University of Dublin

TRINITY COLLEGE

Identifying patterns in comment sections across news groupings

Holly Baker

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

M.Sc. Computer Science (Year 5)

Supervisor: Professor Owen Conlan

Co-supervisor: Dr Brendan Spillane

April 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

April 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this dissertation upon request.

April 19, 2022

Identifying patterns in comment sections across news groupings

Holly Baker, M.Sc. Computer Science
University of Dublin, Trinity College, 2022

Supervisor: Professor Owen Conlan

The toxicity in comment sections unveiled one of the most significant current discussions in today's literature surrounding the online ecosystem. From misinformation to harassment to self-harm to hate speech, it is apparent that moderating comment is one of the most crucial functions on the internet. Currently, moderation combines artificial intelligence (AI) and human knowledge. For human moderators to fully understand the contours of existing norms and the risk of commenter patterns, they must consider research in this area, looking at the potential tools to support these moderation systems.

The primary aim of this dissertation was to investigate comment sections relative to their news grouping and compile a mountain of observations and recommendations for moderation systems. The appeal of examining coefficients relative to news groupings is due to the general lack of research in moderating comments on the premise of a profile made up of assumptions about a news grouping. Machine learning (ML) is used in conjunction with textual analysis to make the process much faster and more efficient than the manual processing of the comments. The aim is reached by conducting two experiments and analysing the coefficients of the logistic and linear regression models. One experiment will probe whether a comment receives a reply or not and then determine the features that provoke this output. At the same time, the other experiment will scrutinise the components that determine the number of recommendations a comment gains.

The significant findings advise directing resources to comments that contain proven influential themes or scandals, targeting comments flagged by argumentation features, distinguishing patterns that vary across news groupings by ignoring repetitive patterns found in

all groupings, investigating engagement metrics with an ambiguousness nature, and moderating subjective comments. Although a chunk of these observations can be ineffectual to moderator systems, there is still significant knowledge gain an awareness that could help moderation systems tackle the cesspools of racism, misogyny, and all other forms of bigotry found in comment sections.

Acknowledgments

I would like to sincerely thank Professor Owen Conlan and Dr Brendan Spillane for their supervision. Their guidance and advice carried me through all the stages of writing my project, and I am incredibly grateful for their encouragement.

Additionally, I would like to thank my family for their patience and continuous support throughout my study period.

Finally, I would like to give special thanks to my fellow peers for the stimulating discussions and fun we have had in the last five years. I am very grateful for our shared academia journey.

Holly Baker

University of Dublin, Trinity College
April 2022

Contents

List of Tables

List of Figures

Glossary

Comment section A section on a website that encourages user participation in online newspapers by enabling readers to comment on the news. xi

News groupings The different sections a news source is divided into based on everyday topics, the typical news groupings are sports, politics, business, fashion, entertainment, and so on. xi

User engagement The measure of readers taking part in user-content and user-user interactivity on an online newspaper, with particular focus on commenting and upvoting. xi

Chapter 1

Introduction

1.1 Background

Online news plays a vital role in mass communication models. One of the most important events of the 1990s was the information revolution, where communication progressed from orality and literacy to computers (Lapham, 1995). Many industries have had to redefine their mission to reap the benefits of state-of-the-art technology. In 1995, the newspaper industry began to see itself as organisations that supply information rather than organisations that provide newspapers. For this smooth transition of digitising information, papers survived by generating extra available information for readers; doing this captivated people with the idea of online newspapers. Audiences became absorbed as they now possessed the ability to use these online news sources as a tool to help them gain more comprehensive knowledge of the surrounding world.

With this digital changeover, a theme emerged where the new communication revolution shifted power to the people. The new one-to-many models assembled a global village, where readers were relying less on information reported by traditional sources. In contrast to the one-to-many model and the dated letters to the editor, a new communication channel has opened between journalists and the audience. This channel has enabled journalists to interact with their audience, understand their audience more with more effective feedback, and readjust the writing to favour readers' tastes. These recent developments at the time heightened the need for a constant tool to elicit these engagements between journalists and readers. The first comment section emerged in 1998, with little knowledge of how it would come under fire in the years to follow. The real solution to the pain point was in 2007 among the top 100 US newspapers, with 33% of these having enabled articles comments. The following year showed the trend dramatically increased to 75%.

1.2 Problem area

Over the past century, there has been a dramatic increase in news organisations killing their commenting function. The toxicity in comment sections unveiled one of the most significant current discussions in today's literature surrounding the online ecosystem. A substantial problem with comment sections is that they can turn into cesspools of racism, misogyny, sexism, and bigotry. These unfortunate traits have outshone the beneficial opportunities in recent years, such as enabling readers to interact, learn from each other, and discuss stories. According to Santana (2011), 64.8% of reporters disagreed with online comments promoting civil or thoughtful discussion. In addition to this, a study conducted by Pew Research Centre showed that comment sections found on news organisations are particularly contentious in comparison to other platforms (Duggan, 2014). One respondent says in the carried-out survey, "Comment sections of news articles often contain some very racist, homophobic, sexist language." Followed by another noting, "people are brutal and seem to feel way too comfortable in their anonymity." When comments are unmoderated, research has consistently shown that they derail meaningful conversations, dehumanise other fellow readers, devalue expertise, distract, and enrage other readers. From misinformation to harassment to self-harm to hate speech, it is apparent that moderating comments is one of the most crucial functions on the internet.

News sources attempted to manage these toxic internet spaces by injecting policies and systems; comments viewed as not conducive to safe and supportive environments were blocked. Moderation systems were used widely in practice, with most systems combining artificial intelligence (AI) and human moderation. For human moderators to fully understand the contours of existing norms or the risk of commenter's patterns, they must consider research in this area, looking at the potential tools that might support these moderation systems. Additional studies will allow human moderators to make more informed decisions based on the dynamics of the community (Gillespie et al., 2020).

AI systems are bound to enable racist hate speech to circulate (Siapera, 2021). While AI systems won't relieve all the misery of human moderators yet, the existing studies on these systems can elicit several more discoveries of tools in attempts to improve the performance of the moderation system. The current machine learning models can be used to draw broad understandings, from which scholars can provide emerging policies that will help moderation systems reflect. The socio-technical system needs continual oversight, with regular assessments of the machine learning techniques and the operators. The investigation into moderation systems is continuous, and there's vast room for studies investigating the complex matters of improvements to moderation systems and perceived bias.

1.2.1 User engagement

Another important problem area to address is the impact of moderation systems on user engagement. Most businesses set out to drive user engagement to increase revenue, create loyal customers, grow the company, and increase the awareness of the organisation's brand. News organisations are no different; they search for ways to engage users and create new and innovative methods such as comment sections. However, these opportunities to engage in a public discussion of current events only prove advantageous to journalists and news organisations when the comment section embodies a civil debate. A relationship between the number of comments and the quality is seen. Such news organisations wish to identify innovative ways of moderation that will allow a high quantity of comments and high quality of comments.

1.2.2 News groupings

And lastly, investigating the pitfall of automating moderation is a continuing concern within moderation. Human moderation is still a significant area of interest within the field. Recent trends in human moderators examining literature in hopes of finding new tools to improve systems have led to a proliferation of studies. Highlighting a knowledge gap in the research area that could present a fresh perspective and unravel new mechanisms to moderators would be deemed an advisable approach to remedy this problem. Most studies focus on comment sections and user engagement metrics; however, few studies investigate comments and user engagement metrics conditioned on news groupings. As mentioned in the glossary, news groupings are the different sections a news source is divided into based on familiar topics; the typical news groupings are sports, politics, business, fashion, entertainment, etc. A comparative study of a small number of news groupings and what patterns can be found in their comment sections is an area that has yet to be explored. These patterns are moulded on features that are influential in user engagement.

1.3 Research objectives

The following research objectives were conceived based on the highlighted areas of concern in the preceding section. The primary aim of this study was to draw out influential features in models across diverse news groupings to identify relative patterns. The appeal of examining coefficients relative to news groupings was due to the general lack of research in moderating comments on the premise of a profile made up of assumptions about a news grouping. The objective is to set up moderators to focus on proven characteristics that cause high engagement or low-quality comments. The aim is reached by conducting two experiments and analysing the coefficients of the selected machine learning models. Experiment One will probe whether a comment receives a reply or not, then determine the features that provoke this output. While Experiment Two will scrutinise the components that are determining in predicting the number of recommendations a comment gains. Although the two experiments differ in models and target variables, there is an overlap of reusable dependent variables; therefore, the input across both experiments is fed the same data sets, reducing the time, and making the experiments achievable. Investigating the objective involved a machine learning background and explicit knowledge of the mass communication tool, online news, which was broadened by reviewing the literature. Furthermore, choosing simplistic models with large amounts of existing literature favoured achieving the objective within the time scope.

Another objective the dissertation intended to determine is to unravel a clear systematic review of why the features are significant corresponding to their news grouping, intending to have a deeper understanding of the logical patterns found from the first objective. The objective was achieved by relating findings to the existing literature as the findings were supported by other studies and made interpretable. An overlap existed between Experiment One and Experiment Two in justifying the engineering of features due to this similar nature of their target variables, shaping the objective to be achievable. A pivotal stage in accomplishing this objective was an amplified review of studies, the domain knowledge was heightened, and prompts were gained for the subsequent chapters.

Due to the observation made in section 1.2.1, where news organisations wish to identify ways in moderation that will allow a high quality of comments and a high quantity of comments, which is supported by Ksiazek (2018) study, an objective of the dissertation was established. When examining these discussions across news groupings, it was decided that it would be best to inspect trends in the comment section concerning the level of engagement they encourage. This creates a bridge between content quality and quantity. Whether the engagement metrics indicated low or high levels, both lead to interesting findings on the suspected content quality in which the moderators are interested. The objective was concluded when the models outputted high-performance metrics on unseen test data, demonstrating the features with the most likelihood of spiking engagement metrics. While one engagement feature required engineering, the opposing independent variable was pre-engineered. This made the objective achievable with fewer resources required in the pre-processing setup. The existing literature shows that engagement practices are highly relevant and, therefore, realistic for further study.

1.4 Contributions of research

This dissertation provided a vital opportunity to advance the understanding of moderating content. Human moderators can attain fewer uncivil comments seeping past their systems by applying pre-moderation techniques on comment threads. Moderators can use the framework provided by the dissertation to derive information on comment sections depending on the related matter. Evaluating features of a model, moderators can cross-examine comments that may lead to a highly populated engaged area, which should be moderated due to the larger audience and more significant impact. In contrast, low active areas, possibly signifying lower quality comments, should be moderated. The ability to draw out these comments means moderators bear enhanced tools to curate civil engagement and confront the avalanche of toxic comments or discussions.

Another vital contribution to content moderation is the functionality of highlighting what features motivate higher engagement to make way for the surfacing of the most valuable comments. This can be of interest to the design of the standard function as comments can be ordered from high quality to low quality, a beneficial design opportunity to relieve the stresses of moderation communities. The ability to note valuable comments also proves to help select a comment for the reader's or editor's most favourable feature.

1.5 Dissertation structure

The overall structure of the dissertation takes the form of seven chapters, including this Introduction Chapter. Chapter 2 begins by laying out the theoretical dimensions of the research. The Literature Review is concerned with reviewing online news media, focusing on the growing body of literature surrounding the comment section, and discussing the methods employed by these existing studies. Chapter 3 details the varying design opportunities throughout the project, incorporating studies mentioned in Chapter 2. The main design opportunities were found in preparing, modelling, evaluating, and interpreting the data. Any decisions made throughout the pipeline are further justified in Chapter 4 with a further in-depth discussion on the implementation of the project. The 5th Chapter presents the findings of the research, displaying cross-validation plots, performance metrics, and influential features for both experiments. Chapter 6 analyses the results of the experiments focusing on the coefficient values of the models. Finally, the Afterword gives a summary and critique of the findings, along with identified areas for further research.

Chapter 2

Literature Review

The first part of this literature review focuses on online news media and disinformation. The chapter then discusses the dynamics of the commenting function by delving into the logistics of why the functionality should be facilitated or abandoned, the motives behind commenting, and the impact of low-quality comments seeping past moderation systems and how they are perceived. Numerous studies have also attempted to explain the importance of news groupings, and several accounts are given on politics, a popular news groupings. Lastly, there is a section devoted to exploring key user engagement features and the most popular machine learning methods that have been used frequently in literature in this area.

2.1 Online news media

The first serious discussion and analysis of online users and their online news-reading behaviour emerged in 2000, by which time 20 years had passed since the first newspaper went online in 1980. Chan & Leung (2005) findings suggest an array of different factors that will influence the likelihood of online news adoption. These were the convenience of retrieving articles, up-to-the-minute information, diversity in news content, access to archival stories, and the available multimedia features. With nearly two decades since this study, it is reasonable to assess whether these predicted influences made online news adoption successful. It is clear from recent reports that online news has been successfully adopted and is now the most favourable form of news consumption. Walker & Matsa (2021) reported that about half of Americans get news on social media, at least sometimes, with nearly a third of Americans regularly getting news on Facebook. Although this illustrates the internet as a new focal point for news consumption, this does not account for news organizations. News found through digital outlets is not yet commonplace, with only 21% of U.S. consumers reporting having paid for online news content in the last year.

A possible explanation for this is given by the anchoring effect mentioned by Staff (2020), where a common human tendency is to rely too heavily on the initial piece of information when making decisions. This theory can be mapped to news consumers hesitating to pay for online news, as they are well accustomed to free access to news content.

Despite Chan and Leung's study analysing traditional news media versus online news adoption, it fails to factor in critical elements where the internet has disrupted traditional media. Due to the dawn of a new digital media era, the business models that journalism relies on have inherited several fundamental problems. In order to survive against internet giants and financial realities that came with the new digital era, news organizations have been led to adopt problematic techniques for survival; these are: prioritizing quantity over quality which facilitates the spread of disinformation, running clickbait headlines, and competing against unfiltered social media platforms (Wales & Kopel, 2019). The study conducted by Chan and Leung does not focus on these new means, particularly fake news. In order to capture this, the factor analysis of online news attributes in Chan and Leung's study should pose further questions about the participants' awareness of disinformation; e.g., do users use other online resources to verify a news source a user is reading.

Traditionally, it has been argued that online news media have a heavier emphasis on maximizing user engagement over increasing information quality (Zhou & Zafarani, 2020; Wales & Kopel, 2019). This prioritization sowed the seeds for the cycle of fake news to continue. The cycle is demonstrated in the findings by Avram et al. (2020), where low credibility content with high user engagement results in lower fact-checking and higher liking or sharing from the reader. Disinformation sources tend to follow a network structure that enables fake news to propagate. Zhou and Zafarani's research encourages analyzing these structures to help predict how fake news will spread in the future. The researchers also insist on exploring News groupingss as it captures topics and events that users are most gullible to due to their political biases or preexisting knowledge.

Most of the recent research has focused on disinformation. This dissertation adopts an alternative approach and instead focuses on discovering insights about the network structures of trustworthy news sources. This is done to expose different angles on user engagement varying across news groupings in the hope that these technological innovations can be mapped back to sources of disinformation and mitigate the spread of fake news. An assumption can be made on trustworthy sources, the antithesis of disinformation. Online information perceived to be trustworthy has the network structure of high information quality with moderated user engagement features that do not impede the quality of information, e.g. online sources have been seen to abandon the commenting functions or put in place moderation systems to disallow the flow of disinformation throughout user commentary.

2.2 Comment sections

One of the most significant current discussions in online news media is the trend of online communications. It is apparent from reports conducted by Stroud et al. (2016) that online comments sections are popular, with 55.0% of Americans posting a comment and 77.9% reading the comments at some point. This way of communicating sprung from the traditions of writing letters to the editor, chat rooms, and bulletin boards, with the common theme of the spreading of information and a sense of every voice being heard (Santana, 2011). In his major study, Santana (2011) claims that online forums have the potential to be viewed favourably as a popular means of communication. They may be regarded as the same subset as telephone and e-mail due to their weighted impact on community building while adding additional aspects such as facilitating online collaboration. The findings also proclaim that a new pipeline between readers and journalists has been put in place due to commenting functions, offering journalists a way of staying in touch with readers, in contrast to the traditional way of the letters-to-the-editors section of the printed newspaper. Although these observations break down how user commentary evolved, several questions remain unanswered, such as the argument of turning comments on or off, what drives the public in these online discussions, and the impact of low-quality comments. This section will review these specific aspects surrounding user commentary.

2.2.1 Deciding on comment sections

Over the past century, there has been a dramatic decline in 'social capital', coined by Robert Putman, an American political scientist. Putnam (1995) refers to social capital as the connection among individuals, the social networks, and the rules of reciprocity that form trust. A decline in social capital is seen in lower levels of trust in government and lower levels of civic participation. Putnam argues that technological trends will increase individualism and further undercut the foundation of cooperative relationships found in communities. A significant criticism of Putnam's work is that he focuses on technology as a reason for the decline in social capital; he fails to mention the opposing argument. If media can relate to people as citizens rather than consumers who are valued for their potential participation over demographics consuming various types of products, then mass media can offer much more support for building social capital (on Capitalizing on Social Science et al., 2001).

The majority of discussion forums have standard features, like upvotes and replies. These can facilitate social interaction between readers and help build an online community. They showcase technology nurturing social capital, as collaboration and social interaction between individuals are allowed. Recent studies by J. R. Williams (2019) elaborate on this by concluding that online social networking sites nurture and cultivate types of social capital. This is seen in the New York Times (NYT), where users value the reader's on-topic commentary, criticism, expertise and in-depth discussion through recommending and replying to feedback. Additionally, some users maintain connections and promote their own weblog in their comments, taking advantage of the visibility of the NYT as a public arena (Ruiz et al., 2011). Together, these studies provide important insights into why enabling user commentary is essential for the glue that holds communities together (Putnam, 1995).

Another encouraging reason to facilitate a space for conversation is that commenting functions can foster democratic deliberation, where democratic processes can expand to communication rather than only aggregation and voting. A democracy that facilitates deliberative communication results in knowledge gain, awareness of the reasons behind conflicting views, opinion change, social trust, and, as already mentioned, an increase in the sense of community (Friess & Eilders, 2015). This is a use case for the NYT, as it is considered one of the most influential public opinion setters globally, with well-argued debates that stay on topic and personal or professional experiences that support users' opinions (Ruiz et al., 2011). Although these listed benefits of embracing digital journalism in a more participatory and inclusive form appeal to serve a democratic society, they are met with solid evidence on why comment sections should be turned off.

Traditionally, it has been argued that journalism could be saved through new online concepts while strengthening deliberative democracy employing citizen participation in a constructive and civil debate. However, these positive outlooks on forms of civic participation are now viewed as outdated, with the heightened threat of online communication. Quandt (2018) coined the term 'dark participation' to describe this phenomenon. With the abandoning of commenting functions, we see news sources distancing themselves from the work and resources involved in hosting and moderating user commentary while avoiding the potential negative impact of poor commenting practices (Nelson et al., 2021).

Although comments of a toxic nature can increase user engagement, for example, Masullo Chen & Lu (2017) notes that uncivil comments spark intentions to participate as people want to defend their political viewpoint politically, it regularly results in a misogynistic, racist, sexist turmoil where the more reasoned responses are drowned out by the noise of offensive comments unless heavily monitored. As a result, there is a large and growing body of literature investigating bad commenting practices and their impact. A thematic analysis of comment removal statements such as that conducted by Nelson et al. (2021) has shown that the most common rationale for news organizations that have limited or removed altogether, their comment sections is an effort to reduce incivility and misinformation among user comments. Anderson et al. (2018) found that uncivil comments tend to increase the perceived bias of a news article; it was also found that short remarks increase the polarization that can occur along with existing divisions in society.

In order to combat these detrimental comments, moderation and policy systems have been put in place to promote civil, intelligent conversations. However, with moderating comments, caution must be applied, as an emphasis placed on incivility tends to privilege the groups that have the power to impose their definitions on a conversation, meaning marginalized groups are more likely to be perceived as uncivil (Masullo Chen et al., 2019). News organizations should further explore these and alternative ways to encourage constructive commenting despite the difficulties surrounding moderating. This dissertation can be used to help identify productive user contributions, capturing best practices and collaboration, which in turn can aid user-driven moderation and policies to encourage a more productive commentary.

2.2.2 Motives for commenting

Before examining low-quality comments, a review was undertaken of studies that help understand what drives people to read and write comments. For Springer et al. (2015), the primary motivation for commenters is social interaction, while reading comments is fueled by cognitive and entertainment motives. This view is supported by Stroud et al. (2016) who wrote that 46.2% of users read to learn about the opinions of others, and 40.1% of users read comments to be entertained or amused by other comments. However, they argue that 56.0% of those who comment on the news say they do so to express an opinion or emotion, mainly when the topic is United States politics or domestic policy. The second study fails to emphasize social interaction as the primary motivator and instead focuses on the expression of emotion. Overall, online comments have similar user motives to traditional forms of communication while holding further motives that suggest user commentary satisfies entertainment-related needs.

2.2.3 Low-quality comments

This dissertation investigates the usefulness of analyzing moderated comments; therefore, a presumption can be made that most comments are of civil and consistent quality, with a small percentage of poorly moderated comments of an uncivil nature. This presumption is supported by Etim (2017) who claims the NYT treats reader submissions like the content. This view is supported by J. Salganik & C. Lee (2020) who writes that the NYT fosters a healthy, safe forum for discussion, as its moderation system uses a combination between machine learning and skilled people. It overcomes the issues surrounding an all-human model of moderation, which would limit the number of stories that are open for user commentary, and an all-machine-learning model, which moderates at a standard that does not reach the NYT expectations. However, a small fraction of uncivil comments may seep past the moderation system.

There is a significant presence of research into low-quality comments; this gives enough reasons to review literature analyzing low-quality comments. The low-quality comments that pass the NYT's moderation model are seen to be well written, authoritative, but offensive. J. Liu et al. (2007) observed that a low-quality comment could have incorrect or misleading information when it talks little about a specific concept but more about general topics. One may suppose topics that tend to have readers with solid viewpoints, e.g. politics, could lead to comments of this nature when they lack insightful experiences and target the general landscape of polarised stances against their political parties. S. Williams & Hsieh (2021) maintains that low-quality comments lead to high-quality articles being overlooked. This significant finding illustrates the influence of comments on readers, as comments can overwrite news found in the article and spread disinformation. Introducing moderation systems would reinforce a higher quality of comments; this is the case for the NYT, by which comments require approval for publication.

