive features on news websites. *Digital journalism*, 4(3):339–358, 2016.

[30] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[31] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. Ranking news articles based on popularity prediction. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 106–110. IEEE, 2012.

[32] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pages 1–8, 2011.

[33] Ori Tenenboim and Akiba A Cohen. What prompts users to click and comment: A longitudinal study of online news. *Journalism*, 16(2):198–217, 2015.

[34] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1765–1768, 2009.

[35] Fei Wang, Hector-Hugo Franco-Penya, John D Kelleher, John Pugh, and Robert Ross. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 291–305. Springer, 2017.

[36] Christian Wartena and Rogier Brussee. Topic detection by clustering keywords. In *2008 19th International Workshop on Database and Expert Systems Applications*, pages 54–58. IEEE, 2008.

[37] Amy Watson. Online news, newspaper and magazine consumption in great britain 2007-2020. `https://www.statista.com/statistics/286210/online-news-newspapers-and-magazine-consumption-in-great-britain/`, 2021.

[38] Marvin Zuckerman and Patrick Litle. Personality and curiosity about morbid and sexual events. *Personality and Individual Differences*, 7(1):49–56, 1986.