GEMA Online® Journal of Language Studies
Volume x (x) xx

The Linguistic Landscape of "Controversial": Sentiment and Theme Distribution Insights

Elizaveta G. Grishechko[1]
grishechko-eg@rudn.ru
RUDN University

ABSTRACT

The study investigates the sentiment and themes associated with controversial topics in news-related social media discourse by analyzing posts containing the evaluative adjective "controversial" in their titles on three news-related subreddits. A mixed-methods NLP approach instrumented via Python was employed, combining VADER-supported sentiment analysis and a qualitative content analysis using n-grams to identify and categorize themes. The sentiment analysis results indicated that the majority of the linked news articles in news-related subreddits had neutral sentiment, which testifies to the predominantly fact-based approach to presenting information with lack of strong emotional connotations. However, the overall compound sentiment scores were negative, which, as the study concludes, suggests a strong negative undertone in the discussions of controversial topics in these subreddits. The theme distribution analysis revealed that Politics and Legislation was the most predominant theme across the three subreddits, followed by Technology and Surveillance, Social Issues and Controversies, Health and Medicine, and Environment and Energy. This distribution highlights the various aspects of society that generate controversy and debate on social media platforms. Study findings have implications for social media and language research and practice by exposing the nature of public discourse on controversial topics. The study also showcases the potential of natural language processing techniques in providing a deeper understanding of complex human language and online interaction.

Keywords: sentiment analysis; natural language processing; social media; online discourse; Python

INTRODUCTION

The rapid proliferation of social media platforms has significantly impacted the manner in which news is disseminated, consumed, and discussed. In this context, the language used to present news stories, particularly in post titles, plays a crucial role in shaping public opinion and discourse. However, despite growing interest in comprehending the linguistic nuances of social media language, there is a research gap in the exploration of commonly occurring evaluative adjectives used in news post titles and their potential influence on sentiment of the corresponding articles.

To address this research gap, the present study employs a combination of web scraping, natural language processing (NLP), and sentiment analysis techniques to investigate the presence of the evaluative adjective "controversial" in post titles on news-related subreddits, analyze the sentiment of the corresponding linked news articles, and identify the key themes associated with its use in post titles.

The evaluative adjective "controversial" is of particular interest in this study due to its unique semantics and application in social media discourse. The token inherently implies disagreement or conflict, which can be linked to the polarization and division present in many

[1] Main author, corresponding author

public discussions, especially on social media platforms. Building on this inherent semantics of the token, the study can expand on the nature of discourse surrounding contentious issues, including how these topics are framed and discussed in different online communities. Beyond that, since social media platforms have become critical spaces for public debate and information sharing, analyzing the use of "controversial" in social media discourse can expose the topics that generate strong reactions, divisions, and debates among users, thereby shedding light on the dynamics of online public discussions.

To that end, the study will attempt to reflect on three research questions.

1. What is the predominant sentiment of linked news articles for posts containing the token "controversial" in their titles?

2. What are the predominant themes associated with the use of "controversial" in news post titles on social media, and how are they distributed?

3. What ensuing implications relevant to social media language research and practice can be deduced from the identified sentiment scores and theme distribution associated with the token?

By addressing these research questions, the study can contribute to a deeper understanding of how controversial topics are discussed, framed, and perceived on social media platforms, specifically in news-related subreddits.

THEORETICAL BACKGROUND

SOCIAL MEDIA LANGUAGE RESEARCH

The rise of social media platforms has led to a surge in research on social media language, as scholars strive to understand how language is utilized and evolves in these digital environments. Earlier studies focused on the unique linguistic features of social media discourse, such as the use of abbreviations, emojis, and unconventional grammar (see Miyake, 2007; Kralj Novak, 2015). More recent research has expanded to examine the effects of social media language on various aspects of interaction, such as identity construction (see Kasperiuniene & Zydziunaite, 2019), persuasion (see Agur & Frisch, 2019), and information dissemination (see Vosoughi et al., 2018; Agapova & Grishechko, 2016).

Furthermore, research on social media language has delved into the role of linguistic precedence in shaping linguistic practices (see Sibul et al., 2019), as well as the impact of platform affordances, such as character limits and multimedia capabilities, on language use (see Zappavigna, 2018). Discourse analysis has also emerged as a valuable approach to studying social media language, enabling researchers to explore the complex interplay between language, power, and ideology in online interactions (see Wu & Pan, 2022). Additionally, studies have investigated the influence of social media language on offline engagement, revealing that digital language practices can both reflect and shape broader linguistic trends (see Lai & Fu, 2021).