To better understand the mechanism of a comment section and the effect of low-quality comments Towne et al. (2017), conducted a controlled experiment through Amazon's Mechanical Turk using a platform designed to measure the attention users paid to the comments. A comment containing conflict lowered perceptions of the content quality; this mirrors the claim made by S. Williams & Hsieh (2021) with a heavier emphasis on conflict. What is not yet clear is the impact of reading and judging separate content in quick succession, e.g. different articles and their comments. Although the authors claim the readers' perceptions of an article and comment section would not be affected by reading similar articles and comment sections immediately before, the experiment does not account for the 'Affective Expectation Model.' If the experiment was conducted according to this theory, a lab study is another feasible choice instead of a Mechanical Turk task. Inconsistencies may arise due to variances such as response time. These variances could be associated with an array of reasons which would be impossible to account for unless monitored in a lab.

Wilson et al. (1989) offer this explanatory theory, “Affective Expectation Model.” It refers to readers having a preconceived idea of the level of quality a piece of content should have, and if the reader’s expectations do not match what is read within the content, they disregard the surrounding content, even if it may be accurate. This applies to articles and comments; if a reader can highlight a low-quality comment section, they may neglect reading the article, although it contains real and factual news. This is proven to be true in the studies mentioned by (Towne et al., 2017) and (S. Williams & Hsieh, 2021). Alternatively, the theory can be applied to a group of similar articles and comments; if a reader reads a low-quality article and comment section, they may neglect reading similar articles published by the same news source.

A recommendation to overcome people’s affective expectations is by actively making attempts to create expectations that the public will like when introducing a new product. An association between this suggestion and the NYT Picks feature can be drawn. A reader can view ‘NYT Picks’ or ‘All’ comments; the ‘NYT Picks’ are interesting and thoughtful points carefully selected by the NYT. This feature creates an expectation in the comments section that people will like as it presents the best comments posted.

2.2.4 Conclusion on user commentary

Overall, several studies in this section suggest that the commenting function should be removed. Putman argued that the deterioration of social capital was at the hands of technology. This can be seen across news organizations with no moderation systems with the heightening trend of uncivil commentary. We see the relationship between online communication and Putman’s concept of Social Capital evolve. The recent studies opposing Putman’s views are enough reason to try to reshape the direction of online communication. A popular trend of moderation systems emerges in light of this, where technology is used to create a safe space for social capital to flourish. Together, these studies outline the dynamics of online news discussions and how comment sections are a powerful vector for online communication.

2.3 News groupings

Although there is a large body of literature analyzing user engagement concerning online media, very few studies investigate user engagement based upon what section of the news website the article is published. Despite its importance to user engagement, only four studies could be identified that focus partially or wholly on the relationship between topic and comments. This view is supported by Diakopoulos & Naaman (2011) who wrote, “there is a lack of research into news topicality and user engagement”. Their study also highlights the importance of future work in this area, where investigating these influential features could inform moderators about sections that need additional moderation attention.

2.3.1 Engagement with topics

Most literature referring to the relationships between content and engagement explores three structures of news content where topics can be tagged; these are topics of an individual article, topics covered in the comment section, or topics of a media’s news groupings. Across these structures, topics determine the levels of user engagement where they have the power to cultivate several tens or several thousands of comments (He, Han, et al., 2020). A vital characteristic of the NYT user population with their interest in topics is that they have similar interests but with various focuses; for example, reader A comments on 90 sports articles and ten political articles, and reader B comments on ten sports articles and 90 politic articles (He, Han, et al., 2020). Furthermore, Lagun & Lalmas (2016) demonstrated, by modelling topics covered by a news article, that a clear presence of different media elements affects the engagement of user levels. Diakopoulos and Naaman’s study reveals that topics that are of a more positive tone tend to receive less user commentary, and negative topics have a higher fraction of deleted comments (Diakopoulos & Naaman, 2011). These studies tell us that the topic of an article is influential in the dynamics of user engagement in news comments, revealing there is a strong relationship between the topic of an article and the volume of comments.

Aldous et al. (2019) found that users’ willingness to post a comment varies depending on the topic. Another observation from this study suggests that specific topics generate high volumes of engagement across social media platforms, and on the other hand, some topics generate low engagement across these platforms. Interesting in this finding is that topics and their associated engagement level can vary across different mediums; for future work, it would be worthwhile to interpret topics within articles and their associated engagement level across the news groupings, politics, sports, and magazine. For example, if one of the topics of a sports article is Donald Trump, will this article receive the same volume of comments as a political article with the same tagged topic, Donald Trump?

2.3.2 Groupings layout

Unlike Lagun & Lalmas (2016) exploration of topics of an article, Kim et al. (2021) examine the topics of the comment section with a filtering feature. They highlight that having a feature to filter different topics is preferred by readers. Although this paper focuses on filtering comments, the study's conclusions can be mapped and given context to the filtering of articles based on their news groupings. The study reveals that users favour a filtering mechanism as it helps them gain a quick overview of a diverse topic; it helps them exclude what they do not want to see, and it also helps improve their awareness of the topic. Since 1851, the NYT has followed a structural approach in designing its newspaper, where articles are grouped based on their topics ("The Comment Section", 2002). Overall, this dissertation reminds readers of the importance of grouping similar articles and why news organizations have traditionally followed this structure.

2.3.3 Key sections

This account seeks to explore articles in news groupings: politics, sports, and magazines producing a comparative discussion on these topics and their effects on the amount of participation and the level of interactivity in an article's comments section (Weber, 2014). A common trend between topics, specifically politics and sports, is journalists producing stories on the tension between two actors; rather than reporting on facts, they reproduce sound bites, fostering a polarisation of positions (Ruiz et al., 2011). Certain commenters may take this as an opportunity to lead public discourse into a pit of incivility, where they are willing to defend their viewpoints regardless of the cost. This is supported by Somasundaran & Wiebe (2010), who claim political and ideological debates on hot issues are widespread on the web, where two polarising stances exist. Much of the current literature on news groupings pays particular attention to politics. The political blogosphere of the NYT is mostly balanced, with a 55 to 40 ratio, favouring liberals over conservatives (Adamic & Glance, 2005). This study further reveals that commenters support their viewpoints by criticizing political figures that argue against their stances. This is evident in the NYT commenters with an extensive array of democratic opinions criticizing the republican figure Donald Trump. In the same vein, Diakopoulos & Naaman (2011) points out that politics has a higher proportion of deleted comments indicating that users whom moderators already blocked were attracted to this topic. By drawing on these studies, we are left with a large growing body of evidence of why politics, particularly, receives special attention in research in online news media over other topics. Politics often attracts heavily opinionated, passionate individuals; moderation systems must be more attentive toward political topics to prevent these discussions from turning toxic.

Adamic & Glance (2005) examination of the political blogosphere was centred around the 2004 U.S Presidential election. Since then, the political scene has undergone a great deal of change; this questions whether the findings of this paper are relevant to this dissertation. However, the studies' foundational concepts on the dynamics of political parties and their behaviours on online media remain the same and can be mapped to the 2017 U.S Presidential election. Kim et al. (2021) reported on a procedure of a within-subjects experiment and obtained results this way. The study could have further modelled the observed data to reveal more trends regarding user engagement with and without the filter. Collectively, these studies highlight the need for an investigation into the contrasting news groupings and their relative user participation levels. This was exposed by reviewing three main areas: user-engagement about topics, the importance of media structures based on topics, and the most prominent topic, politics.

2.4 Textual analysis on comments

Turning now to the techniques applied to gain new information about the raw comments and better understand each news grouping. Most recent attention has focused on text analysis for literacy studies as a research method in an attempt to help understand the cultural, social, and historical context of the text (Arya, 2020; Smith, 2017). Machine learning (ML) will be used in conjunction with textual analysis to make the process much faster and more efficient than the manual processing of texts. Arya found that the strength of textual analysis lies in eliciting a multiplicity of meanings, as it leads the researcher to examine all the possible elements of the text, drawing in diversity in that particular domain. Given this, the study outlines the importance of interpreting text and that many explanations should be explored when possible.

2.4.1 Engineering features

A crucial process in the textual analysis and ML pipeline is feature engineering: creating features suited for ML models from raw text data based on the objectives of the task. Natural Language Processing (NLP) analyses the raw comments and extracts features of interest for the models. Prior to engineering features, an initial step was necessary for generating ideas based on intuition and inventing variables that were relevant for the study (Veeramachaneni et al., 2014). Much of the current literature on features in relation to user engagement pays particular attention to features: the number of comments, the number of recommendations, and stylistic and argumentative aspects.

Zong et al. (2017) affirmed the relevance of the number of comments and how they can visually reflect the influence of online news. The study illustrates the extreme difference in quantity distribution of comments concerning their news groupings. It supports this with data stating that military, social, and sports news have more comments as they arouse massive public opinions, whereas financial and economic news generally generate fewer comments. Rather than exploring the popular relationship between the content of an article and the number of correlated comments, the study examines the quantity of comments a topic receives. This metric can be used to analyze and identify patterns across news groupings.

Within the NYT news and comment sections, the recommendation button allows users to express whether they favour a user's comment. As noted by Ksiazek et al. (2016), a strong positive relationship exists between popularity and commenting; the relationship indicates the influence of these metrics on the likelihood of engaging in user-content interactivity. This can be seen in the case of an article with popular content that receives much attention; it can also be applied to user-user interactivity, where a comment of a widespread nature with many recommendations will more than likely attract more users reply. A different perspective has been adopted by Ksiazek et al. (2016) on user-user interactivity, who argues that the negative relationship between popularity and user-user interaction may indicate that negative metrics, such as downvotes, inspire more discussions among commenters due to the perception of polarised tastes. This study highlights the need to downvote features in comment sections due to acting as an informative tool in moderating by possibly indicating low-quality comments or in-depth discussions.

More metrics are revealed by Russo (2020) study, where he examines the trend of emotions that follow low-quality information circulating online sources. The study draws on an extensive range of features to assess the comments, and engineer new features with stylistic aspects, e.g., punctuation marks in a comment, as this reveals the presence of an emphatic element in the text. Overall, the evidence indicates that emphatic features will provide knowledgeable insights; thus, tailored research questions can be derived from this concerning the correlation between user engagement and the empathic tone and the variation of an empathic tone between news groupings.

In contrast to empathic features, features of an argumentation nature hold valuable insights into user engagement. Ruiz et al. (2011) claim that the NYT shows a more significant deal of argumentation, respect among commenters, and diversity of ideas than other online media sources. This study, therefore, affirms that the data obtained from the NYT website will favour the engineering of features that expose the argumentative dynamics, e.g. a feature based on whether a comment receives a reply or not, which demonstrates the interactivity between users. This is supported by Ksiazek et al. (2016) who note that analyzing the conversations would contribute to a richer understanding of interactive engagement practices. Unlike the NYT, many other online news sources tend to be a coherent collective reproduction of the same positions, whereas the NYT platform is a combination of alternative minority perspectives expressed and discussed.

2.4.2 Machine learning techniques

The idea of a problem output not fitting into a yes or no category is assumed in today's world, favouring regression models for more accurate results, as they capture the dynamics of a problem with a broader knowledge of the data. Siirtola & Röning (2020) reiterate this point as they identify results obtained from a regression model to be more desirable than the results from a classification model. This claim is justified by stating that regression models are fed more knowledge by receiving continuous targets, whereas classification models receive discrete targets. Another factor in their results is their area of study, stress, and how it is not a binary problem. Several studies oppose traditional binary thinking; they focus on deconstructivism against structuralism, where the nature of a system is constantly changing, and it is impossible to box a problem into an association of a black and white mindset seen in classification problems. However, traditionally the way of thinking was binary, and it is still heavily present and seen throughout variations of structuralism (Elbow, 1993). Elbow (1993) argues that binary thinking can yield a better model for understanding complex activities. For this paper, to better understand the variation in user engagement across news groupings, questions of a binary and continuous nature are posed.

Supervised classification

Understanding what influences a comment to receive a reply throughout different news groupings can be exposed through modelling a classification problem and analyzing the obtained results. This area will require models to interpret a large number of features, as individual words in comments may act as an influential predictor. Logistic regression, support vector machine (SVM) and k-nearest neighbours (KNN) have been identified as reasonable models to be used in classification tasks, especially text classification; there is also a large volume of published studies that identify using these models for fake news detection (Khan et al., 2021). Kowsari et al. (2019) highlighted limitations of each of these models; for example, the KNN is a data-dependent algorithm. Therefore, it is limited by data storage constraints when finding the nearest neighbours for significant search problems. In addition to this, Kowsari et al. listed optimal improvements to these models that elevate the simplistic nature. Together, these studies outline that starting with simple models and progressively diving into more complex models enables solving problems efficiently through reasoning at the most superficial helpful level. This is supported by Albert Einstein, as he points out: ‘Everything should be made as simple as possible, but not simpler.’ The logistic regression model stands out in this research area due to its simplistic nature. Gharibshah & Zhu (2021) draw attention to the logistic regression model and its scalability; the handling of a large set of features makes logistic regression a favourable model over other classification models. Another remark noted by Gharibshah and Zhu on logistic regression models is the model’s requirement for feature engineering and how it is a disadvantage in contrast to other models. This disadvantage does not appear to be a plausible reason for disregarding the logistic regression model, as feature engineering is simply a way of expressing domain knowledge and does not always have to be viewed as a hindering task.

Q. Liu et al. (2015) suggests that a logistic regression model helps build a powerful tool for explaining the number of comments. Liu and colleagues selected this model due to the dichotomous nature of the independent variables. Although logistic regression is a favourable model, this study’s justification for model selection is irrelevant, as one may assume models can handle categorical and continuous independent variables, with the condition that they are pre-processed and normalised accordingly. Therefore, the dichotomous nature of the independent variables is insignificant; however, what is significant are the dependent variables. The structure around the dependent variable can be a deciding factor in selecting a model, where a model is selected based on the kind of data available for the dependent variable. This is supported by Field (2009) who declares logistic regression should be used to predict a dependent variable, that is, a categorical dichotomy from one or more continuous or categorical independent variables.

Supervised regression

Unlike a logistic regression model, having a categorical dichotomy as the dependent variable violates the assumption of linearity in ordinary regression according to Field (2009). Therefore, linear regression models are preferred in exploring the number of recommendations. Several studies have presented experiments with linear regression models for predicting a numerical output based on a piece of text (Joshi et al., 2010; Alessa et al., 2019; Nguyen et al., 2011). A detailed examination of regression models and their accuracies by He, Shen, et al. (2020) showed that no matter which feature space is considered, the performance of linear regression is considerably worse than the other nonlinear algorithms. These were: random forest (RF), support vector regression (SVR), and neural networks (NN). He and colleagues approved of more complex models over linear regression for their study. Researchers should not rule out linear regression models on nonlinearity problems straight away; after readjusting the linear model to fit a curvilinear relationship between dependent and independent variables, this can be determined. If the adjusted polynomial regression model still under-performs nonlinear regression models, then nonlinear models should be favoured and adopted.

In another study dealing with a nonlinearity relationship between variables, Z. Liu (2017) found that the best result was obtained by Generalized Additive Model (GAM) over linear and lasso regression by comparing the mean squared error. As linear and lasso regression are both linear models, poorer performance is anticipated when engaging with nonlinear data unless readjusting the feature space enhances the model's accuracy. With this common knowledge, the study should have investigated more regression models that stepped away from linear data, having a higher ratio of nonlinear models to linear. This could have enabled a more conversant comparative analysis among models. Unlike Chen et al. (2019) compare a more extensive range of nonlinear models; these include SVR, NN, RF, GAM, and more. GAM computational time is significantly more prominent than the other models and is harder to interpret with multiple predictors; therefore, it is worth analyzing the other suggested models when there are more independent variables and the data is nonlinear.

Chapter 3

Design

This dissertation is formed based on being a helpful resource for researchers in the area of moderation systems; it aims to identify features in the comment section that inform moderators about news groupings that need additional moderation attention. Focusing on features linked to user engagement indicators is reasonable due to the existing literature confirming the potential of these features varying across news groupings. This chapter will precisely establish how the research will be carried out and provide thorough justifications for the approaches taken. The main processes involved are collecting, preparing, modelling, evaluating, and interpreting the data. The following section displays an overview of the proposed system with a brief description; the subsequent sections dive further into each component, explaining and justifying approaches.

3.1 System overview

The below diagram gives an overview of the system.

Stage 1 The source of the data is Kaggle. The data requires preparation before being fed to models; this is a combination of cleaning and pre-processing. Besides the tokens pre-processed from raw comments, features based on domain knowledge are engineered. Once this stage is complete, a CSV file is produced with preexisting and newly engineered features.

Stage 2 The CSV file is split into training, and testing data sets, to provide trained models with unseen data. Cross-validation is performed for the model's hyperparameter. The type of model depends on the nature of the problem; it is trained with the carefully selected hyperparameter.

Stage 3 With the trained model, the results are evaluated against the model's performance on unseen data. Evaluation metrics differ from classification to regression; suitable metrics are used. If several models are evaluated, the best performing one is selected

Stage 4 The model's results are then interpreted. If one model is trained, the corresponding results are interpreted.

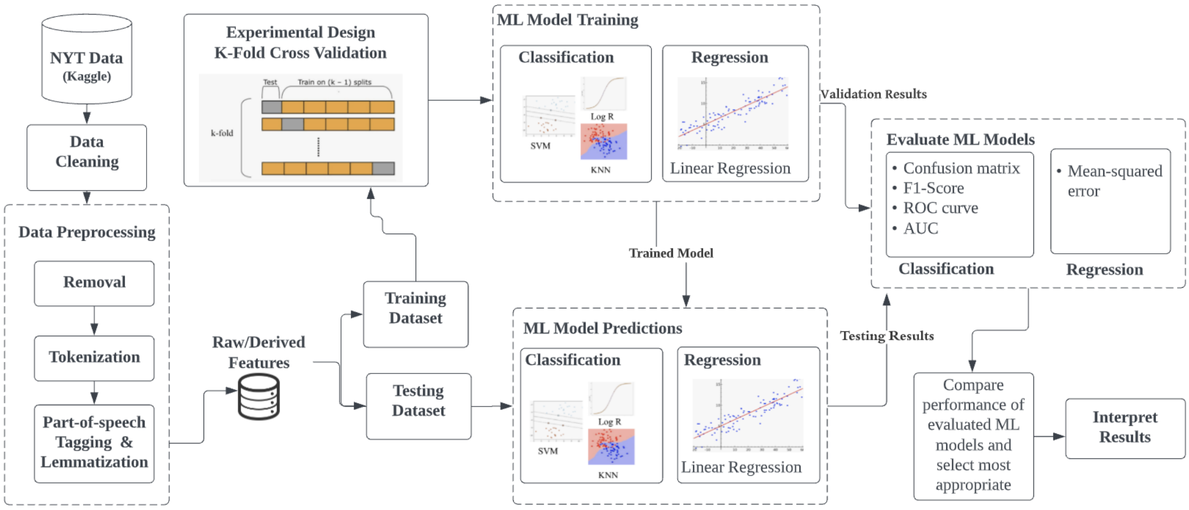

Figure 3.1: A high-level overview of the system

3.2 Data preparation

This is Stage 1, which explores the various components of data preparation. An in-depth discussion of the system's design centred around data collection, data cleaning, pre-processing, and data transformation. The data is carried through each stage, resulting in a set of features fit for a model. Each subsequent section showcases robust decision making with justifications behind each choice.

3.2.1 Data collection

News source

It was decided that the best news source to adopt for this investigation was The New York Times (NYT). This news source was chosen for the following reasons: it produces a broad scope of news coverage on all topics, it offers multiple engagement features, and it attracts an astonishing level of engagement with 150 million monthly global readers ("The

New York Times Company", 2022), in comparison to the Irish Times with 58 million monthly page views ("The Irish Times - Media Solutions", 2022). These advantages create solid building blocks in determining user engagement across news groupings, which is imperative to achieving research objectives.

Although the NYT has a moderated comment section that provides well-informed, civil user commentary; the work does not end there (J. Salganik & C. Lee, 2020). The moderation system requires regular assessment, as low-quality comments can seep past the system. A possible explanation for this might be that the NYT would not offer the flag feature if they were not sceptical of low-quality comments existing, as human moderator decisions can be subjective and could potentially overlook low-quality comments ("The Comment Section", 2002). Therefore, there is an opportunity for models to identify low-quality comments while additionally identifying informative natured comments.

Data set

The community-orientated platform, Kaggle, was explored in an attempt to find a potential data set on user commentary. This online community, with over eight million data scientists and machine learning practitioners, allows users to find and publish data sets ("Kaggle", 2022). Although data quality is not monitored by Kaggle, drawing on the high engagement with experienced Kaggle members on a posted data set, the popularity is seen as an indicator of the data's reliability. With this in mind, a good data set was discovered called 'New York Time Comments,' a data set on comments on articles published in the NYT. Besides receiving over 345 upvotes, another reliable indicator is that this data set was collected with the help of the NYT application programming interface (API). This API is provided by the NYT, illustrating the data's authenticity as it is obtained straight from the news source. The available data set was chosen over interacting with the NYT API directly to save time and resources at this stage in the pipeline and reallocate these resources to more crucial phases, e.g. pre-processing and modelling. Additionally, the decision to cease exploring any comments from years prior to 2017 was down to the restructuring of the moderation system in the NYT at the start of 2017, which allowed more articles to be open for comments (J. Salganik & C. Lee, 2020). This facilitated the gathering of more data, enriching data sets. Alternatively, the reason to ignore more recent years was to avoid the process of gathering additional data through building a web scraper. Although this was not carried out, it is potential for future work.

The data collected on Kaggle contained information about the user commentary from articles published in the NYT from January to May in 2017 and the months January to April in 2018. All months were included in the collection to have enough data for each news grouping. If data from all months were utilised, it would lead to a data set of 2,186,894 comments, which in turn creates a situation with an immense number of computationally expensive features unless considerable computational power is retained, which is not the case for this dissertation. However, as extracting the comments from only the three news groupings was required, this was not an issue.

The necessary content was loaded. The months-wise data was a combination of two CSV files: a CSV file for the articles and a CSV file for the comments. For example, the CSV for the related articles of May 2017 had 996 articles with a total of 16 features, and the CSV for the corresponding comments of May 2017 had 275,493 unique comments with a total of 34 features.

Selecting news groupings

The most purposeful news groupings to examine for this study were: politics, magazine, and sports. This selection was made from a list of 28 different groupings. The selected sections were most eligible due to the following reasons: the requirement of a combination of topics that gave more insights into how engagement metrics vary across topics, the availability of data for a topic such that there were enough comments to accurately train models and the choice of two groupings that contrasted the most popular subject, Politics.

3.2.2 Data cleaning and feature engineering

Data cleaning and feature engineering were imperative time-intensive processes; in-depth accounts of implementation are given in the subsequent chapter. Data cleaning was recommended to achieve a reliable data set; this was done by removing null and redundant values. Features were engineered for predictive modelling to get the most insights out of the available data. Features were designed to represent the underlying problem by transforming raw data into features. This resulted in improved model accuracy on unseen data. Based on the insight into the task and the experience gained from reviewing the literature in Chapter 2, features were brainstormed and created before the raw comments were pre-processed into lists of keyword tokens.

3.2.3 Pre-processing

Pre-processing the raw comments found in the news article is necessary, as, without this, the posted comments will remain highly unstructured, containing redundant and counterproductive information (Bica, 2021). The following design choices were made to address the noisy data set of comments and remove meaningless information.

Lower casing is applied to ignore the presence surrounding the use of uppercase but rather represent the two cases in the same vector space, resulting in a reduced dimensional space. Numbers and URLs are removed for clustering and analysing key phrases as they do not carry importance when extracting keywords. Expanding contractions reduces the dimensionality of the matrix and avoids the splitting of words in the removal of punctuation steps. Additionally, words of length two or less add more noise to the data set than value; therefore, they are removed. This reasoning is similar to justifying the removal of punctuation, and tags, such that this study is not interested in the presence of punctuation within a comment.

After removing punctuation and unnecessary material, tokenisation is applied to get tokens from the comments, followed by part-of-speech (POS) tagged and then lemmatised with WordNet. Lemmatization reduces the high volume of words by reducing the words to an existing word in the language. Lemmatization was chosen over stemming, as it performs morphological analysis of the words. POS tagging is required for lemmatisation and resolving the word to its lemma, such that it transforms the word into a root form. Although this approach is more computationally expensive than stemming, it is worth it in this case as large volumes of keywords are handled, with the need for high quality pre-processed comments, as the focal point of this research is analysing comments. Moreover lastly, irrelevant words, such as stop words, are removed due to not providing valuable information on insights for this analysis. These lists of key tokens and the engineered features were concatenated into one large data frame and saved as a CSV file.

3.2.4 TF-IDF

Within the list of tokens, various words are more informative. Term frequency-inverse document frequency (TF-IDF) is applied to the tokens to identify these critical words and assign further weight. Term frequency (TF) is the standard protocol for text features and is widely used to reveal the more interesting tokens. TF,$tf(t,d)$, is given a token, t, in a document, d, where the document is a sequence of words. It measures how often the token appears in the document. The TF value is more significant when the token occurs multiple times in the document. The document frequency (DF) is the number of documents that contain a token. The inverse of this is the inverse document frequency (IDF):

$$idf(t) = 1 + log\frac{1 + No.of documents in D}{1 + df(t)} \tag{3.1}$$

The IDF is significant when the DF is slight, such that when a token rarely appears in the collection of documents, it is naturally engaging. TF-IDF is then multiplied by the two:

$$tfidf(t, d) = tf(t, d)xidf(t) \tag{3.2}$$

where TF-IDF is of significant value for a token when the token occurs relatively rarely in the comprehensive collection of documents but instead occurs a lot in documents. This is a reasonable, widely used heuristic for identifying informative tokens, and this method is adapted for that reasoning.

3.2.5 Normalising

Normalising handles variables with different ranges of values by scaling the inputs between a particular range. Specific models require normalised data to work. For example, gradient descent in a linear model has numerical problems when the input features are not normalised, as larger values can heavily skew the step size. In this dissertation, numerical values are min-max normalised; this approach transforms values between the range of zero to one by the following equation:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.3}$$

where x represents a single feature vector

3.2.6 Balancing data

A balanced data set is desirable for classification models as it generates higher accuracy in comparison to an unbalanced data set. Therefore, for an imbalance between the two classes, two re-sampling techniques were taken to handle the failure to capture the minority class. Under-sampling deleted samples from the majority class, and over-sampling duplicated samples from the minority. Both techniques were combined to achieve improved performance compared to sampling techniques in isolation (Pykes, 2020). Oversampling was applied to the minority class in moderation, improving the minority's associated bias, while at the same time, bias on the majority was reduced by under-sampling the majority class. A risk of over-fitting or losing valuable information is taken with these techniques. However, achieving the balanced trade-off, these techniques are favourable due to their simple implementation and quick execution time that works well with large complex data sets.

3.3 Data modelling

The following section is Stage 2, it details the machine learning modelling process. Models for both classification and regression problems are explored due to the diverse nature of the experiments. The section heavily concentrates on machine learning techniques that select, adjust, tune, and validate models.

3.3.1 Regression

A large body of prior work uses linear regression with text and non-text features to directly predict a numerical output (Joshi et al., 2010; Gaus and Meng, 2018). Multiple parallel models were not run for the regression experiments based on this and time constraints. Unlike the classification problem, a singular model was selected; this was the linear regression model.

Linear Regression

The following equation is for a linear regression model, based on the input feature x, from which predictions are made. This equation is structured as a line due to the linear nature:

$$prediction : \hat{y} = h_\theta(x) = \theta_0 + \theta_1 x \tag{3.4}$$

where $\hat{y}$ indicates an estimate of the output, θ_0 is the intercept, and θ_1 is the slope of the line. Both θ's are unknown parameters. Initially, the structure is assumed to depend on the model; then, the unknown parameters are discovered by training. In choosing these parameters, the values that make the predictions closest to the training data values are selected. The cost function measures the error between the predicted values and the actual values over all the data points:

$$J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{3.5}$$

The cost function can be used to determine the unknown parameters, based on what parameters minimise the function, as this will result in the predictions with the least squared error.

This dissertation has a linear regression model with multiple variables; however, the principles from a linear regression model with one feature can be carried over. Therefore the model is extended to the following equation based on the number of features:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n = \theta^T x \tag{3.6}$$

$$J(\theta_0, \theta_1, ..., \theta_n) = J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{3.7}$$

where each feature is being multiplied by its corresponding weight.

3.3.2 Classification

A systematic literature review of studies that suggested potential classifying models was carried out. Models were only included in the comparative study if they were easy to implement with a simplistic nature due to delegating time resources to experiments evenly and not exhausting one experiment while the remaining are left undetermined. This compromise was made as to the NYT offer's many user commentary engagement features, e.g. recommendations on comments and replying to other comments; it was worthwhile exploring the maximum number of these in order for a broader perspective to be gained on the commenting function. Based on this and the review of studies, notably Khan et al. (2021) study, three models were selected: logistic regression, support vector machine (SVM) and k-nearest neighbours (KNN). These were trained and evaluated, followed by a comparative study of results, enabling the best performing model to be chosen. The logistic regression was the highest performing model with a large number of features, whereas the KNN and SVM performed well with a small number of features. This aligns with the analysis conducted by Gharibshah & Zhu (2021), favouring logistic regression for its scalability and simplistic nature. Given these statements, logistic regression was selected to model the experiment. As this dissertation's main objective is to draw out influential features across news groupings, putting strenuous efforts into bettering the models is not a priority. Furthermore, as the best performing model is only interpreted, an in-depth account of logistic regression is solely given.

Logistic regression

Similarly to linear regression, logistic regression is formed on a function that maps an input feature vector X to an output prediction; in addition, it is quantised using the sign function to produce -1 or $+1$ labels. The sign function assigns $+1$ when Tx is greater than 0 and -1 otherwise. The following equations illustrate the models form:

$$h_\theta(x) = sign(\theta^T x) \tag{3.8}$$

where X is the feature, θ is the weight, determining the importance of X, and $\theta^T x$ is the weighted combination of the input feature vectors, which can also be written as:

$$\theta^T(x) = \theta_0 x_0 + \theta_1 x_1 + ... + \theta_n x_n \tag{3.9}$$

where n is the number of features, for this study, the total number of features is expected to be substantially high. Therefore reporting all features parameters is unreasonable.

3.3.3 Hyperparameter tuning

Structural aspects of the model are known as hyperparameters, e.g. penalties or the number of features. In contrast to model parameters chosen with optimisation, hyperparameters tend to be chosen more manually, where cross-validation is selected as the tool of choice. This is achieved by scanning across a range of values for the hyperparameter, and then cross-validation is used to estimate the spread of accuracy and variation for each hyperparameter. From this, the best-selected hyperparameter is chosen. To reach a reasonable value, a widespread of hyperparameters with logarithmic jumps are chosen to quickly span an extensive range to give an idea of whether a small or large value is required.

3.3.4 Cross-validation

A hold-out method is taken in cross-validation, where predictions on new data are made to avoid being optimistic about the model's performance. The most straightforward approach to implementing this is splitting the data into two, one for testing and the remaining for training the model. Based on existing studies, the most common ratio is 20% for test data and 80% for training data, which is applied in this dissertation. The trade-off surrounding the data divide is influential, as a steep amount of test data will result in insufficient data for training the model. However, equally, enough test data is required to evaluate the model. Additionally, if the training data is ordered, the test data is taken randomly to avoid bias.

Building on this hold-out method, there is a slight variation known as the conventional way when performing cross-validation, called k-fold cross-validation. This deterministic way implement splits by dividing the data into k equal-sized parts, where the first part is used as the test data and the remaining for the training data, which is again evaluated. Iteratively, the second part is held out as the test data, the remaining as the training data, this is evaluated, and so on. This results in k estimates of the $J(\theta)$, which can be used to work out the mean and the spread of values.

To use k-fold Cross-validation, a value for k has to be selected. The most common choice in similar past models for k is five or ten, as they offer a reasonable compromise. Five folds are used for the two experiments.

3.3.5 Model assessment

After carefully choosing a model and tuning the hyperparameters, the model was assessed on unseen data to minimise generalisation errors. Due to training a model on a specific data set, the model is tuned to that data, and its predictions can lean towards being too optimistic. To avoid being overconfident about the model's accuracy, a train validates test setup is required.

3.3.6 Model selection

The model selection covers the choice of features and the complexity of the model. There are two main ways, sequential model selection, where features are repeatedly added. If they improve performance, this process is continued, and then there is regularisation, which is a reverse version of sequential model selection. Regularisation is the chosen approach, as this method begins with all the features. Similar to the model's structure in this dissertation, the process begins with deleting features and checking for improvements regularly. Regularisation adds a penalty to the cost function in the case of a regression model:

$$J(\theta) = \frac{1}{n}\sum_{n}^{j=1}(h_\theta(x^i) - y^j)^2 + \theta^T\theta/C \tag{3.10}$$

It encourages a simpler model and a selection of improved parameters to manage over-fitting and under-fitting. Cross-validation is applied to determine these penalties; there are two main penalties: L2 penalty and L1 penalty. The L2 penalty assists elements of theta to take small values by penalising large values of theta. It can be combined with any cost function and takes the form:

$$R(\theta) = \theta^T\theta = \sum_{n}^{j=1}\theta_j^2 \tag{3.11}$$

The other penalty, the L1 penalty, is the sum of the absolute values of the parameters:

$$R(\theta) = \theta^T \theta = \sum_{n}^{j=1} |\theta_j| \tag{3.12}$$

The absolute value is taken to avoid the cancellation of positive and negative values. The L1 penalty penalises larger values of parameters less severely than the L2 penalty; it also encourages sparsity of a solution, where a high number of parameter values are set to zero. The L2 penalty is considered more suitable for this research due to accommodating the visualisation of the model in a way that exhibits the entire distribution, so the magnitude of the influential features can be compared to the features of less importance and understood better.

3.4 Evaluating model performance

This section represents Stage 3. It is necessary to be clear about the overall purpose of this dissertation and how to measure how well these proposed machine learning systems meet the overall objectives. There are multiple metrics of interest for both regression and classification problems. Goodhart's Law states that a measure becomes a target; it ceases to be a good measure; for example, suppose the time between an article published and a comment was posted was used as a metric for user engagement; this can be reduced by posting it at a time when most users are active, or the contents of the article are viewed as a popular topic, this favours easy cases and focuses on the value of this variable as an indicator. When selecting suitable metrics, it was essential to be mindful of Goodhart's Law to avoid losing sight of what outcome matters. Baselines were established for each regression and classification problem, where the performance of existing solutions was examined, along with similar problems and how they were solved.

3.4.1 Baseline

Baselines are introduced for a sanity check to know if something nontrivial was achieved. A dummy baseline model tells whether the new models are doing anything of value; e.g., if the MSE of the dummy model s is lower than the regression model, the conclusion can be made that something trivial is being achieved. For this paper, selecting a constant baseline model is a reasonable choice. This engineering judgement call was made on the lack of current statistics on predicting user engagement in the comments in the NYT across the news groupings, which is an interest of this dissertation. Therefore, there are no baselines to compare against, and a constant baseline can be deemed suitable.

3.4.2 Regression metrics

In exploring the regression problem, the choices of metrics were: mean square error(MSE), root mean square error(RMSE), mean absolute error, and R^2. The mean square error was selected for model tuning using cross-validation due to it dominating the listed options based on the existing body of knowledge. Furthermore, no notable differences were highlighted when comparing metrics that favoured the choices. The mean square error is the sum of the error of all data points divided by the number of data points; the square root of the result can be taken to adjust the result to be the same as the values, which will be easier to interpret. These two metrics take the following form:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\theta^T x^i - y^i)^2 \tag{3.13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta^T x^i - y^i)^2} \tag{3.14}$$

3.4.3 Classification metrics

Classification metrics are not as relatively straightforward as regression. In predicting the label, +1 or −1, if the prediction is correct, a true positive for the plus label is assigned, or a true negative for the minus label. If the prediction is incorrect, then a mistake occurred. This gives a false positive, where +1 was predicted but the correct label was −1, and vice versa for false negatives. The classification problem in this dissertation, whether a comment receives a reply or not, is investigated; if a comment receives a reply and the comment is marked as receiving a reply, a true positive is given. A true negative is given if a comment does not receive a reply and is marked as not receiving a reply. In contrast, if a comment is marked as receiving a reply, but it does not, a false positive is given, and if a comment is marked as not receiving a reply, but it does receive a reply, a false negative is given. These different outcomes have different consequences; they form a trade-off between true and false negatives. Therefore, a balance is required, where the different errors are given different weights and consequences. Finding different ways to balance these is why evaluating the performance of classification is harder than regression.

A tool that has to be used with caution in classification problems is accuracy, where the number of correct predictions is counted over the total number of instances. Due to the data being initially imbalanced, where most comments do not receive a reply, accuracy is not a sensible metric. Instead of accuracy, a handful of tools are available which are much more robust.

Confusion matrix

A Confusion Matrix is highly informative, as it captures all aspects and notes the four possible outcomes in a table in the following form:

	Predicted Positive	**Predicted Negative**
True Positive	TP	FN
True Negative	FP	TN

Table 3.1: Confusion maxtrix structure

In the confusion matrix, the diagonals are the true positives and true negatives, the values that the model obtained correctly, whereas the off diagonals represent the errors. From this table, other measures besides accuracy can be calculated: true positive rate, also known as recall, or false positive rate, also known as specificity. The recall is the number of $+1$ labels predicted correctly over the total number of positives, written as:

$$\frac{TP}{TP + FN} \tag{3.15}$$

Similar to recall, specificity is the number of mistakes made, FP, over the number of negative examples, written as:

$$\frac{FP}{TN + FP} \tag{3.16}$$

Additionally, precision is another means to summarise the confusion matrix, giving the number of positive predictions which are correct; it takes the form:

$$\frac{TP}{TP + FP} \tag{3.17}$$

One limitation of the confusion matrix is the ability to not plot its values by cross-validation, unlike the mean squared error in regression. To overcome this, a single value is required; one option is the f1-score, a combination of false positives and false negatives, which is seen below:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.18}$$

ROC curve

The ROC curve varies the threshold, beta, between the range 0 to 1, the values accepted by the confidence function. The true and false favourable rates are then calculated. By varying beta, several points have produced that form the curve. An ideal classifier has a curve that reaches the left-hand corner where no false positives are present but

100% true positives. Based on the curve, a threshold can be chosen that represents the point closest to the coordinates (1,0); this will give a reasonable compromise. A random classifier runs diagonally from (0,0) to (1,1) and shows a balance between the positive and negative labels; the formed ROC curve should steer away from this random classifier in the plot. If the model's curve drops below the random classifier, the resulting performance is unreasonable, and the outcomes should be flipped, where if it predicts -1, +1 should be reassigned and vice versa. Similar to the confusion matrix, the ROC curve is difficult to plot for cross-validation, so one value is desired instead. The standard way to get one value from a ROC is a metric called the area under the ROC curve (AUC). A desirable AUC value is between 0.5 and 1, with a value closer to 1 as this represents an area with a higher number of true positives.

3.5 Experiment layout

Several experimental hypotheses were derived based on domain expertise, the preliminary research undertaken, and the observations made from the exploratory data analysis (EDA). These experiments aimed to determine offbeat revelations in hopes that arranging these experiments based on news groupings would unravel new information to moderators, making the moderation systems more resilient. The two sets of experiments focused particularly on unravelling user engagement trends across news groupings, and both experiments aim to explore clear indicators of user engagement in comment sections. Although many studies in online news have focused on user engagement practices, researchers have not examined engagement features across news groupings. The number of experiments fluctuated throughout the project timeline, initially beginning with ten, then three, and finally settling at two. Ten experiments were reduced based on time constraints and the irrelevance between the purpose of the experiment and the research objectives. Both experiments consisted of the pre-processed data, feeding the data to the most reasonable model to achieve favourable performance, then interpreting the resulting coefficients in the model and drawing conclusions on these with the posed experiment.

Experiment	**Description**
1	Predicting whether a comment will receive a reply
2	Predicting the number of up-votes a comment receives

Table 3.2: List of experiments

Experiment One was conducted by investigating the relationship between whether a comment receives a reply based on the varying aspects found in a comment. This setup seemed to be a good indication of the interactivity between commenters. Conclusions from this experiment can prove to be an effective tool for moderation systems. One possible explanation for this is that moderation systems can target comments that point to high interactivity, as perhaps the comment may dwell on offensive content infuriating the community and causing them to take a stance. Another possible explanation would be a comment that interacts with more people should automatically be examined, as it has a higher likelihood of reaching more people, and thus a higher risk on more commenters if it has a toxic nature.

Experiment Two was carried out in an attempt to identify what elements make comments highly favoured. This experiment proved to be a reliable inkling of popular comments among the community. Identifying popular comments for moderation systems would be advantageous, based on a similar justification given for Experiment One, where comments with a high number of recommendations have a higher likelihood of reaching more people and thus a higher risk on more commenters if it has a toxic nature.

For future work, a possible third experiment could be carried out by analysing the overlap between the two experiments to understand better the possible reasons for a comment receiving a reply. Notably, a comment that receives a reply and a high number of recommendations are more likely to have the commenters agree, whereas a comment with no recommendations and a large quantity of comments shows more signs of commenters disagreeing with the content.

3.6 Result analysis

This is the last section, Stage 4. As the primary objective of the research was to draw out influential features in models across diverse news groupings, the results must be analysed efficiently. All features reassembled a coefficient value; this signified the weight that feature has on the model, as mentioned in X. Interpreting the features of importance in the models was considered the most reasonable way due to the in-depth findings this method would make available, while also requiring a low amount of implementation. In addition, the distribution of coefficient values was visualised in an attempt to understand where influential features lie among the remaining coefficients.

3.7 Limitations of the design

Several limitations in this design need to be acknowledged. A lack of data existed across most news groupings; for example, the news grouping 'Art' consisted of only 4000 comments compared to 'Sports' with 20000. A large amount of data is required through the copious amount of training necessary for machine learning models. Although the selected news groupings possessed considerably more data, investigating other news groupings could potentially produce significant insights. Web scraping tools may overcome this to gather additional data for the appropriate news groupings. Secondly, in hopes of moderation systems using this dissertation to help data-driven decision-making, it would be more of a benchmark study if the data were more recent to capture patterns or correlations of today's world. This limitation is due to the design of the method used to collect the data. The measured variables were gathered in the years 2017 and 2018, very soon after Trump's presidential election in November 2016. This may lead to highly saturating the data in Trump-related news and scandals. Therefore, further data collection is regarded to determine the impact of news groupings in more recent years. In the same vein as the previously mentioned limitation, implementing a web-scraping component using the available NYT's APIs would overcome this. Moreover thirdly, despite the existing literature favouring more complex algorithms, the study evaluated simplistic models only. This was due to finding the quickest path to getting the pipeline working end-to-end. This resulted in very few optimal solutions throughout the system; therefore, future work would be building on the simple models and seeking the best possible solution.

Chapter 4

Implementation

In this chapter, the implementation of the system for drawing out patterns across news groupings is discussed. It intends to determine how the research objectives were achieved. As Chapter 4 gives an in-depth detail of all dimensions of the research, the Implementation Chapter focuses on the vital methodology that arose from these design choices. The overall structure of this chapter takes a similar form as Chapter 3, with a focus on the essential details.

4.1 Data handling

As data gathering was not an intensive resourceful process, due to the data being pre-collected, not many implementations were necessary for Stage 1. It is important to note that the data was not stored in a database because the amount of data was manageable through CSV handling. However, if moderators planned to apply this research to more recent content with a web-scraping tool, a database would integrate well as the scale of the data would proliferate, and in which a robust database system can help manage the increase in data. The only substantial work in handling the data before pre-processing was the manipulation of the original datasets to the news grouping form. This required the separation of news groupings of interest. The two original datasets, comments, and articles each produced three separate datasets conditioned on the pre-engineered feature, 'News Desk'. This was used to filter news groupings, which helped achieve a primary objective.