Researchers have also looked into the role of social media language in the construction and negotiation of social relationships and group identities. For instance, Chau & Lee (2021) has explored how users employ linguistic strategies, such as code-switching, to navigate and signal belonging in diverse online communities, while Schoenebeck et al. (2023) addressed the challenges posed by online harassment and hate speech, highlighting the need for effective moderation tools and policies that account for linguistic nuances.

Moreover, the analysis of social media language in crisis communication (see Splendiani & Capriello, 2022) and health-related discussions (see Fernández-Luque & Bau, 2015) has proven valuable in understanding public opinion, misinformation, and the role of language in shaping collective responses.

The rise of social media platforms has also led to an interest in researching memetic language and the factors that contribute to the virality of content. Memetic language refers to the use of memes, catchphrases, and other forms of easily shareable content that often combine text, images, and multimedia elements. This area of research explores how language is utilized and evolves in these digital environments to create relatable, humorous, or thought-provoking content that resonates with users and encourages sharing. In this field, linguistic research investigates the unique linguistic features of memes and other viral content, such as the use of humor, irony, intertextuality, and visual elements (see Shifman, 2013). Researchers examine how these features contribute to the appeal and shareability of memetic content across social media platforms.

Sentiment analysis has become a popular research area in social media language studies, as it allows for the examination of emotions, opinions, and attitudes expressed in digital texts. Researchers have employed a variety of computational techniques, including machine learning and natural language processing, to analyze sentiment in social media content (see Kavitha et al., 2022). Applications of sentiment analysis in social media research range from understanding consumer preferences (see Păvăloaia et al., 2019) to monitoring public opinion on political events (see Belcastro et al., 2022) and detecting mental health issues (see Benrouba & Boudour, 2023).

Thus, the growing body of research on social media language reflects the diverse and complex ways in which language is employed, adapted, and transformed in digital contexts, unmasking the broader implications for society, culture, and discourse.

EVALUATIVE ADJECTIVES IN MEDIA DISCOURSE

Evaluative adjectives play a crucial role in shaping public opinion and discourse in media contexts. They contribute to framing, a process in which certain aspects of a topic are emphasized to convey a particular interpretation or evaluation. Research on evaluative adjectives in media discourse has investigated their use in various contexts, such as political discourse (see López-Rabadán, 2021), news reporting (see Ash et al., 2019), and opinion articles (see Johannessen, 2015). These studies have found that evaluative adjectives can influence readers' perceptions and attitudes towards the topics being discussed, often by implicitly conveying the author's stance or opinion.

In addition to their impact on readers' perceptions, evaluative adjectives can also shape the overall tone and sentiment of media texts. According to Bednarek (2006), evaluative adjectives are central to the expression of appraisal, which encompasses the linguistic resources used to convey judgments, evaluations, and emotions. Using evaluative adjectives, media outlets can subtly convey positive, negative, or neutral evaluations of events, people, or issues, thus potentially influencing public opinion and sentiment.

In the context of social media, evaluative adjectives take on added significance due to the highly interactive and dynamic nature of online discourse, as well as their highly influential role in determining user engagement and virality (Berger & Milkman, 2012). In addition to framing, evaluative adjectives can serve various pragmatic functions, such as expressing solidarity with or opposition to a particular viewpoint (Grašič, 2022), or signaling group membership and identity (Mostafa, 2019). Research on evaluative adjectives in social media discourse remains limited. However, a few studies have examined their use in specific contexts, such as online comments (see Zubbir et al., 2021) and online reviews (see Azhari & Fang, 2018). The selection of evaluative adjectives in social media discourse has been found to contribute to the spread of particular perspectives or stances, as well as the polarization of online discussions (Alkhammash, 2021). Furthermore, the use of evaluative adjectives has been proved to impact the credibility of news sources, as readers may perceive content framed with

certain evaluative adjectives as more or less objective or biased (Tandoc et al., 2018). These studies suggest that evaluative adjectives can have a significant impact on the sentiment, tone, and dynamics of online discussions.

Despite the growing recognition of the importance of evaluative adjectives in media discourse, there is still a need for more research on their specific use and impact in social media contexts, particularly in relation to the sentiment and framing of controversial topics. By examining the use of "controversial" as an evaluative adjective in news-related subreddits, the present study aims to contribute to this emerging area of research, giving exposure to the linguistic mechanisms that underpin public discourse on contentious issues in online environments.