Overall, the number of comments for each news grouping was around 20,000. The sports grouping contained just over 20,000 comments, whereas the remaining two news groupings contained more. To level out the data sets, 20,000 comments were sampled. This meant each model had a similar input size of data, and therefore, no model could learn more about the behaviour than the others. Additionally, 20,000 comments were enough data to train the models effectively.

4.2 Feature engineering

Table 4.1 shows the features engineered, and how they were implemented:

Features	***Implementation***
article_word_count	Existing variable in the data set
recommendations	Existing variable in the data set (number of likes on a comment)
comment_word_count	The built-in function split() was used to spilt the string into a list, the number of elements were then counted
start_question	Used startWith() method of String class to check the prefix of the comment and whether it starts with a typical start word
question_mark	Used the 'in' operator in python to check if the comment contained a '?'
exclamation_mark	Repeated same action as question_mark
sentiment	Applied the TextBlob library to return the subjectivity of the comment
pub_length	Used the datetime library to calculate the difference between the approved date of the article and the date the comment was posted
get_reply	Used the depth feature to find comments with the value 2.0, indicating it is a reply to a comment. Then search for the parent_id of this comment
comment	The tokens extracted from the posted comment with nltk tokenizer

Table 4.1: Independent features for Experiment One and Experiment Two

4.3 Cleaning

An in-depth EDA was performed to enable the unmasking of the data set's characteristics in cleaning the datasets. Out of the 34 original features of the comments data sets, 21 of these were redundant for this study. This was also the case for articles, with nine redundant features out of 16. The initial step was deleting features with a dramatic quantity of null values, given they were irrelevant to the study, to avoid the adverse effect they would have on models' performance and accuracy. For features with a low number of null entries, rather than deleting the whole feature, the null values were filled with the average of that particular independent variable, which avoided the significant loss of data. Secondly, several columns were removed by considering the association between the feature and the dependent variable, noting the values that were not important in this study's evaluation. Another step in the EDA was the removal of outliers to avoid decreasing statistical power, as outliers can increase the variability of the data. The approach to highlighting outliers was visualising the data with scatterplots and box plots. The last step of the EDA involved analysing relationships between variables. This allowed us to quickly understand the relationship between the two variables and point out high correlations, which must be accounted for in the models.

4.4 Pre-processing

To determine whether words and comments tend to influence engagement practices, the comments were pre-processed and transformed into text features. Once the data was collected and cleaned, the libraries, Natural Language Toolkit (NLTK), and contradictions were imported to pre-processing. The key stages of carrying out pre-processing are as follows:

Lemmatization

Lemmatization downsized the high volume of words by reducing words to an existing words in the language. Lemmatization was chosen over stemming because it performs morphological analysis of the words. The imported library NLTK has a built-in lemmatiser. NLTK was chosen over the spaCy library because spaCy only handles the English language. This implementation would be more valuable if it could be mapped over to other languages.

Lowercase

In python, Strings have a method, lower(), to convert text to lowercase. Lowercase is applied, as attention is not being paid to the intensities of presence surrounding the use of uppercase. Therefore, the two cases are represented in the same vector space resulting in a reduced dimensional space. A possible future work would be, exploring the level of enthusiasm implied by using uppercase letters in the comments.

Punctuation

Stripping punctuation is necessary, as the presence of punctuation is not of interest in this dissertation. A string was defined as all the punctuation marks that should be removed. A brute force approach was applied by iterating through a comment, checking each character against the defined punctuation dictionary, and then re-building the string after the punctuation was removed. This approach is the most simplistic and avoids additional libraries.

Contractions

Expanding contractions reduces the dimensionality of the matrix, which makes the process less computationally expensive, such that it accounts for "we are" rather than "we're" and "we are." Furthermore, expanding contractions will remove the apostrophe, which will not be removed in the punctuation step, causing issues. The contradictions library was imported. Although this library does not take into account the ambiguity of the words in comparison to pycontractions, such that "ain't" only has one meaning which is "are not," rather than "am not / are not / is not / have not." These contradictions are not viewed as highly important when analysing keywords. Therefore, using the less computationally expensive library with lower accuracy, which works in most cases, is acceptable.

4.5 Machine learning

4.5.1 Cross-validation

For each k-fold, the training data fit the model, and the remaining test data was used for evaluation. The evaluation metrics were noted, and the overall skill of the model was summarised based on the cumulated evaluation scores. The means and prediction errors are plotted for the range of hyperparameters, revealing how well the model performs on unseen test data. From cross-validation plots, the most optimal hyperparameter is selected. A reasonable hyperparameter is chosen based on a small cost close to zero to avoid a too strong penalty, conveyed in the figure by smaller error bars. Furthermore, the simplest model is ideally chosen to avoid over-fitting, selected by assigning the smallest hyperparameter value when possible. The cause of over-fitting is due to complex model fitting noise in the training data, resulting in poorer predictions on unseen data. A model is under-fitting when its nature is too simplistic, failing to capture the behaviour of the data, resulting in poor predictions. The objective of cross-validation is to reach a reasonable hyperparameter demonstrating the right balance between under-fitting and over-fitting.

K-Folds

If the k-value is very large, e.g. 90, the accuracy will be estimated by a small number of data points, e.g., two or three, where noise has a much more significant impact. For example, if one data point is considered noise, then the data spread among the other two data points will be negatively impacted, creating estimates that fluctuate between each split. Therefore, a smaller k considers more data points that average out the noise leading to more minor fluctuations and revealing the fundamental behaviour of the model. Five folds are chosen over ten folds, as ten folds will increase the computational time to double the amount; therefore, keeping k to a minimum is desirable.

Performing cross-validation on C

The logistic and linear regression models have an associated hyperparameter, C. As mentioned in 3.3.6, an additional term, the L2 penalty, is added to the linear function for regularisation. C represents the strength of the penalty. To span a wide range of potential values, as described in X, the hyperparameter range rises in multiples of ten, starting at 0.001 up to 1000. In all cross-validation figures, a logarithm spread was applied to the x-axis.

4.5.2 Other ML techniques

Balancing the datasets was coded to reduce the severe skew in class distribution for the classification problem, where the minority class was a comment receiving a reply. The framework imbalanced-learn was leveraged to combine both random sampling techniques. An over-sampling sample strategy of 0.5 was chosen, while a sampling strategy of 0.8 was selected for under-sampling. This resulted in a modest amount of under-sampling of the majority class, reducing the bias of the majority and oversampling the minority class to improve the bias of the minority.

The best modelling library adopted for this investigation was scikit-learn. This python library has all the mainstream algorithms for machine learning tasks essential in this dissertation. Due to not requiring complex and production scale models found in sophisticated libraries such as TensorFlow or PyTorch, scikit-learn was favoured over these. The data was normalised using the MinMaxScaler functionality from the scikit-learn library.

Although many studies addressed different normalisation techniques, MinMaxScaler was simplistic in implementing them. Future work would be attempting to normalise with StandardScaler, as it does not suppress the effect of outliers.

As mentioned in 3.3.5, to minimise generalisation errors, the dataset must be split into training and testing. The train_test_spilt() function was used within the scikit-learn. This involves keeping a percentage of the data, e.g. 20%, to one side left untouched. This is not used in choosing the model. Although this can be an expensive approach, this is not the case for this dissertation if there is not a large amount of data. The remaining is the training data, e.g. 80%; this is then split into training and testing for when executing k-folds, cross-validation, and the validation data. Therefore, when the final model is reached, the test data is available to act as unseen data; this will give a more realistic idea of the model's performance. Based on this domain knowledge and similar studies, the train validates test setup was fixed to a test size of 20%, using the parameter test_size=0.20.

4.6 Evaluation metrics

After the model was fitted, the performance was evaluated with several metrics extracted from the scikit-learn library. One of the contributing factors in selecting the scikit-learn library was the availability of the required metrics; these were: mean_squared_error, accuracy_score, f1_score, recall_score, precision_score, roc_curve, roc_auc_score, and confusion_matrix. Incorporating these metrics made the evaluating process effortless, as no scores required in-depth calculations.

As mentioned in 3.4.1, a constant baseline was deemed a suitable choice, as there was a lack of statistics on predicting user engagement in the comments in the NYT across the news groupings in which is an interest of this dissertation. The DummyClassifier was adopted from the scikit-learn library. The strategy selected was 'most_frequent', which returns the most frequent class label in the target variable. As the imbalanced class issue was resolved, this strategy is appropriate due to no existing bias towards the majority class.

4.7 Results analysis

Once the best performing model had been decided upon, the coefficient values of the positive and negative influential features were obtained. To display a comparison plot across news groupings, the top 15 coefficients from each positive and negative set of features were selected. This number enabled the full display of coefficients without the figure becoming unreadable. Moreover, tables are given with the top 25 positive and negative coefficients of text features, along with the coefficient values of the engineered features. Presenting the results by this means displays a broader range and allows the reader to grasp a deeper understanding of the significance of a coefficient compared to the remaining.

The code in listing 4.1 is an approach to determining the significant coefficients of a model given. A data frame is formed with the use of the panda's library. It is shaped by all the model coefficient values and the related tokens. The top 15 and 25 positives are derived from sorting the data frame in descending order and slicing it down to the first 15 values. This is repeated for the negative coefficients, but with the data frames sorted in ascending order.

```
coef = pd.DataFrame(model.coef_.T, feat, columns=['coef'])
coef_p = coef.sort_values(by='coef', ascending=False).head(15)
coef_n = coef.sort_values(by='coef', ascending=True).head(15)
```

Listing 4.1: Code for determining the influential coefficients

Chapter 5

Evaluation

This chapter presents and describes the experimental results systematically and comprehensively. The results were produced by two major experiments, Experiment One and Experiment Two. The first section gives detailed accounts of the experiments and their setups; the subsequent section presents experimental results.

5.1 Experiments

The experiments were designed to derive informative features from different comment sections. They aim to expose characteristics of a comment that influence user engagement features and vary across news groupings. Experiment One does this by identifying what features lead to interactivity between two users across news groupings, where the user engagement feature is whether a comment receives a reply. Experiment Two achieves this aim by identifying features that lead comments to be considered valuable or not by other readers, where the user engagement feature is the number of recommendations a comment receives. The experiments are listed in the following table:

Experiment	**Description**
1	Predicting whether a comment will receive a reply
2	Predicting the number of up-votes a comment receives

Table 5.1: List of experiments

Both experiments have dependent variables, the conditions measured in the experiments, that translate into user engagement features. For Experiment One, the engineered feature, get_reply, is tested, while for Experiment Two the preexisting feature, recommendations is measured. The tables 5.2 and 5.3 show example values of features along with descriptions;

the state of these values is after the pre-processing stage but before normalisation to make the values comprehensible.

Features	***Description with examples***
get_reply	1 if the comment receives a reply, otherwise 0 - e.g. 0
recommendations	The number of likes a comment receives from other reader - e.g. 100

Table 5.2: Independent features for Experiment One and Experiment Two

The independent variables, the values attempting to explain the output, are identical for Experiment One and Experiment Two. Table 5.3 shows an overview of the features and example data. Note that each token accounts for a feature; however, reporting all tokens is unreasonable due to the sheer amount. Therefore the comment feature represents them as a whole.

Features	***Description with examples***
article_word_count	Count of words in the article - 1200
recommendations	The number of likes a comment receives from other reader - e.g. 100
comment_word_count	The number of words in the posted comment - e.g. 50
start_question	1 value if the comment starts with a question word, a 0 otherwise - e.g. 1
question_mark	1 if the comment contains a question mark, otherwise 0 - e.g. 1
exclamation_mark	Same as the previous feature, but for exclamation mark - e.g. 0
sentiment	A float between the range [0.0, 1.0], where 0.0 means the comment is very objective and 1.0 is very subjective - e.g. 0.5
comment	The tokens extracted from the posted comment - e.g. ['great','article']

Table 5.3: Independent features for Experiment One and Experiment Two

5.2 Results

An identical setup of the cross-validation components, range, and graph was extended to all news groupings in both experiments. All the cross-validation figures virtually possessed interchangeable characteristics. The optimal hyperparameter values remained constant throughout news groupings. This carried on into the model's performance and the ROC curves for each news grouping, where there was a standard census across all news groupings. Therefore, the following section only presents the results for the news grouping, sports. The remaining results are found in the Appendix Chapter. Displaying the magazine and politics news grouping only repeated the same hyperparameters, metrics, and justifications.

5.2.1 Experiment One

Cross-validation

In figure 5.1, the initial cost was distant from the y value, zero, with the cost centring around 0.32. This space closed around the $log_{10}(10)$ mark, with a newly centred cost value of 0.18. The remaining C values linger around a similar mark. Using fivefold cross-validation, specified in 4.09, ten was selected as the most reasonable value for C. The value, ten, resulted in the simplest model with a small cost range closest to zero.

After performing cross-validation on the logistic regression model and noting the most appropriate C value, then, the subsequent model, the SVM, continues the process of hyperparameter tuning. In figure 5.2, the majority of prediction errors obtain a small spread. A clear exception is seen for the C value, 0.001, with a higher cost value, making it an inadvisable option compared to the remaining. The most optimal C value was one where the lowest cost was achieved for nearly all gamma values along with small reasonable variances. The optimal value is translated to one for the second hyperparameter, γ. The value generated the simplest model lessening the computational time while producing a small variance and a prediction error close to a zero value on the y-axis.

The last model, KNN, followed a similar pattern to the previous models, where news groupings produced comparable cross-validation figures. As seen in the plot 5.3, with the increase in the number of neighbours in the KNN model, the cost values move further from zero. Based on all three figures, one was the most reasonable choice for the hyperparameter. This decision was made based on the prediction error being closest to zero, the simplest model is generated, and the relatively small variance.

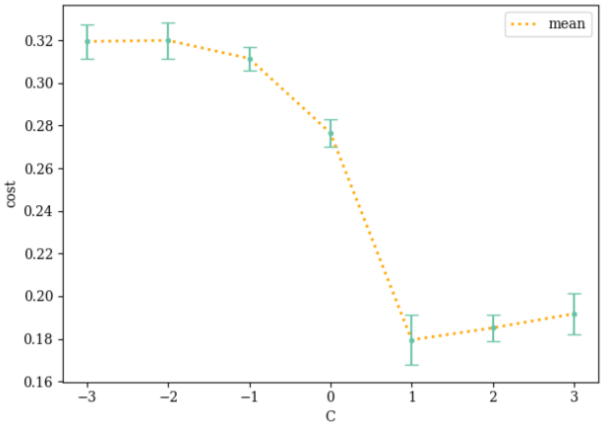

Figure 5.1: Logistic reg. sports model with a varying C

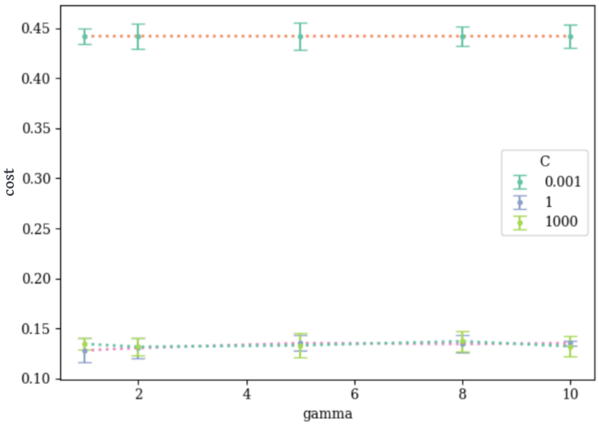

Figure 5.2: SVM sports model with a varying C, and γ

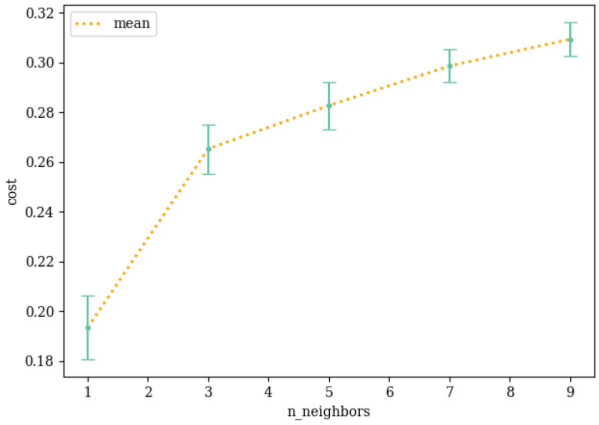

Figure 5.3: KNN sports model with a varying neighbours

Models performance

On average, logistic regression was shown to have out-performed the other models. As shown in Table 5.4, the results indicate that the logistic regression performs slightly better than the SVM and KNN. The next chapter discusses these tables in further detail with justifiable explanations on why selecting the logistic regression model was suitable.

	Logistic Reg.		**K-Neighbours**		**SVM**		**Baseline**	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**
Precision	0.7779	1.0	0.7580	1.0	0.7574	0.993	0.0	0.0
Recall	0.8990	1.0	0.9053	1.0	0.8003	0.9980	0.0	0.0
Accuracy	0.8276	1.0	0.8147	1.0	0.9056	0.9964	0.5459	0.5579
F1-Score	0.8313	1.0	0.82653	1.0	0.7783	0.9960	0.0	0.0

Table 5.4: Performance metrics for sport models

Confusion matrices

	Logistic Reg.	**K-Neighbours**	**SVM**	**Baseline**
Test	$\begin{bmatrix} 1105 & 0 \\ 66 & 852 \end{bmatrix}$	$\begin{bmatrix} 898 & 207 \\ 87 & 832 \end{bmatrix}$	$\begin{bmatrix} 858 & 247 \\ 102 & 817 \end{bmatrix}$	$\begin{bmatrix} 1105 & 0 \\ 919 & 0 \end{bmatrix}$
Train	$\begin{bmatrix} 4516 & 0 \\ 0 & 3578 \end{bmatrix}$	$\begin{bmatrix} 4516 & 0 \\ 0 & 3578 \end{bmatrix}$	$\begin{bmatrix} 4494 & 22 \\ 7 & 3571 \end{bmatrix}$	$\begin{bmatrix} 4516 & 0 \\ 3578 & 0 \end{bmatrix}$

Table 5.5: Confusion matrix values for sport models

As mentioned in 3.4.3, the diagonals of the confusion matrix are the true positives and true negatives. Table 5.5 presents the breakdown of confusion matrices for each of the three models. From the data in the table, the logistic regression and KNN models appear to have the highest number of true negatives, 4516, for the training set. However, logistic regression has the superior number of true negatives for the test set, 1105. This means logistic regression has a greater likelihood of predicting correctly if a comment does not receive a reply on unseen data, and the KNN and logistic regression have more probability of predicting a comment not receiving a reply on the seen data. A similar presence of models can be mapped to the true positives for the unseen and seen data, where the same conclusions are drawn.

ROC curve and AUC scores

Figure 5.4 provides the ROC curves for the baseline and the three models. In the figure, the dashed turquoise line represents the performance of a random classifier, and it can be thus seen that the baseline classifier lies upon this line. As mentioned in Y, a model with good performance displays a ROC curve that reaches the top left corner of the grid. This is the case for all three models as the coloured lines are drawn out to this corner. The top left corner can be explained by a model achieving no false positives and all true positives, an ideal performance for a model where the model predicts correctly on the unseen data. The reasonable choice of a classifier based on this figure alone would be logistic regression, as the model is seen to reach furthest to the (0,1) for most of the plot.

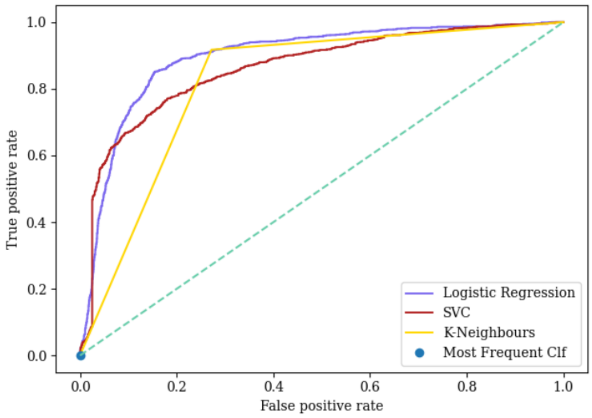

Figure 5.4: ROC curves for the sports models

	Logistic Reg.	**K-Neighbours**	**SVM**
AUC Score	0.8959	0.8227	0.86217

Table 5.6: AUC Scores for the sport models

As specified in section 3.4.3, how thoroughly a model can distinguish between the two classes, positive and negative, can be betrayed by AUC score. From the data in table 5.6, it its apparent the logistic regression has the leading AUC score.

Overall, from the results and the existing literature, logistic regression was finalised as the ultimate model. It was apparent from the performance metrics, confusion matrices, and ROC curves that the logistics somewhat exceeded the other two models. Another important observation made in Chapter 2 was that those regression models tend to gain particular interest for their scalability. This is a favourable attribute, as if the dissertation were to process an excessive amount of data in the future, the logistic model would adapt to this situation effectively while still outputting reasonable performance metrics.

Features and coefficients

Reporting on all features is illogical due to the vast amount of features the problem is modelled on; therefore, the following figures 5.5 and 5.6 show the top 15 influential positive coefficients and negative coefficients for all news groupings. The figure is the first significant plot that permits a clear comparison across the three news groupings and thus provides essential insights across similar and differentiating trends that moderators could potentially review. Figure 5.5 displays the most influential features that cause a comment to receive a reply. Strong evidence of the recommendations feature is found across all news groupings; the remaining displayed features are all of a textual nature with no overlap. These features are examined and discussed thoroughly in Chapter 6.

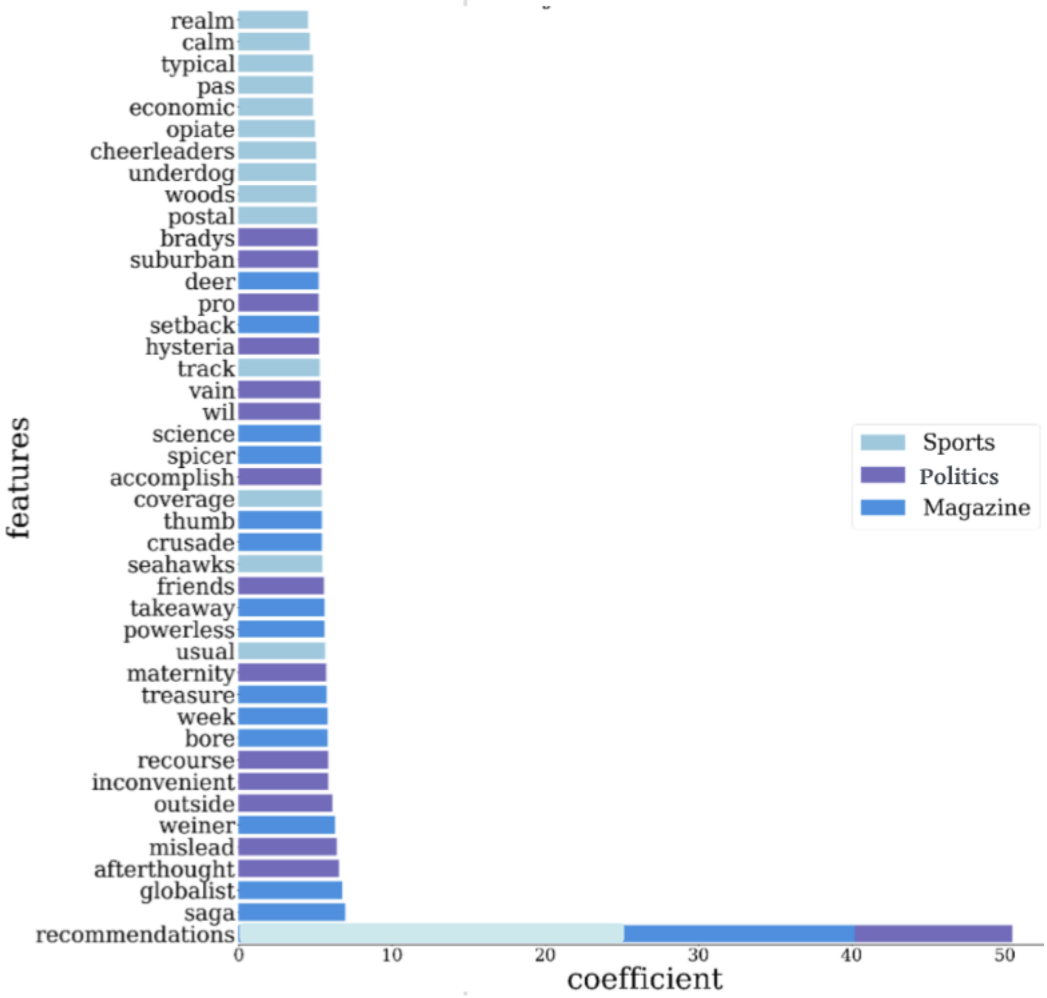

Figure 5.5: Most influential positive features of the logistic regression clf.

If we now turn to the feature that illustrates the influential features that cause a comment not to receive a reply. A negative correlation was found between when a comment was posted and when the article was published for both the politics and sports news groupings. The remaining features were textual, like the figure 5.6, with most of the magazine features having the greatest influence on a comment not receiving a reply. Moderators could examine comments with these indicators in hopes of locating low-quality comments. These are explored further in the subsequent chapter and discussed in-depth.

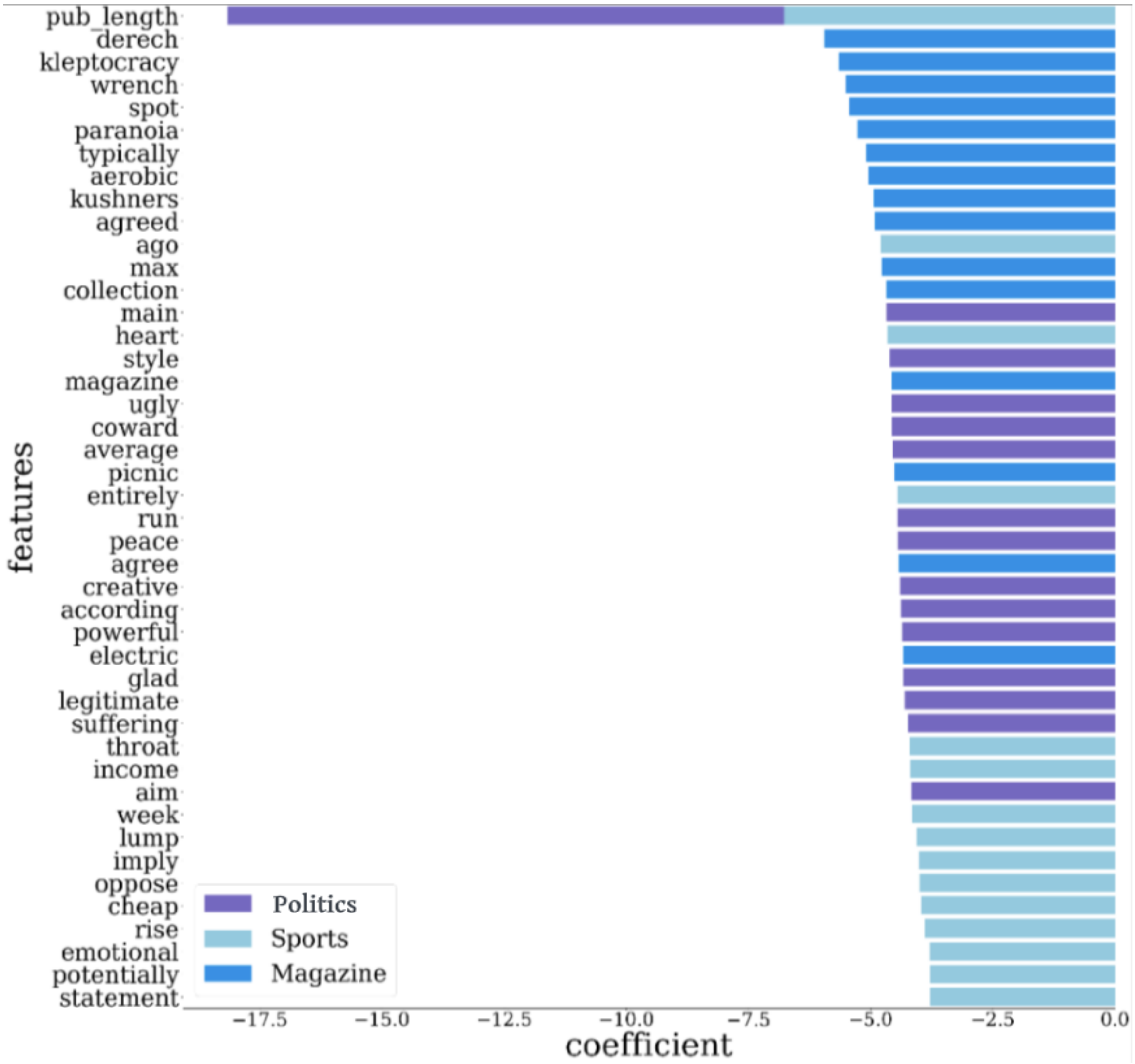

Figure 5.6: Most influential negative features of the logistic regression clf.

To better understand the features, a distribution plot was constructed. It visualises a summary of the arrangement of each news grouping and highlights how significant coefficients fall among the remaining features. In Figure 5.7, all three news groupings have a unique curve. The mean of the plotted distributions can be seen as the peak in the curve, which appears to be zero. This can be interpreted as a multitude of features having no impact on whether a comment receives a reply or not. The remaining features are more likely to have a positive influence over a negative, as it is apparent from all three news groupings that a broader range of positive coefficient values exists. Additionally, there is a close negative range found in comparison to the vast positive range for the sports and politics groupings.

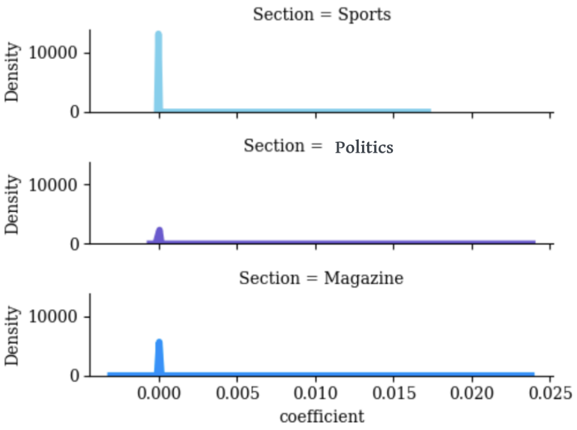

Figure 5.7: Distribution of logistic regression clf. based on news grouping

5.2.2 Experiment Two

Cross-validation

In figure 5.8, the initial cost was distant from the y value, zero, with the cost centring around 0.34. This space closed around the log10(10) mark, with a newly centred cost value of 0.09. The remaining C values linger around a similar mark. Using fivefold cross-validation, specified in 4.09, ten was selected as the most reasonable value for C. The value, ten, resulted in the simplest model with a small cost range closest to zero.

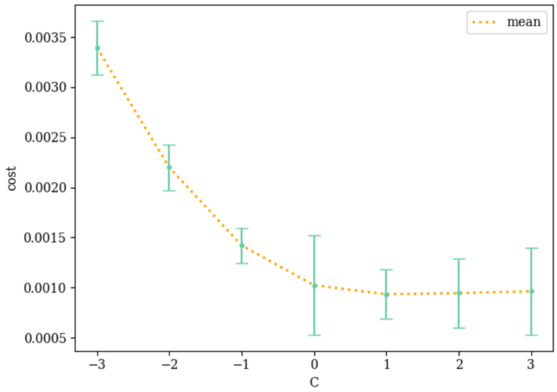

Figure 5.8: Linear reg. sports model with a varying C

Models performance

The linear regression is the stand-alone model for Experiment Two. It was selected based on existing literature as mentioned in Chapter 2; so, therefore, a comparative study across models or baselines was not necessary. More time and resources could be re-delegated to discussions around news groupings and patterns. As stated in section 3.4, the lower the value of the MSE, the better. Therefore, the model performs reasonably on both seen data and unseen data.

	Linear Regression	
	Test	**Train**
MSE	0.00083	0.00084
RSE	0.02887	0.02894

Table 5.7: Performance metrics for sports linear regression models

Features and coefficients

As in Experiment One, influential coefficients across news groupings were displayed. Figure 5.9 shows the influential features that lead a comment to receive a high number of recommendations. The magazine coefficients appear to have the greater influence, followed by sports, then politics, bearing in mind this is based on where most of the news groupings are found in the figure. Chapter 6 provides an in-depth discussion surrounding these coefficient values.

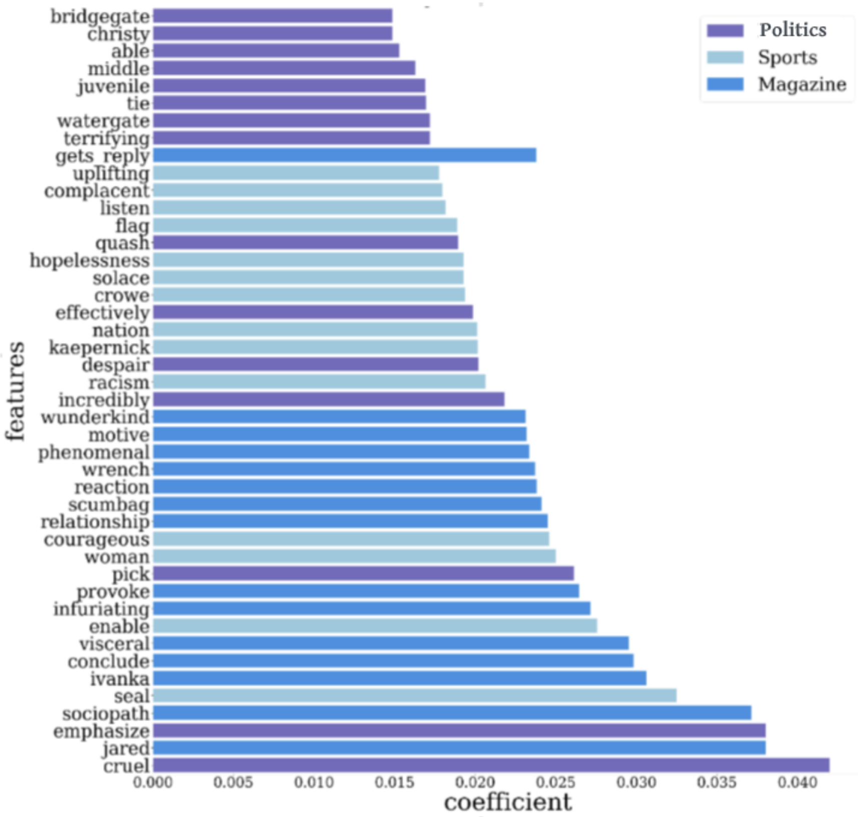

Figure 5.9: Cross-Validation for the linear regression sports model with a varying C

Figure 5.10 displays the influential features that cause a comment to have no recommendations. A similar pattern is found in Figure 5.9, where magazine has the highest coefficients, followed by sports than politics. All but one of the features are textual; the non-textual feature is pub_length and can be seen to have the highest coefficient value among the rest. This is only the case for politics and will be discussed further in the following section.

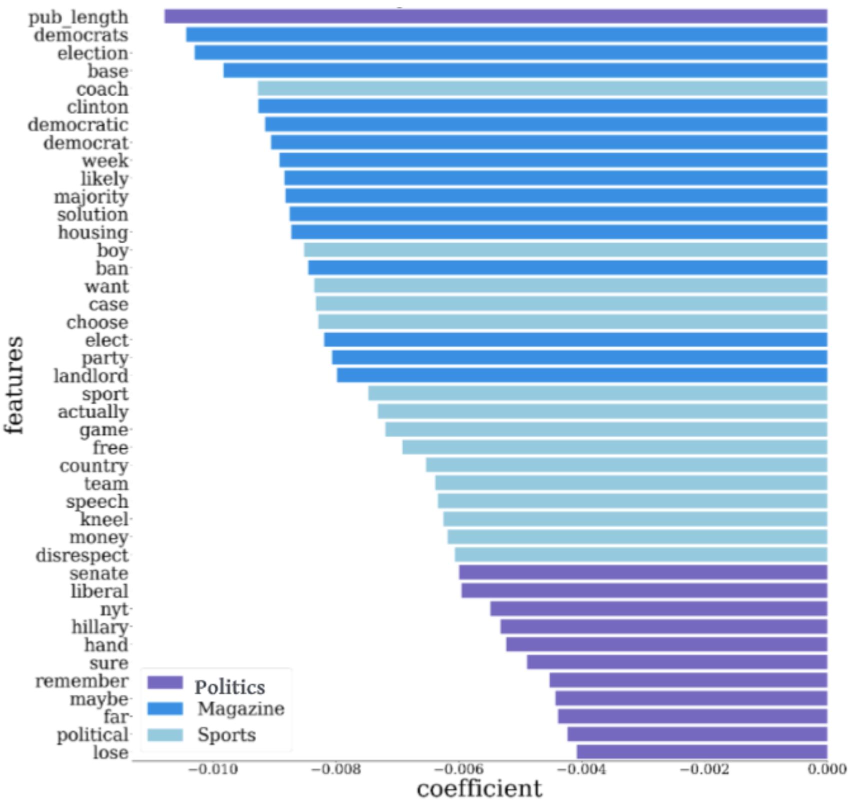

Figure 5.10: Cross-Validation for the linear regression politics model with a varying C

From Figure 5.11, the distribution of features for the three news groupings is shown. All three distributions are centred around zero, with a high density of values situated around this mark. In contrast to Experiment One, the plots in Figure 5.11 bear a higher degree of similarity among them. There is no significant difference between the negative and positive coefficient ranges or the density areas. Therefore, the distribution for Experiment Two delivers similar characteristics to each other.

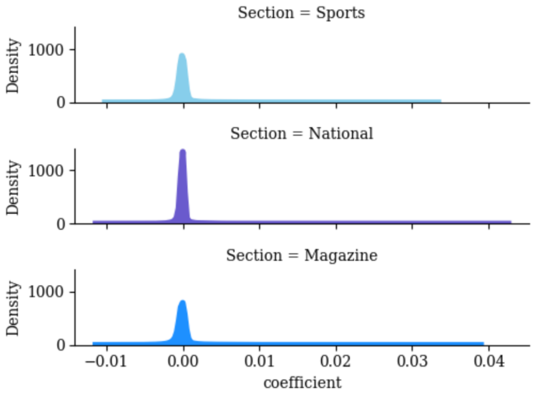

Figure 5.11: Cross-Validation for the linear regression politics model with a varying C

Chapter 6

Discussion

The results presented in the earlier section will be discussed in-depth in this chapter. The subsequent conversation intends to explore numerous explanations for the findings, when possible, before concluding, followed by examining the implications of the findings. The chapter is situated on the two experiments; it is important to note that Experiment One is discussed significantly more. Many remarks made when analysing Experiment One can also be applied to Experiment Two due to the similar user engagement nature of the target variables. For understanding the features in Experiment Two, they are examined with the claims contrived in Experiment One. This approach is taken as Experiment Two tends to reiterate similar patterns in justifying the findings. In addition, highlighting in-differences between the two experiments can shed more light on moderators. For example, if a comment is predicted to receive no replies but a high number of recommendations, it can be possibly assumed that the likelihood of the comment being low-quality diminishes, as it was perceived well by the community. Therefore, moderators could redirect their resources to areas in higher demand for moderating. However, as the distributions differ between each experiment in terms of range and high dense areas on the plot as presented in Chapter 5, caution must be applied to the scales of the coefficient values when comparing the experiments. Therefore, rather than measuring the coefficients, the sign of the coefficients is used in comparison across experiments.

The discussion focuses on understanding the coefficient values and why they vary across news groupings rather than how moderators would integrate this factual piece of information into their system. In presenting the findings, several explanations for the results are given, along with general hypotheses regarding analyzing the comment section conditioned on news groupings. It is assumed that these general hypotheses collectively would help inform moderators in fully understanding the contours of existing norms or the risk of commenter's patterns. It is also important to note that the discourse of this chapter is centred on the assumption that the comment sections have reached a point where they are considered to have reached the total amount of expected comments. This means that the findings found, although they are on comment sections that are crowded with comments, moderators can use them to identify what areas are essential to look at a comment section is initially open and commenting commences. This is supported by the purpose of the machine learning models, such that the main objective is to predict some posed questions. Moreover, due to the high accuracy the models achieved, the tool would be more beneficial to moderators.

6.1 Experiment One

Based on the findings in Chapter 5, it was concluded that the logistic regression model was the most reasonable to carry out Experiment One. This section analyses all the non-textual features based on their coefficient value in the model of predicting whether a comment receives a reply or not. Additionally, the section probes the various textual features that bear relevance by highlighting trigger words and themes.

6.1.1 Get reply

In Experiment One, the dependent variable, the output of the process, is the get_reply feature. This variable holds the phenomena which Experiment One is centred around. The subsequent analysis of independent variables enables the identification of important elements in the posed classification problem. This feature was selected as the dependent variable to explore the ideas of Ruiz et al. (2011), who suggested that the NYT shows a greater deal of argumentation. It is important to remember that an assumption is made, where get_reply is a tool in understanding the argumentative dynamics. When the target variable is assigned a positive plus label, the heavily weighted independent features can be possible facets in starting a debate or interactivity between readers. Highlighting features that influence a comment receiving a reply or not was by far the most reasonable approach in attempting to reveal features that vary across news groupings. This also accords with earlier observations, which showed that analysing conversations would contribute to a richer understanding of interactive engagement practices (Ksiazek et al., 2016).

6.1.2 Recommendations

As mentioned in the literature review, the recommendation feature exhibits a strong positive relationship between popularity and comments. The recommendation feature was a result of this focal point of Experiment One due to it acting as an informative indicator for user engagement based on prior studies claiming a comment of a popular nature with a multitude of recommendations will likely lead to more users replying. Changes in the recommendation feature across news groupings were compared using tables and a diagram showing the coefficients, values, and the degree of influence the recommendations have over a comment receiving a reply.

The results obtained from the model evaluation of the recommendation feature for each news grouping can be compared in Table 6.1. This table is quite revealing in several ways. First, the coefficients are more statistically significant than the other coefficients. A more comprehensible overview of these coefficients is given in Figure 5.5. The large difference between recommendations and the other coefficients can be attributed to recommendations being much more influential in leading comments to receive a reply. This regards the importance of other features but rather shines a light on the gravity of recommendations in a comment.

	Sports	**Politics**	**Magazine**
recommendations	25.68520	50.495751	40.22179

Table 6.1: Coefficient values for recommendations feature in Experiment One

A second observation from the table is the high positive correlation between recommendations and receiving a reply across all three news groupings. Therefore, it seems that comments with several recommendations are more likely to encourage interactivity between users in the comment section, no matter the news grouping. This result can be perceived as anticipated; generally, on all media platforms, the trend of comments receiving a high number of replies is typically complemented with a high number of upvotes or downvotes. These findings further support the idea of a high correlation existing between popularity and commenting, which corroborates the attitudes expressed by Ksiazek et al. (2016).

In comparing the recommendation coefficient among news groupings, it is apparent that politics has the largest value, followed by magazines and then sport. The result of the leading politics coefficient may be explained by the earlier observation made by Ruiz et al. (2011), stating that a common trend of polarisation of positions emerges in political articles. This can lead to a user commentary where readers engage more with upvoting comments of similar political stances as they are more heavily opinionated and passionate about supporting their polarised position, which could potentially explain the findings of this experiment with the recommendation coefficient being largest for politics. Contrary to Ruiz's expectation of extending the remark to sports articles, this dissertation did not show this suggested result, as the magazine recommendation value appeared greater than the sports. A possible explanation for this may be other external factors not accounted for in this experiment, which may lead the news grouping, magazine, to receive higher engagement via recommendations over the sports news grouping.

A limitation comes to light when attempting to understand a reply being a predictor in the number of recommendations received instead of the recommendations being a predictor in receiving a reply, which is the basis of this experiment. Although these findings are rather disappointing when exploring the inverse of this experiment, the second experiment is designed to capture any potential results on a reply being a predictor of the number of recommendations.

Another issue that emerges from these findings is whether the number of recommendations is due to readers being swayed by how other readers perceive a comment, where a comment receives a high number of replies or recommendations and, therefore, the reader's perception, especially readers that lack media literacy, is influenced by the popular response. It is encouraging to compare this issue with that found by Wilson et al. (1989), who found that readers have a preconceived idea of the level of quality a piece of content should have, and if their expectations are not matched, then they disregard the surrounding content, even if it may be accurate. A minor comparison can be drawn between this 'Affective Expectation Model, and the high correlation between replies and recommendations reported in this dissertation. The reader's perception can be influenced effortlessly by the articles surrounding content that can be mapped to this finding, as a popular comment may easily influence readers' perceptions. For this dissertation, reader's perceptions being swayed by user engagement metrics are not accounted for, similarly to the study conducted by Towne et al. (2017), where a lab study is a feasible alternative option in order to account for reader's perceptions by capturing response times, as mentioned in section 3.3 in Chapter 2. A lab study for users interacting with the comment sections could capture their perceptions, which then, in turn, could be factored into the models. This is an important issue for future research.

Published length

Another engineered feature in this research was 'pub_length,' the time difference between the article being published and the posted comment. An initial objective of the project was to identify variables based on the intuition that was relevant to the study. A similar attitude was taken by Veeramachaneni et al. (2014) , as mentioned in section 2.3 in

Chapter 2. This time feature was engineered based on this premise and using intuition; the time feature was identified as an influential potential indicator in whether a comment receives a reply or not, hoping that the longer time passes by, the less likely a comment receives a reply.

	Sports	**Politics**	**Magazine**
pub_length	-6.764866	-18.16156	-1.16999

Table 6.2: Coefficient values for pub_length in Experiment One

The results in the table for the three news groupings indicate that the length of time between a published article and a posted comment can negatively impact the likelihood of a comment receiving a reply. Keeping this in mind, the model attempts to minimise coefficient values as mentioned in section 2.2 in Chapter 2. Therefore, the lowest value among the three is still considered influential with a heavily weighted figure.

All three features negatively correlate with the coefficients seen in Table 6.2. A possible explanation for this is that a reader replies to comment to interact socially, an observation made by Springer et al. (2015), stating the main motivation for commenters is social interaction. A response from the recipient is ideal when readers are replying with the motive to fulfil social needs. Therefore, it is more probable for a reader to reply to a recently posted comment, as it implies signs of an active recipient with a higher possibility of responding while the topic is still fresh for them. Rather than an older comment, where time has passed, the recipient feels less inclined to reply as their initial reaction to the topic had lessened compared to when the comment was initially posted. Another possible explanation for time coefficients having a significant negative correlation is that commenters reply to comments regardless of peak hours. The findings indicate that timeliness does not play an important role in a comment receiving a reply, which articulates commenters replying to comments regardless of when most commenters or readers are active.

Sentiment

Objectivity and subjectivity are two metrics used in calculating sentiment, where zero represents objectiveness and one for subjectivity. All topics have subjective elements, especially the comment section, where topics are open to interpretation. For example, the rules for sport create an objective environment, where the team who scores the most points wins the game. Commenters interpret these game results as bad or good, creating a subjective statement. Although commenters can interpret game results, commenters can also demonstrate objectiveness through objective player performance measures (McIntosh et al., 2019). Therefore, engineering the featured sentiment, developing the ability to capture the contrasting elements, and then deciding whether they encourage user engagement can prove to be a useful tool.

	Sports	**Politics**	**Magazine**
sentiment	-0.773019	0.73669	-0.519698

Table 6.3: Coefficient values for sentiment in Experiment One

It can be seen from the data in Table 6.3 that the politics group reported significantly higher sentiment than the other two news groupings. This finding corroborates the ideas of Somasundaran & Wiebe (2010) who suggested that political and ideological debates on hot issues are popular on the web, where two polarising stances exist. The high political sentiment, therefore, can be explained by a high number of commenters being drawn to politics to get involved in debates where debaters tend to be more subjective in delivering their arguments (Yuyun & Putri, 2016). Although polarisation stances were identified as an element in the comment sections of sports, the sentiment was most negatively correlated between the three news groupings. It seems possible that these results are due to subjective judging in the sports comments not leading to the starting of conversations. Magazines can be viewed in the same light, whereas political subjectivity invites conversation and arguments.

Question mark and exclamation mark

Another approach to investigate the trend of emotions was engineering features based on whether a question mark or exclamation mark was present. As mentioned in 3.3 in Chapter 2, identifying comments with stylistic aspects was to show any variations of an empathic tone between news groupings. A question mark indicates a direct question and whether the commenter is uncertain of an instance of a topic, and it can be tagged as a conversation starter as they wait for replies from others in the online community. Furthermore, an exclamation mark illustrates strong feelings while emphasising a particular matter; identifying a strong feeling comment and examining whether they elicit more responses is enough reason to model it as a predictor in this experiment.

	Sports	**Politics**	**Magazine**
question_mark	0.22983	0.13307	0.27503
exclamation_mark	-0.37716	-0.13907	-0.34407

Table 6.4: Coefficient values for question_mark, and exclamation_mark in Experiment One

From the data in Table 6.4, it can be seen that the coefficient values are relatively similar in comparison to the previously mentioned features. A significant positive correlation exists between the 'question_mark' feature in each news grouping and a comment receiving a reply. Interestingly, the magazine was observed to produce the highest positive correlation among the three groupings, the first occurrence of the magazine coefficient being greater than politics. The weakened value arising out of politics can be associated with the higher concentration of debates found in political comment sections. Although questions can be posed in an effective crossfire, there is a lack of inquisitive general questions between two opponents, as this is not a common strategy in winning an argument. In general, therefore, it seems that posed questions in politics cause less interactivity between readers, whereas readers in sports and especially magazines are more supportive and helpful as they tend to engage more in answering questions.

From the second row of data in Table 6.4, it is apparent that the exclamation mark feature resulted in negative coefficients across news groupings. These results differ from the previously mentioned stylistic aspect; this may be explained by the fact that exclamation marks can be used in numerous ways to fabricate space for flaws. Several authors advise cautious use of exclamation marks, as readers often use them improperly with tendencies to overuse them to manage the recipient's feelings. This poor grammar can be the making of a low-quality comment. If this is the case, then the table's findings suggest that low-quality comments can discourage interactivity between readers. This assumption that exclamation marks can be associated with low-quality comments must be interpreted with caution because there are cases where they can be effectively used. In terms of comparing the coefficients across news groupings, this finding, while preliminary, suggests that low-quality comments possibly occur due to poor use of exclamation marks, which is seen to a greater extent in sports and magazine, and in turn, harms user engagement in these sections. This can be applied to politics to a lesser extent as the coefficient value is not as significant, which may result from proper grammar use throughout the associated comments as commenters acclaim to a higher standard of writing to succeed at delivering their point in a debate.

Article word count and comment word count

In reviewing the literature, no data was found on the association between article word count and whether it helps elicit commenters' interactivity among one other. Stroud et al. (2016) study found that fewer than half of readers who post comments spend more time reading articles than leaving and reading responses. It can be suspected that many commenters do not read the full article, affecting the quality of comments. In reading the headline and scanning the article, low-quality comments are more likely to be constructed. Therefore, the article's length could be a potential tool to retain the reader's attention and, in turn, structure the quality of comments.

	Sports	**Politics**	**Magazine**
article_word_count	-0.199498	0.293701	-0.21426
comment_word_count	0.82438	0.56398	0.79907

Table 6.5: Coefficient values for article_word_count, and comment_word_count in Experiment One

From Table 6.5, sports and magazine articles seem to negatively impact engagement between readers when the article has a high word count. This is the opposite for political news groupings, where longer word counts seem to perform better in encouraging conversation between readers. It seems possible that the results surrounding the sports and magazine are due to readers, whereas, for politics, a possible explanation is that journalists need more space to develop an argument by which the readers have a mutual understanding, and a lengthy article does not dampen the reader's attempts at reading the whole article. This results in having the ability to articulate their views into comments of a much higher standard as they have read the entire article thoroughly. Hence, it is likely that such a connection between the length of the article and a comment receiving a reply exists because the length of political articles does not influence the quality of comments, such that articles of long length do not stop commenters from reading the article and gaining knowledge on the topic producing quality comments. On the other hand, magazines and sports seem to revert to this. Their coefficient values can be interpreted as readers finding the length of an article discouraging. Therefore, readers are more likely to scan the article and swiftly move on to the comment section. This results in a lack of knowledge gained as readers post comments; these low-quality uninformative comments, in turn, will be less engaging to other readers. It is important to bear in mind the possibility of other reasons for low-quality comments, some of which carry heavier weight.

For the comment word count, the feature was engineered by similar means to the article word count. It can be seen that comment word count coefficient values increased with sport leading, followed by magazine, then politics. The theme emerges that all news groupings positively correlate the length of comment and whether that comment receives a reply. A possible explanation for these results is that if a comment is lengthy with high content quality, then the length does not play a negative role on whether that comment will receive a reply, in a way that fellow commenters understand the space needed to develop an argument and do not disengage with a comment on this basis.

Start question

Similar to the reasoning behind engineering question_mark, start_question is engineered on the premise of identifying whether questions elicit responses from other users. Sentences in a comment that started with a question word were tagged as questions being posed. From Table 6.6, the coefficient values can be seen as positive and within a close range, with politics taking the lead. These findings further support the idea of the question_mark feature and the significant positive correlation that exists between a question and a comment receiving a reply. However, the coefficient value for politics is assigned a greater weight than the other two news groupings, which was not the case for the question_mark feature. It insinuates the opposite of the previous discussion. A lack of inquisitive general questions between two opponents in an effective crossfire debate was listed as a possible explanation for why the question marks did not receive the same weight in eliciting replies in comparison to the other news groupings. This rather contradictory result may be due to rhetorical questions being written without question marks ("The Blue Book of Grammar and Punctuation. 2022. No Question About It - The Blue Book of Grammar and Punctuation", 2022). Therefore, it is possible that the use of rhetorical questions encourages the coefficient value to spike, among others, especially in the context of persuasive writing used to engage in debate.

	Sports	**Politics**	**Magazine**
start_question	0.15608	0.29059	0.22629

Table 6.6: Coefficient values for start_question in Experiment One

Text features

All comments are pre-processed into a lengthy list of features where a token represents a feature; therefore, reporting on all feature parameters would be unreasonable. Alternately, the most influential features are reported, as seen in Figure X and the table Y. These feature parameters will be discussed in better detail to help understand the significance of a reader's comment on other readers' engagement practices and whether differing trends emerge across news groupings.

Trigger words

An effective word choice while communicating can convey the commenter's feelings more thoroughly and significantly increase the quality of content, which can determine to what degree readers take action with the comment. The use of trigger words can invoke and engage emotion among readers; by grabbing and keeping the reader's attention, a comment may go on to resonate with a wider audience and attract other high-quality comments. There is abundant room for further progress in determining influential trigger words in comments and gaining a broader view of these across news groupings. These variables could be considered by investigating a wider scope of influential features rather than the top 25 positive and negative. Trigger words can be divided into categories based on the recipient's reaction they attain. Clark (2021) article delivers context for several categories of trigger words; the most prevailing category found among the textual features was powerlessness.

<u>Powerlessness</u>

Commenters that embody powerlessness compose comments with trigger words exhibiting levels of frustration and intense resentment. This permits other readers to empathise with the comment as an unattainable solution is conveyed. By empowering other readers who feel similar, comments can ignite engagement. The table below illustrates the breakdown of words associated with the feeling of powerlessness. These results should be interpreted with caution as words can be used in different contexts, and some words tend to capture the category more.

Strong evidence of powerlessness trigger words was found in the news grouping magazine. These positive correlations suggest that comments with this category variations can increase commenters' engagement practices. Although the other two news groupings show trigger words of negative emotions, no word choice captured the essence of powerlessness compared with the magazine. A possible explanation for this might be that a key characteristic of individuals drawn to the magazine section may be motivated by feeling empowered by others. It is therefore likely that these commenters will engage more with other commenters when they express powerlessness, for example, in the form of being oppressed, as they feel oppressed and feel the need to do something about it (Pease et al., 2013). Although this characteristic may surface across other news groupings and their commenters, it is possible that it is not as statistically significant as magazine.

Themes

Various comparisons across the themes of text features in news groupings can be made. The listed themes in the subsequent table were noted based on several studies. Not all words stood with significance, as they could not be assigned to a theme even when the original context was analysed. Words were assigned a theme based on domain knowledge and existing literature; therefore, the table must be interpreted with caution as space for human error exists. Machine learning models could potentially remove this for future work.

From the date in Table 6.1.2, it can be seen that the greatest quantity of words appointed to themes is magazine. The series of nouns outweighs all the other news groupings, which can be perceived as different concepts or themes. A noun paradigm was expected due to the nature of the pre-processing stage, where natural language techniques were applied to capture object-centred syntax (Stuart et al., 2013). Although a ratio of nouns to verbs exists in table 6.1.2, it is apparent that the news grouping, magazine, is noun heavy. This result may be explained by the fact that readers tend to engage more with commenters when content words are used in common ways to refer to common concepts and experiences (Stuart et al., 2013).

Another observation can be drawn from Table 6.1.2, political references arise more or less the same as in magazines as they do in politics. However, in sports, commenters' attention seems to be drawn to sports establishments, and key sport figures tend to have a problematic nature, as drug use and sexual harassment are popular indicators that encourage interactivity between readers.

Themes based on influential textual features for Experiment One		
Sports(+)	Cultures differences	diversity
	Health	opiate, caught
	Commercialisation	economic, publicly
Sports(-)	Equality	sexist
Politics(+)	Foreign Policy	european
	Governance	arrest, 'pro-'
	Safety	arrest, maternity
	Social and cultural	brady (Tom Brady)
	Social program	maternity, 'pro-', suburban
	Science	physics
Politics(-)	Military	peace
	Governance	policy, rep (republican)
Magazine(+)	Health	yoga, establishment
	Entertainment	saga (Metropolis saga)
	Environment	globalist

	Governance	weiner(Anthony Weiner: congressman), establishment, thetvdems(democratic)
		spicer(Sean Spicer: Trump's press secretary), jack (Louis Jack: politician)

	Science	science
	Wildlife	deer
	Commuting	thumb, bikers
	Travel	bayless (skip bayless)
Magazine(-)	Cultural	drench (rude context)
	Governance	kleptocracy, kushners, ballot
	Guns	paranoia, magazines
	Fitness	aerobic
	Economic	collection
	Health	donor
	Crime	bankrupt

6.1.3 Experiment Two

As mentioned at the beginning of the chapter, Experiment One is discussed significantly more than Experiment Two. The subsequent section expresses similarities and differences between the experiments regarding the non-textual features and the textual features themes. This can signify information presented in Experiment One, for example, where if a comment is predicted to receive no replies. However, with a high number of recommendations, it can be possibly assumed that the likelihood of the comment being low-quality diminished, as it was perceived well by the community. It can thus be said that moderators could redirect their resources to areas in higher demand for moderating.

Get reply

Unlike Experiment One, the engineered feature, gets_reply, acts as an influential indicator of the number of recommendations a comment should receive. This difference is due to the varied structuring of the experiments based on objectives, where both are attempting to predict and evaluate distinct target variables. This feature was engineered on the premise of wanting to create a feature of an argumentation nature to expose informative insights on user engagement. Several authors' studies supported this design choice as they claimed the NYT shows a greater deal of argumentation, and by analysing these conversations, a greater understanding of interactive engagement and practices would be drawn (Ruiz et al., 2011; Ksiazek et al., 2016). Therefore, an additional variable requires evaluation. Table 6.7 presents the correlations among a comment with a reply across news groupings and the number of recommendations. It is apparent from this table that a significant positive correlation exists, with magazines setting the trend, followed by sports, then politics.

A possible explanation for the positive correlation across all news groupings is related to a remark mentioned in section 2.2, where Aldous et al. (2019) found that users' willingness to post a comment varies depending on the topic, and different topics can generate different levels of engagement. Although this statement seems irrelevant in understanding this feature, the perspective can be shaped and applied. For instance, the comment variable was substituted with recommendations and topics were perceived at a level of the content of the comments; then, this remark made by Aldous, An and Jansen has some value in inducing an explanation. Based on this, it is likely that such connections exist between a reply and the number of recommendations, as a comment receiving a reply indicates that the comment's content is thought-provoking as it reaps engagement from other users. Hence, it could conceivably be hypothesised that when a reader sees a comment with a reply, they are led to believe that the comment's content may be of interest relative to the topic and, therefore, is more likely to engage with the comment by possible means of up-voting.

	Sports	**Politics**	**Magazine**
gets_reply	0.017712	0.01169	0.02377

Table 6.7: Coefficient values for gets_reply in Experiment Two

In comparing the news groupings, magazine generated the highest coefficient. It is difficult to explain this result, but it might be related to a claim made by Dutceac Segesten et al. (2020). The study claims that although comment disagreement generates higher dwelling time on a common section, common disagreement does not preclude cue participants' likelihood to share a comment. This connection must be interpreted with caution as the statement is based on sharing rather than liking a comment, which may entail divergent factors. However, comments lacking the ability to be shared on the NYT, we may interpret this claim as if it is centred around recommendations. A clear trend in the news grouping, politics, is debited from existing literature and domain knowledge. Although this may occur in magazines and sport, it seems to surface more frequently in politics. It can thus be suggested that comments with replies to a magazine article are more likely to receive recommendations as the comment is less likely to be of an argumentation nature during a debate. Therefore, based on the Dutceac Segesten et al. study, the lack of potential disagreement in the comment will positively sway a participant's likelihood to recommend a comment.

Question mark

Interestingly, a positive correlation was found between a comment containing a question mark and the comment receiving several recommendations. This result contrasts with the correlation observed in table 6.8, which had a striking negative correlation. Therefore, this feature is heavily impacted by the target variable. In general, it seems that comments that pose a question are more likely to receive a reply rather than obtain many recommendations.

	Sports	**Politics**	**Magazine**
question_mark	-0.00212	-0.00026	-0.00118

Table 6.8: Coefficient values for question_mark in Experiment Two

Article word count

For article_word_count, a positive correlation is present for all news groupings for Experiment Two. This slightly differs from the negative correlation found for sports and magazines in Experiment One. However, the positive news grouping still advances in the leading position; therefore, the same assumption for the feature can be made as in Experiment Two. For Experiment One this assumption must be applied with caution, as the remark made is not as significant in this case but still applies to a certain degree.

	Sports	**Politics**	**Magazine**
article_word_count	0.00078	0.00427	0.002085

Table 6.9: Coefficient values for article_word_count in Experiment Two

Sentiment

The sentiment is another feature that differs across experiments and is worth further investigating. Magazine can be interpreted the same way as in Experiment One, but the remaining news groupings cannot. For sports, comments are highly unlikely to receive a reply when the comment bears high sentiment, in contrast, a comment is likely to gain several recommendations when the relative sentiment is of high value. A possible explanation for this is that commenters like to agree with subjective comments through favouring; however, they do not feel the need to reply to display their approval. This is the opposite case for politics, where commenters are likely to reply to a highly subjective comment, whereas they are unlikely to favour a subjective comment. This can be understood by the reiterated debate trend found in the politics comment section.

	Sports	**Politics**	**Magazine**
sentiment	0.00036	-0.00034	-0.00382

Table 6.10: Coefficient values for sentiment in Experiment Two

Other features

The remaining features in Experiment Two complement coefficient values found in Experiment Two. Therefore, justifying why the features were engineered is not required. In the same vein, an in-depth analysis of similar results to Experiment One is unnecessary; attention is focused on significant differences. In Experiment One, the comment_word_count feature did not show any further significant differences between the feature across news groupings and the target variable; the same understanding from Experiment Two can be applied in the following way. News groupings have a positive correlation between the length of comment and, in this case, the number of recommendations. Furthermore, the pub_length feature conveys a comparable trend to comment_word_count across the two experiments. In Experiment One and Two, pub_length was negatively correlated across news groupings for both target variables, where the longer the time, the less likely the engagement levels were high. Another feature with a similar impact across Experiment One and Two is exclamation_mark. The coefficient values for sports and magazine are both negatively correlated. However, Experiment Two slightly differs for politics as it reveals a positively correlated value. A possible explanation for this is that readers tend to engage we comment containing exclamation marks by up-voting rather than replying; this may be a result of the nature of the exclamation mark such that comments with indications of forceful utterance or strong feelings are less likely to start a conversation but rather receive recommendations as other readers can relate. Moreover, the last feature that follows the mentioned pattern is start_question, where both sports and politics share similar attributes of being positively correlated across both experiments, with one slight difference present in magazine. Magazine elicits a negative correlation for start_question in contrast to a positive correlation found in Experiment One. However, this negative coefficient is very subtle, and therefore the difference is not worth investigating.

6.1.4 Themes

The subsequent table follows a similar setup to the table of themes in Experiment One, as it illustrates some of the main characteristics found in the text features. The table must be interpreted with caution based on assigning themes as there is space for human error. It is apparent from the table that there is a higher density of meaningful text features in predicting the number of recommendations, in contrast to Experiment One, where several more words are considered significant when attempting to assign the text featured to a theme. It is important to bear in mind that the themes are signed based on the majority of which that text feature was used within the comments. There is a clear difference in the populated tables, with around 60 words found in the following table and 39 words found in Experiment One Themes table. A possible explanation for this is that the audience is drawn to topical comments in terms of recommendations, whereas short comments with a lack of personal opinion on a topic are less likely to have readers agree with them, whereas comments with informative remarks or personal experiences may tend to engage the readers more. The mention of topics in the comments has a greater influence on the number of recommendations over the replies to comments.

Another observation is that there is a high mention of political references in magazines and politics for the negative coefficients. It is apparent that blatant references to parties, particularly Democrats, harm the number of recommendations. However, these political references do not have the same negative influence on whether a comment receives a reply. These findings further support the idea of debating between users, as a possible explanation is that comments with a polarised political stance tend to lack recommendations as other commenters do not agree, but they still engage through replying as they want to defend their viewpoint.

Another tendency across both experiments was the speaking of scandals regarding the key members of the public that were intertwined in the scandal. For example, in 2017, there was a scandal surrounding Anthony Weiner, a former congressman involved in sexting scandals. The acknowledgement of this man proved to start a conversation in the magazine grouping. Scandals seem to prevail as a particularly engaging method in the

Themes based on influential textual features for Experiment Two		
Sports(+)	Legal System	judge, incarceration, crow (Jim Crowe Law, anti-black racism) aquilina (Judge Rosemarie Aquilina)
	Equality	racism, woman, protest, citizenship, lynch
	Science	scientist
	Military	troop
	Governance	seal (Bobby Seale - American political activist), kaepernick (Colin Kaepernick, civil rights activist), colin (Colin Kaepernick)
	Health	knee
Sports(-)	Sport structure	coach, sport, game, team, player
	Establishment	money, boy(sexual harassment)
	Military	veteran
Politics(+)	Governance	watergate(scandal), christy (Chris Christie), bridgegate(scandal),
	Legal System	juvenile
	Economic	bank, cyprus (Bank Involved in Scandal)
	Social and cultural	masquerade
Politics(-)	Health	hospital
	Governance	senate, liberal, democrats, nyt, hillary, democratic, political, clinton, lose, win , supporter, government
Magazine(+)	Economic	jared (Senior advisor to Trump), wunderkind(marketing) kushners , kushner (Jared Kushner)
	Governance	ivanka (Trumps wife), win(political win)
	Entertainment	mogul(person in media industry)
Magazine(-)	Governance	democrats, election, clinton, democratic, elect, party, liberal, obama, bernie
	Social Program	housing, landlord

magazine. A possible explanation for this could be the nature of magazine news groupings and the higher newspaper coverage of scandals. In Experiment Two, scandals became apparent as an influential feature, considerably more than in Experiment One. A feasible motive for this is that references to scandals engage the NYT community in recommending the comment over replying. Thus, scandals arise in the comment sections of all news groupings; therefore, moderators should be made aware of this constant condition and advised accordingly in terms of moderating this content.

A suggested starting point for moderators to explore these scandals would be to moderate comments with features that appeared to engage interactivity but received no recommendations. This approach would identify comments that were not favoured but initiated a conversation; several possible meanings for the type of comment warrant these results. In one case, the comment could be of low quality and encompasses an element of disinformation. The vigilant community of the NYT replies in an attempt to add factual remarks to the comment. There are, however, other possible explanations; for example, the comment could be of high quality and have a controversial viewpoint that starts an in-depth discussion but with no need for moderation. Overall, these results are guidelines and therefore need to be interpreted with caution.

A thorough examination of this was not possible due to the scope and time constraints of the project. Puglisi & Snyder Jr (2011) study found that Democratic-leaning newspapers provide relatively more coverage of scandals involving Republican politicians than scandals involving Democratic politicians. It would be interesting to explore these newspapers and their comment sections and whether political scandals prevail to be an engaging topic for interactivity when they are composed with these motives.

Chapter 7

Afterword

7.1 Work complete

The present study set out to determine potential ways in aiding moderation systems by examining comment sections based on news groupings in hopes that a new angle will reveal more breakthroughs in tuning out toxic comments. Several noteworthy contributions to improving moderation systems were found through applying machine learning techniques. The machine learning pipeline consisted of preparing, modelling, evaluating, and analysing the data. Data Preparation was a combination of cleaning and pre-processing, features were engineered during this stage, and a resulting CSV file was produced. Stage 2 splits the CSV file into training and testing data sets to provide the succeeding trained models with unseen data. Cross-validation was performed on the model's hyperparameters. The results were evaluated against the model's performance on unseen data in evaluating the trained models. For Experiment One, logistic regression, KNN, and SVM were evaluated. Logistic regression was selected as the final model due to its performance, simplistic nature, and scalability. Due to the large body of prior work that uses linear regression with textual and non-textual features to directly predict a numerical output, it was decided to implement a linear regression model for Experiment Two. The model's target variables were both user engagement metrics, as it was decided in Chapter 1 that it would be best to inspect trends in the comment section concerning the level of engagement they encourage. The last stage consisted of interpreting the models' coefficients and understanding the findings.

7.2 Significant findings

This investigation demonstrates that news groupings can elicit similar and different reader responses; therefore, it is worthwhile for moderators to understand these characteristics. Bringing awareness to these characteristics, a moderator can identify what areas in a news grouping need attention and create an online discussion space facilitating a safe conversation while allowing freedom of expression. The news groupings selected are only examples to set up the framework. Moderators could potentially use this on different topics to retrieve similar results as found in this dissertation. With the three news groupings, sports, magazine, and politics, it is apparent that there is a range of findings, revealing similar characteristics and varying characteristics. These findings are:

The evidence from Experiments One and Two suggests that by identifying the themes of influential words in a news grouping, the range of these themes, and the frequency, a profile proved to be helpful to a moderator can be made up. This profile would list themes found in comments that cause highly engaged areas or themes found in comments that cause shallow engaged areas. Integrating this into a moderation system would involve ML and AI techniques to identify comments with themes listed on the profile. Human moderators can then review these comments, as they risk reaching a larger audience or being of low quality. This explanation assumes that highly engaged areas will have a larger audience and a more significant impact and the moderation of comments. Moreover, the assumption is that low engaged areas will signify lower quality comments.

From concluding both experiments, the findings suggest that, in general, comments referencing scandals, particularly crucial figures involved in political scandals, seem to elicit some reader response across all news groupings. The current findings add to a growing body of literature on the political blogosphere (Adamic & Glance, 2005; Diakopoulos & Naaman, 2011; Kim et al., 2021). While this finding did not confirm attributes unique to news groupings that moderators should be aware of, it partially substantiated how moderators should tackle scandals. It is suggested that predicting comments of a scandalous nature to have many replies but no recommendations could be a possible practical starting point in attempts to round up low-quality comments. Furthermore, a clear trend emerged was the argumentation nature of comments between commenters. This finding suggests that, in general, engineering features based on the intuition of the underlying problem act as a helpful tool. Several engineered features, such as the exclamation mark, and start question, perform well as precise indications of a crossfire debate. Depending on the moderators' primary objectives, it can assist moderators in locating debates, which poses the threat of passionate readers verbally attacking each other.

The analysis of coefficients undertaken here has extended our knowledge of which non-textual features influence user engagement practices. Although the current study is on three news groupings, the findings suggest that evaluating the coefficients can lead to informative insights into which aspects of the comments are worth further examining, depending on the news grouping. This research has revealed several valuable similarities and differences when understanding news groupings and how evaluating them from this perspective can reinvent moderation systems to advance their performance levels. Firstly, it points to features that had a similar influence on both interactivity and up-voting comments for all news groupings; these were: the word count of a comment and the time difference between when an article was published and the posted comment. These findings enhance our understanding of what indicators in the comment section remain constant across news groupings; whether the correlation is positive or negative, the coefficient sign remains the same. Therefore, moderation systems should take these as a given for investigating all news groupings. If a moderator's objectives are to highlight areas of moderation that differ between these news groupings, these constant features should be the last means taken.

Secondly, it points to features that slightly differ across the two-target variables. Whether the comment contains an exclamation mark and whether the comment starts with a question word. These differences illustrate their varying impact on the two target variables and insinuate that both metrics are not always perceived as positive interactions between readers. The recommendations feature only conveys positive interactions, whereas the reply feature does not necessarily mean the replier favours the comment, as they may sceptically reply, causing a dispute. Moderators should initially review information relating to engagement metrics with ambiguousness; it holds more insights into the fluctuation of comment types.

Thirdly, it points to features that vary significantly across experiments and news groupings. These were: the sentiment of a comment, the article's word count, and whether the comment contained a question mark. While this study did not confirm sentiment to be constant throughout all news groupings, it substantiated how subjectivity plays in a comment section and how it varies based on the news grouping. In news groupings, with a higher level of argumentation throughout the comment section, it is apparent that subjectivity can lead to more interactivity. However, it does not drive a higher number of recommendations. Therefore, moderators need to note common sections with highly subjective comments as a larger conversation may follow, which sparks higher possibilities of uncivil remarks between readers.

The main goal of the current study was to determine influential features in models across diverse news groupings to identify relative patterns for moderators. It is now possible to state that investigating comment sections relative to their news grouping can compile a mountain of observations and recommendations for moderation systems. The significant findings advise directing resources to comments that contain proven influential themes or scandals, targeting comments that were flagged by argumentation features, distinguishing patterns that vary across news groupings by ignoring repetitive patterns found in all groupings, investigating engagement metrics with an ambiguousness nature, and moderating subjective comments. Although a chunk of these observations can be ineffectual to moderator systems, there is still significant knowledge gain an awareness that could help moderation systems tackle the cesspools of racism, misogyny, and all other forms of bigotry found in comment sections.

7.3 Future work

This research has thrown up many questions in need of further investigation. Although there is a large and growing body of literature on user commentary, there is a lack of research into potential ways moderation systems can improve. What is now needed are add-ons and the moderation systems in place that will enhance the system's ability and make removing comment sections a distant thought. This study is an example of a different tool moderation systems can use; further work needs to be done to integrate this analysis into a system to gain these valuable insights the models propose effortlessly. It would be interesting to build a content moderation system based on the premise of this study and assess its performance.

Several possible future studies using the same experimental setup are apparent. Firstly, an alternative approach can be taken in data collection to secure current knowledge from respective news sources or miscellaneous news groupings. Carrying out the study with a web scraper rather than a pre-collected data set. If moderating comments move forward, a better data gathering framework needs to be developed. In addition to this, the storage of data must also be revisited. Secondly, experiments with other reputable machine learning models could be conducted. A great focus on the use of unsupervised learning would help establish a greater degree of clarity on the matter of themes within a comment. A good starting point would be applying K means clustering to the text features and comparing these clusters to the influential text features. This is supported by Holm (2016) who claims clustering algorithms can find subtopics within a discussion given the textual content. Holm also highlights a relevant future work about clustering comments based on information besides textual context, e.g., recommendations and number of replies; this would cause the clusters to contain comments of equal importance, and it would stress which comments lead to an avalanche of more comments. Introducing more suitable classification and regression models would also be interesting to investigate; selecting models that are more capable of handling linear data, and a multitude of text features would be more appropriate. From the review of literature, potential models highlighted for a regression problem were random forest, support vector regression, generalised additive models, and neural networks.

In contrast to future studies using the same experimental setup, adjusting the experiment's variables and objectives are also worthwhile. Conducting more experiments could provide more definitive evidence. It is suggested that the association of features from an article and the comment section should be investigated in future studies. This could overcome the barrier of moderation systems questioning which areas are initially important in the initial moments after an article has been published. Lastly, a natural progression of this work is to analyse more features to which moderation systems need to pay attention, e.g., commenter's track record in commenting.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on link discovery* (pp. 36–43).

Aldous, K. K., An, J., & Jansen, B. J. (2019). View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 47–57).

Alessa, A., Faezipour, M., et al. (2019). Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR public health and surveillance*, *5*(2), e12383.

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, *30*(1), 156–168.

Arya, A. (2020). An overview of textual analysis as a research method for cultural studies. *International Journal for Innovative Research in Multidiciplinary Field*, *6*(3), 173–177.

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682*.

Bica, J. (2021). Tweet topic modeling part 2: Cleaning and preprocessing tweets.

The blue book of grammar and punctuation. 2022. no question about it - the blue book of grammar and punctuation. (2022).

Chan, J. K.-C., & Leung, L. (2005). Lifestyles, reliance on traditional news media and online news adoption. *New Media & Society*, *7*(3), 357–382.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., ... others (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, *130*, 104934.

Clark, B. (2021). 65 trigger words and phrases for powerful content.

The comment section. (2002).

Diakopoulos, N., & Naaman, M. (2011). Topicality, time, and sentiment in online news comments. In *Chi'11 extended abstracts on human factors in computing systems* (pp. 1405–1410).

Duggan, M. (2014). *Part 2: The online environment. the pew research center, internet & technology.*

Dutceac Segesten, A., Bossetta, M., Holmberg, N., & Niehorster, D. (2020). The cueing power of comments on social media: how disagreement in facebook comments affects user engagement with news. *Information, Communication & Society*, 1–20.

Elbow, P. (1993). The uses of binary thinking. *Journal of Advanced Composition*, 51–78.

Etim, B. (2017). Why no comments? it'sa matter of resources. *The New York Times*, *27*.

Field, A. (2009). *Discovering statistics using spss.* London: SAGE.

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339.

Gharibshah, Z., & Zhu, X. (2021). User response prediction in online advertising. *aCM Computing Surveys (CSUR)*, *54*(3), 1–43.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., ... West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4), Article–number.

He, L., Han, C., Mukherjee, A., Obradovic, Z., & Dragut, E. (2020). On the dynamics of user engagement in news comment media. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1), e1342.

He, L., Shen, C., Mukherjee, A., Vucetic, S., & Dragut, E. (2020). Cannot predict comment volume of a news article before (a few) users read it. *arXiv preprint arXiv:2008.06414*.

Holm, R. (2016). *Cluster analysis of discussions on internet forums.*

The irish times - media solutions. (2022).

Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010). Movie reviews and revenues: An experiment in text regression. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 293–296).

J. Salganik, M., & C. Lee, R. (2020). To apply machine learning responsibly, we use it in moderation.

Kaggle. (2022).

Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*, 100032.

Kim, T., Kim, H., Kim, J., & Ma, X. (2021). Improving readers' awareness of divergent viewpoints by displaying agendas of comments in online news discussions. In *Companion publication of the 2021 conference on computer supported cooperative work and social computing* (pp. 99–103).

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

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

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society*, *18*(3), 502–520.

Lagun, D., & Lalmas, M. (2016). Understanding user attention and engagement in online news reading. In *Proceedings of the ninth acm international conference on web search and data mining* (pp. 113–122).

Lapham, C. (1995). The evolution of the newspaper of the future. *CMC Magazine*, *1*(7).

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)* (pp. 334–342).

Liu, Q., Zhou, M., & Zhao, X. (2015). Understanding news 2.0: A framework for explaining the number of comments from readers on online news. *Information & Management*, *52*(7), 764–776.

Liu, Z. (2017). Statistical models to predict popularity of news articles on social networks.

Masullo Chen, G., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media*, *61*(1), 108–125.

Masullo Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media+ Society*, *5*(3), 2056305119862641.

McIntosh, S., Kovalchik, S., & Robertson, S. (2019). Comparing subjective and objective evaluations of player performance in australian rules football. *Plos one*, *14*(8), e0220901.

Nelson, M. N., Ksiazek, T. B., & Springer, N. (2021). Killing the comments: Why do news organizations remove user commentary functions? *Journalism and Media*, *2*(4), 572–583.

The new york times company. (2022).

Nguyen, D., Smith, N. A., & Rose, C. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th acl-hlt workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 115–123).

on Capitalizing on Social Science, C., Behavioral Research to Improve the Public's Health, D. o. H. P., & Disease Prevention, I. o. M. (2001). Promoting health: Intervention strategies from social and behavioral research. *American Journal of Health Promotion*, *15*(3), 149–166.

Pease, B., Grey, M., & Webb, S. (2013). The new politics of social work.

Puglisi, R., & Snyder Jr, J. M. (2011). Newspaper coverage of political scandals. *The journal of politics*, *73*(3), 931–950.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, 65-78.

Pykes, K. (2020). Oversampling and undersampling.

Quandt, T. (2018). Dark participation. *Media and communication*, *6*(4), 36–48.

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? the democratic qualities of citizen debates in online newspapers. *The International journal of press/politics*, *16*(4), 463–487.

Russo, I. (2020). Sadness and fear: Classification of fake newsspreaders content on twitter. In *Clef (working notes)*.

Santana, A. D. (2011). Online readers' comments represent new opinion pipeline. *Newspaper research journal*, *32*(3), 66–81.

Siapera, E. (2021). Ai content moderation, racism and (de) coloniality. *International Journal of Bullying Prevention*, 1–11.

Siirtola, P., & Röning, J. (2020). Comparison of regression and classification models for user-independent and personal stress detection. *Sensors*, *20*(16), 4402.

Smith, J. A. (2017). Textual analysis. *The International Encyclopedia of Communication Research Methods*, 1–7.

Somasundaran, S., & Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 116–124).

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information, Communication & Society*, *18*(7), 798–815.

Staff, P. (2020). The anchoring effect and how it can impact your negotiation. *PON-Program on Negotiation at Harvard Law School*, *21*.

Stroud, N. J., Van Duyn, E., & Peacock, C. (2016). News commenters and news comment readers. *Engaging News Project*, 1–21.

Stuart, L. M., Taylor, J. M., & Raskin, V. (2013). The importance of nouns in text processing. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Towne, W. B., Rosé, C. P., & Herbsleb, J. D. (2017). Conflict in comments: Learning but lowering perceptions, with limits. *The 2017 CHI Conference*, 655–666.

Veeramachaneni, K., O'Reilly, U.-M., & Taylor, C. (2014). Towards feature engineering at scale for data from massive open online courses. *arXiv preprint arXiv:1407.5238*.

Wales, J., & Kopel, O. (2019). The internet broke the news industry—and can fix it, too. *Foreign Policy*, *234*, 36.

Walker, M., & Matsa, K. E. (2021). News consumption across social media in 2021.

Weber, P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New media & society*, *16*(6), 941–957.

Williams, J. R. (2019). The use of online social networking sites to nurture and cultivate bonding social capital: A systematic review of the literature from 1997 to 2018. *New Media & Society*, *21*(11-12), 2710–2729.

Williams, S., & Hsieh, G. (2021). The effects of user comments on science news engagement. *Proceedings of the ACM on Human-Computer Interaction*, *13*, 1-29.

Wilson, T. D., Lisle, D. J., Kraft, D., & Wetzel, C. G. (1989). Preferences as expectation-driven inferences: effects of affective expectations on affective experience. *Journal of personality and social psychology*, *56*(4), 519.

Yuyun, I., & Putri, S. M. (2016). Arguments in university-level debating: Subjective or objective. *Journal of Language and Literature*, *16*(1), 29–35.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1–40.

Zong, H., Wu, X., Xue, C., & Chen, F. (2017). Distribution law of user comments on hot news. In *2017 international conference on progress in informatics and computing (pic)* (pp. 461–465).

Chapter 8

Appendix

Including all results for each news grouping in Chapter 5 proved to be cumbersome. Therefore, the subsequent section presents this supplementary material that is not an essential part of the experiments but can be helpful in providing a more comprehensive understanding of their research problem.

8.1 Experiment One

8.1.1 Cross-validation

Logistic regression

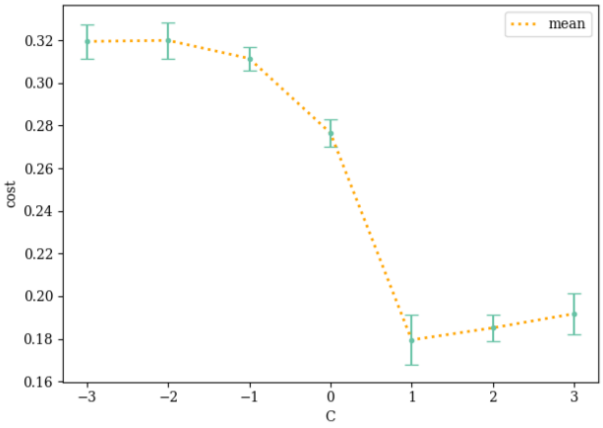

Figure 8.1: Logistic reg. sports model with a varying C

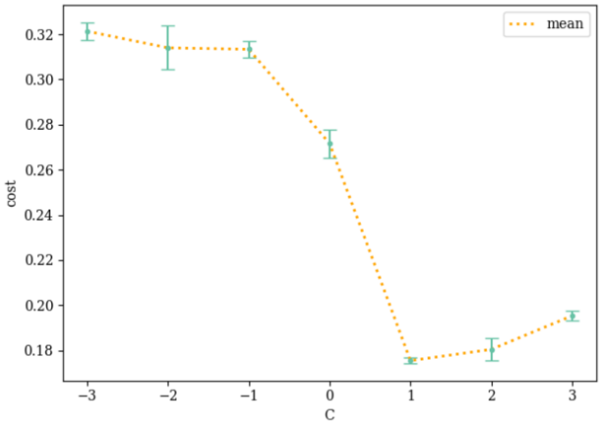

Figure 8.2: Logistic reg. politics model with a varying C

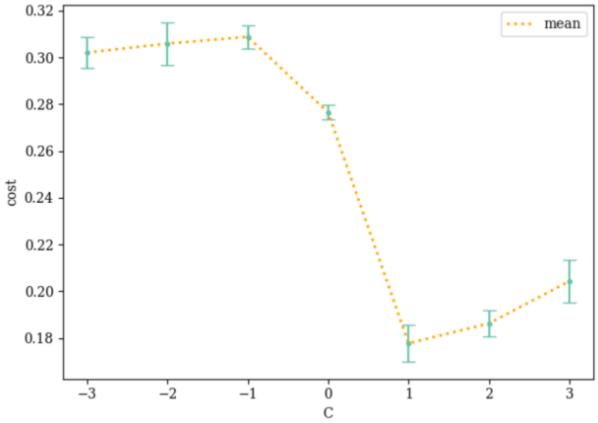

Figure 8.3: Logistic reg. magazine model with a varying C

SVM

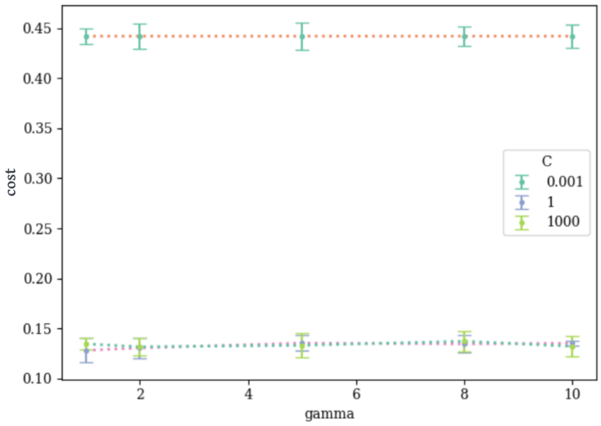

Figure 8.4: SVM sports model with a varying C, and γ

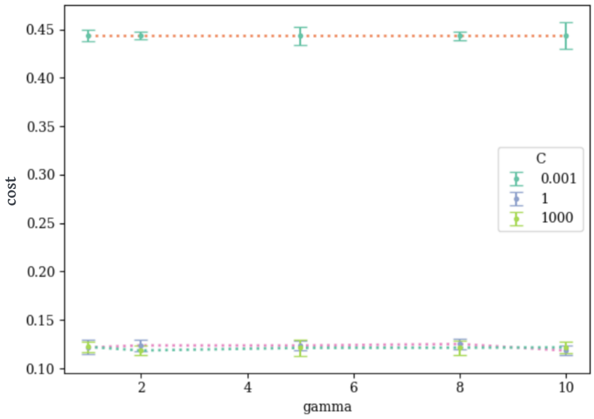

Figure 8.5: SVM politics model with a varying C, and γ

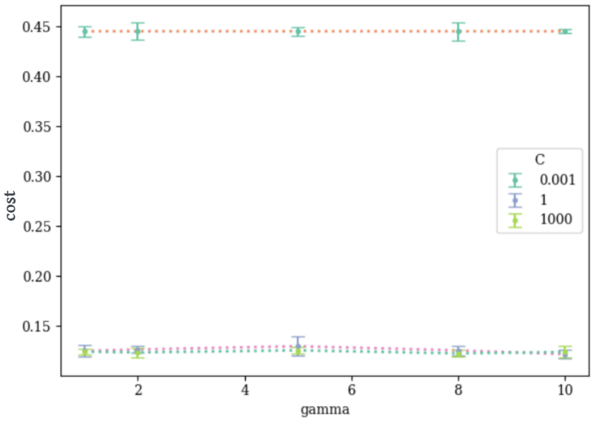

Figure 8.6: SVM magazine model with a varying C, and γ

KNN

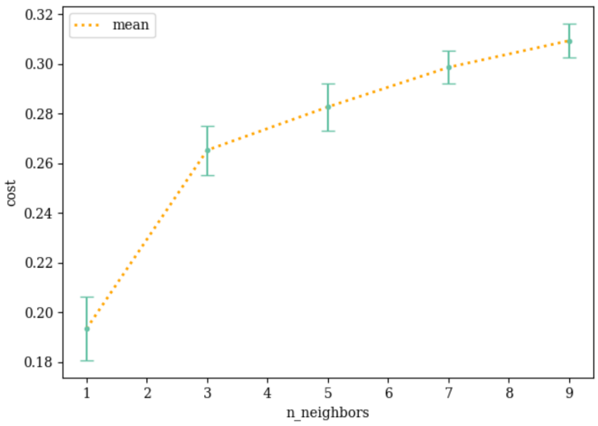

Figure 8.7: KNN sports model with a varying neighbors

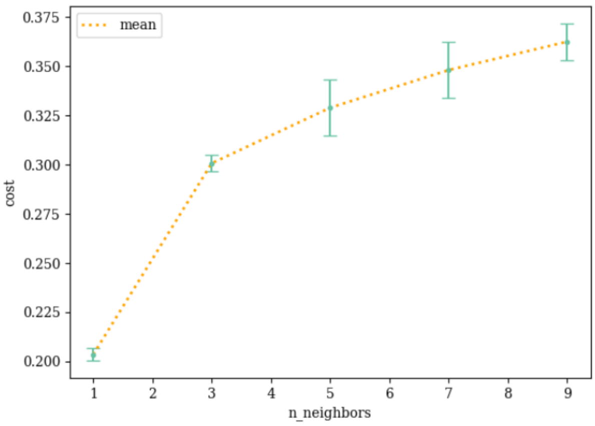

Figure 8.8: KNN politics model with a varying neighbors

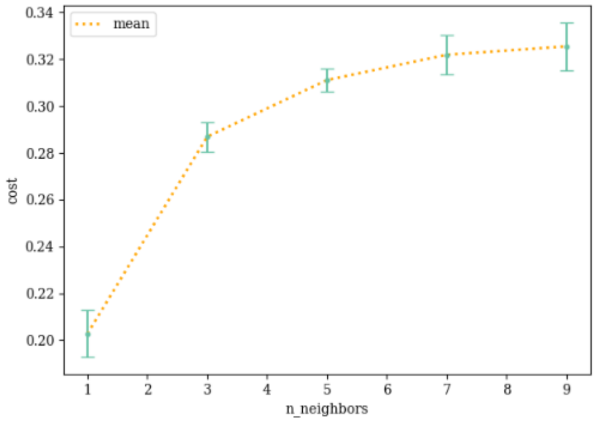

Figure 8.9: KNN magazine model with a varying neighbors

8.1.2 Performance metrics

Politics

	Logistic Reg.		**K-Neighbours**		**SVM**		**Baseline**	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**
Precision	0.7744	0.9818	0.8002	1.0	0.7272	1.0	0.0	0.0
Recall	0.8881	0.9886	0.9175	1.0	0.7622	1.0	0.0	0.0
Accuracy	0.8340	0.9868	0.8543	1.0	0.7548	1.0	0.5521	0.5564
F1-Score	0.8274	0.9852	0.8590	1.0	0.7443	1.0	0.0	0.0

Table 8.1: Performance metrics for politics models

	Logistic Reg.	**K-Neighbours**	**SVM**	**Baseline**
Test	$\begin{bmatrix} 2896 & 974 \\ 810 & 2697 \end{bmatrix}$	$\begin{bmatrix} 3089 & 781 \\ 277 & 3130 \end{bmatrix}$	$\begin{bmatrix} 1611 & 428 \\ 185 & 1489 \end{bmatrix}$	$\begin{bmatrix} 2039 & 0 \\ 1654 & 0 \end{bmatrix}$
Train	$\begin{bmatrix} 8218 & 0 \\ 0 & 6552 \end{bmatrix}$	$\begin{bmatrix} 8218 & 0 \\ 0 & 6552 \end{bmatrix}$	$\begin{bmatrix} 8098 & 120 \\ 75 & 6477 \end{bmatrix}$	$\begin{bmatrix} 8218 & 0 \\ 6552 & 0 \end{bmatrix}$

Table 8.2: Confusion matrix values for politics models

	Logistic Reg.	**K-Neighbours**	**SVM**
AUC Score	0.8876	0.8584	0.84179

Table 8.3: AUC Scores for the politics models

Magazine

	Logistic Reg.	**K-Neighbours**	**SVM**	**Baseline**
Test	$\begin{bmatrix} 3228 & 884 \\ 732 & 2848 \end{bmatrix}$	$\begin{bmatrix} 1696 & 483 \\ 211 & 1514 \end{bmatrix}$	$\begin{bmatrix} 1668 & 511 \\ 201 & 1524 \end{bmatrix}$	$\begin{bmatrix} 2179 & 0 \\ 1725 & 0 \end{bmatrix}$
Train	$\begin{bmatrix} 8664 & 0 \\ 0 & 6950 \end{bmatrix}$	$\begin{bmatrix} 8664 & 0 \\ 0 & 6950 \end{bmatrix}$	$\begin{bmatrix} 8594 & 70 \\ 57 & 6893 \end{bmatrix}$	$\begin{bmatrix} 8664 & 0 \\ 6950 & 0 \end{bmatrix}$

Table 8.4: Confusion matrix values for magazine models

	Logistic Reg.		**K-Neighbours**		**SVM**		**Baseline**	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**
Precision	0.7789	0.9899	0.7581	1.0	0.7631	1.0	0.0	0.0
Recall	0.8835	0.9918	0.8777	1.0	0.7955	1.0	0.0	0.0
Accuracy	0.8222	0.9919	0.8192	1.0	0.7899	1.0	0.5581	0.5549
F1-Score	0.8201	0.9909	0.8130	1.0	0.7790	1.0	0.0	0.0

Table 8.5: Performance metrics for magazine models

	Logistic Reg.	**K-Neighbours**	**SVM**
AUC Score	0.9053	0.8311	0.8728

Table 8.6: AUC Scores for the magazine models

8.1.3 Influential negative coefficients

Sports		**Politics**		**Magazine**	
ago	-4.787498	main	-4.670984	derech	-5.954843
heart	-4.654206	style	-4.603198	kleptocracy	-5.650461
entirely	-4.454366	ugly	-4.566174	wrench	-5.516528
throat	-4.195058	coward	-4.558166	spot	-5.441951
income	-4.184682	average	-4.546636	paranoia	-5.271704
week	-4.157813	run	-4.448475	typically	-5.088920
lump	-4.065357	peace	-4.438555	aerobic	-5.054121
imply	-4.018079	creative	-4.395209	kushners	-4.928556
oppose	-4.000655	according	-4.375456	agreed	-4.915262
cheap	-3.964087	powerful	-4.365421	max	-4.780083
rise	-3.902544	glad	-4.335875	collection	-4.690145
emotional	-3.788329	legitimate	-4.302404	magazine	-4.567389
potentially	-3.785001	suffering	-4.232254	picnic	-4.511900
statement	-3.773328	aim	-4.161376	agree	-4.425959
box	-3.709762	presume	-4.151273	electric	-4.341533
motivation	-3.703898	enhance	-4.126568	ballot	-4.271044
teach	-3.695933	deaf	-4.095110	scummy	-4.266978
conviction	-3.688049	rep	-4.093435	donor	-4.219789
courts	-3.686123	firing	-4.078352	shout	-4.201235
result	-3.685424	speed	-4.048016	budget	-4.193392
provide	-3.634665	leaking	-4.038191	nest	-4.177913
grade	-3.634487	renew	-4.007508	bankrupt	-4.121014
sexist	-3.626985	unconscionable	-3.976296	genuinely	-3.999838
agree	-3.625384	policy	-3.949883	privileged	-3.989831
enablers	-3.590291	sweet	-3.945180	window	-3.968715

Table 8.7: Influential negative coefficients for all news groupings

8.1.4 Influential positive coefficients

Sports		**politics**		**Magazine**	
usual	5.612868	afterthought	6.516426	saga	6.942363
seahawks	5.450343	mislead	6.381895	globalist	6.738814
coverage	5.405586	outside	6.112193	weiner	6.263705
track	5.270600	inconvenient	5.837163	bore	5.813140
postal	5.097800	recourse	5.828033	week	5.809108
woods	5.069369	maternity	5.689405	treasure	5.747612
underdog	5.047697	friends	5.548987	powerless	5.604238
cheerleaders	5.013951	accomplish	5.373914	takeaway	5.584616
opiate	4.969258	wil	5.324035	crusade	5.423007
economic	4.835276	vain	5.294745	thumb	5.410660
pas	4.814031	hysteria	5.232012	spicer	5.370914
typical	4.806868	pro	5.204479	science	5.351307
calm	4.622705	suburban	5.167766	setback	5.231909
realm	4.513821	bradys	5.135443	deer	5.201220
caught	4.510920	obtain	5.070309	yoga	5.167484
willingness	4.503645	miss	5.063395	iota	5.071178
text	4.503560	arrest	5.050159	establishment	5.052444
diversity	4.482167	european	5.044301	thetvdems	5.018357
publicly	4.431530	execution	5.019572	jack	4.990352
speech	4.428248	separate	4.967763	polar	4.977421
hop	4.419077	peril	4.958180	accepting	4.930640
chickens	4.411696	haste	4.935599	appalled	4.849029
anybody	4.392850	physics	4.930197	encompass	4.815196
unhappy	4.303714	scout	4.906492	bikers	4.814333
grass	4.288571	soft	4.904950	bayless	4.782070

Table 8.8: Influential positive coefficients for all news groupings

8.1.5 Coefficients value for engineered features

Features	**Sports**	**Politics**	**Magazine**
article_word_count	-0.199498	0.293701	-0.21426
recommendations	25.68520	50.495751	40.22179
comment_word_count	0.82438	0.56398	0.79907
start_question	0.15608	0.29059	0.22629
question_mark	0.22983	0.13307	0.27503
exclamation_mark	-0.37716	-0.13907	-0.34407
sentiment	-0.773019	0.73669	-0.519698
pub_length	-6.764866	-18.16156	-1.16999

Table 8.9: Engineered features and coefficients for news groupings

8.1.6 ROC Curves for all news groupings

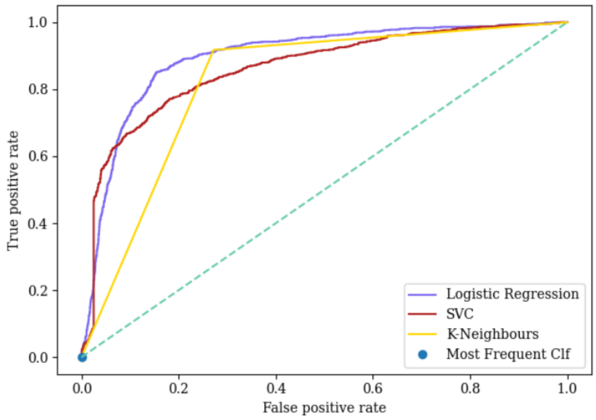

Figure 8.10: ROC curves for the sports models

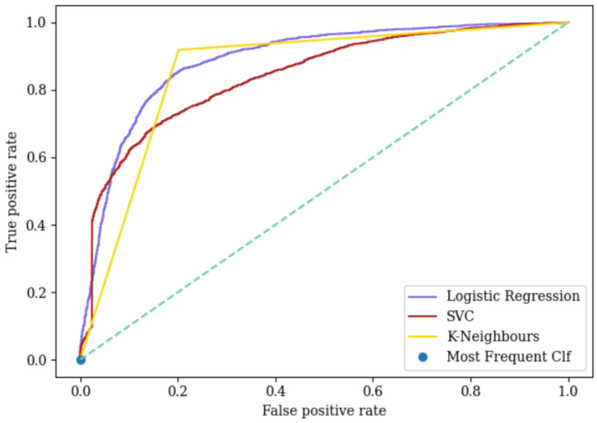

Figure 8.11: ROC curves for the politics models

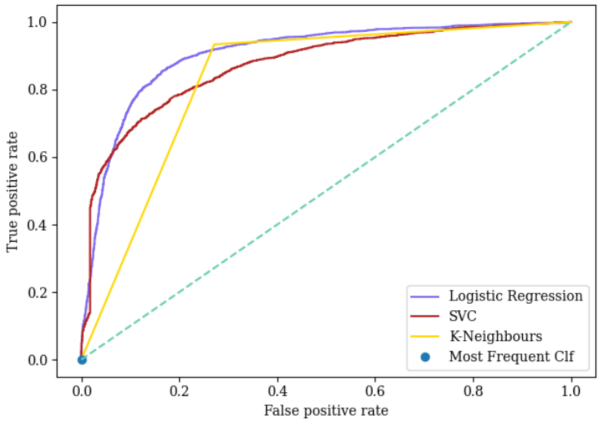

Figure 8.12: ROC curves for the magazine models

8.2 Experiment Two

8.2.1 Cross-validation

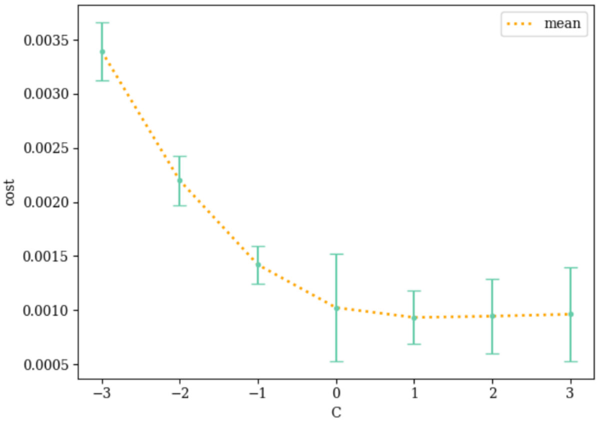

Figure 8.13: Linear reg. sports model with a varying C

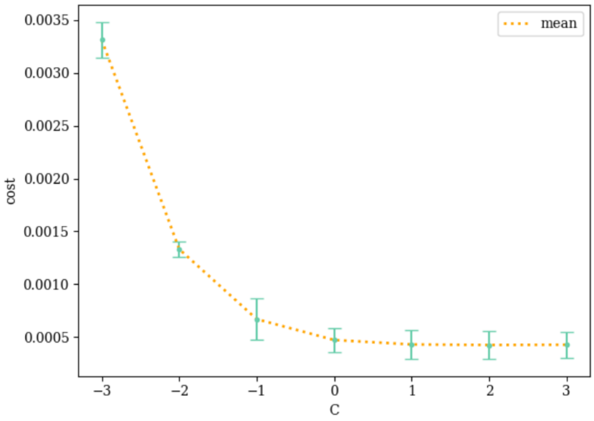

Figure 8.14: Linear reg. politics model with a varying C

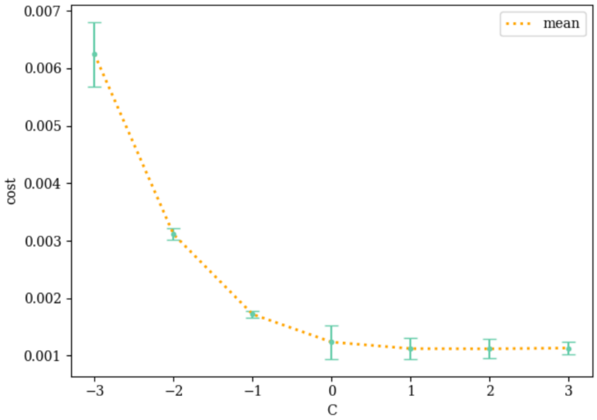

Figure 8.15: Linear reg. magazine model with a varying C

8.2.2 Performance metrics

Politics Model

	Linear Regression	
	Test	**Train**
MSE	0.00048	0.00035
RSE	0.02193	0.018832

Table 8.10: Performance metrics for politics linear regression models

Magazine model

	Linear Regression	
	Test	**Train**
MSE	0.00126	0.00092
RSE	0.03556	0.03033

Table 8.11: Performance metrics for magazine linear regression models

8.2.3 Engineered features

Features	**Sports**	**Politics**	**Magazine**
article_word_count	0.00078	0.00427	0.002085
comment_word_count	0.00368	0.00044	0.00352
start_question	0.001128	0.00136	-0.00014
question_mark	-0.00212	-0.00026	-0.00118
exclamation_mark	-0.00131	0.06414	-0.00015
sentiment	0.00036	-0.00034	-0.00382
gets_reply	0.017712	0.01169	0.02377
pub_length	-0.004358	-0.01078	-0.00558

Table 8.12: Engineered features and coefficients for news groupings

8.2.4 Influential positive coefficients

Sports		**politics**		**Magazine**	
seal	0.032490	cruel	0.041983	jared	0.038017
enable	0.027557	emphasize	0.038001	sociopath	0.037124
woman	0.025008	pick	0.026129	ivanka	0.030624
courageous	0.024578	incredibly	0.021808	conclude	0.029815
racism	0.020608	despair	0.020182	visceral	0.029526
kaepernick	0.020124	effectively	0.019837	infuriating	0.027129
nation	0.020102	quash	0.018917	provoke	0.026445
crowe	0.019372	terrifying	0.017190	relationship	0.024497
hopelessness	0.019261	watergate	0.017187	scumbag	0.024093
solace	0.019261	tie	0.016953	reaction	0.023813
flag	0.018858	juvenile	0.016893	wrench	0.023707
listen	0.018177	middle	0.016264	phenomenal	0.023335
complacent	0.017956	able	0.015288	motive	0.023156
uplifting	0.017736	christy	0.014835	wunderkind	0.023123
troop	0.017696	bridgegate	0.014835	mogul	0.022503
colin	0.017526	heat	0.014457	jeff	0.021884
protest	0.017489	masquerade	0.014347	grifter	0.020839
unexpectedly	0.017320	approve	0.013958	genius	0.020791
citizenship	0.017182	baseless	0.013492	secret	0.019627
aquilina	0.017063	development	0.013381	sessions	0.019434
lynch	0.016418	bank	0.013081	soulless	0.018460
scientist	0.016094	firing	0.012947	win	0.018423
judge	0.015944	tough	0.012947	kushners	0.018222
knee	0.015928	upton	0.012875	kushner	0.018100
incarceration	0.015813	cyprus	0.012570	cod	0.017985

Table 8.13: Influential positive coefficients for all news groupings

8.2.5 Influential negative coefficients

Sports		**politics**		**Magazine**	
coach	-0.009267	senate	-0.005991	democrats	-0.010429
boy	-0.008517	liberal	-0.005957	election	-0.010294
want	-0.008344	democrats	-0.005652	base	-0.009820
case	-0.008321	nyt	-0.005484	clinton	-0.009256
choose	-0.008278	hillary	-0.005312	democratic	-0.009145
sport	-0.007466	hand	-0.005226	democrat	-0.009053
actually	-0.007321	sure	-0.004885	week	-0.008914
game	-0.007197	remember	-0.004519	likely	-0.008826
free	-0.006912	maybe	-0.004423	majority	-0.008816
country	-0.006526	far	-0.004382	solution	-0.008752
team	-0.006377	democratic	-0.004362	housing	-0.008727
speech	-0.006336	political	-0.004222	ban	-0.008443
kneel	-0.006249	clinton	-0.004170	elect	-0.008187
money	-0.006180	lose	-0.004075	party	-0.008061
disrespect	-0.006057	win	-0.004068	landlord	-0.007981
business	-0.006021	hear	-0.004041	firearm	-0.007972
care	-0.006014	supporter	-0.004025	owner	-0.007937
think	-0.005930	excuse	-0.003996	half	-0.007668
action	-0.005908	confirm	-0.003896	hard	-0.007655
hope	-0.005777	cost	-0.003890	discuss	-0.007582
course	-0.005748	government	-0.003871	yes	-0.007460
wrong	-0.005729	leak	-0.003858	liberal	-0.007352
veteran	-0.005686	afford	-0.003805	obama	-0.007351
opinion	-0.005679	turn	-0.003756	bernie	-0.007172
player	-0.005623	hospital	-0.003733	fix	-0.007126

Table 8.14: Influential negative coefficients for all news groupings