METHODOLOGY

DATA COLLECTION

The methodology employed in this study involves a combination of web scraping, natural language processing (NLP), and sentiment analysis techniques to collect and analyze data from news-related subreddits. The Python Reddit API Wrapper (PRAW) was used to access and collect data from the Reddit platform. PRAW facilitates the extraction of Reddit posts, along with their metadata, such as post titles, creation time, and subreddit. In this study, news post titles containing the token "controversial" were fetched from multiple subreddits: r/news, r/worldnews, and r/uplifting news covering the period of 3 years to ensure higher relevance of the findings. Additionally, the study collects the full texts of the linked news articles using the BeautifulSoup library, which assists in extracting content from the URLs.

DATA PREPROCESSING

Prior to data analysis, preprocessing was performed to clean and prepare the data for further inspection. The following steps were applied during the preprocessing stage:

a) removal of brackets and their contents from post titles using regular expressions;
b) tokenization of post titles using the NLTK library;
c) extraction of n-grams containing the term "controversial" using the NLTK library.

SENTIMENT ANALYSIS

The sentiment of the linked news articles was analyzed using the VADER sentiment analysis library, which is part of the NLTK library. VADER is specifically designed to analyze social media text and can effectively capture sentiment in short pieces of text. The library provides polarity scores for positive, negative, and neutral sentiment, as well as a compound score representing the overall sentiment. The compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative and +1 (most extreme positive). Positive sentiment: compound score >= 0.05. Neutral sentiment: compound score > -0.05 and < 0.05. Negative sentiment: compound score <= -0.05. Thus, the compound score will increase as the intensity of the text increases towards positive, and, inversely, it will decrease as the intensity of the text decreases towards negative. Further on, average compound sentiment was calculated for the datasets per subreddit under investigation.

r/news and r/worldnews were selected based on their relevance to the study, as they represent news-related social media discourse. This choice aimed to provide insights into the sentiment and themes associated with controversial topics in general news discussion forums. On the other hand, r/upliftingnews was chosen as a control pool of data, as it purposefully focuses on positive news. The hypothesis was that r/news and r/worldnews would yield

different sentiment results compared to r/upliftingnews, as the latter subreddit emphasizes uplifting stories by design. This difference in focus between the subreddits was expected to be reflected in the sentiment scores, allowing for a more comprehensive understanding of the sentiment distribution in the context of controversial topics across different types of news forums.

THEME DISTRIBUTION

To identify the key themes associated with the use of "controversial" in post titles, n-grams containing the token were examined. Specifically, the study focused on trigrams, which are strings of three uninterrupted tokens. Trigram collocations were identified using the TrigramCollocationFinder from the NLTK library.

STATISTICAL ANALYSIS

Descriptive statistics, such as percentages, mean, median, and standard deviation, were calculated for the sentiment scores of the linked news articles using the NumPy library. These statistics provide insights into the distribution of sentiment across the collected data in addition to the overall average sentiment scores per subreddit.

DATA EXPORT

Finally, the collected data, including post titles, linked text, sentiment scores, n-grams, statistical data, and subreddit, were exported to a CSV file for further analysis, presentation and interpretation.

RESULTS

SENTIMENT ANALYSIS OF LINKED NEWS ARTICLES

The sentiment analysis of linked news articles containing the token "controversial" in their titles yielded the following results for the three subreddits under investigation.

A total of 321 titles were analyzed in r/news. The distribution of sentiment scores is presented in Table 1.

TABLE 1. Distribution of sentiment scores in r/news

	Percentage	Mean	Median	SD	Average compound sentiment score
Negative Sentiment	12.55%	0.0881	0.08	0.0569	-0.1661
Positive Sentiment	7.44%	0.0744	0.063	0.0529	
Neutral Sentiment	80.01%	0.8001	0.837	0.1706	

Note: sample size n=321; positive sentiment: (compound score >= 0.05); neutral sentiment: (compound score > -0.05) and (compound score < 0.05); negative sentiment: (compound score <= -0.05)

In r/worldnews, 270 titles were analyzed. The sentiment scores were distributed as shown in Table 2.

TABLE 2. Distribution of sentiment scores in r/worldnews

	Percentage	Mean	Median	SD	Average compound sentiment score
Negative Sentiment	11.9%	0.0745	0.073	0.0543	-0.1228
Positive Sentiment	6.72%	0.0672	0.0695	0.0561	
Neutral Sentiment	81.38%	0.8138	0.85	0.1872	

Note: sample size n=270; positive sentiment: (compound score >= 0.05); neutral sentiment: (compound score > -0.05) and (compound score < 0.05); negative sentiment: (compound score <= -0.05)

A total of 80 titles were analyzed in r/upliftingnews. The sentiment scores distribution is presented in Table 3.

TABLE 3. Distribution of sentiment scores in r/upliftingnews

	Percentage	Mean	Median	SD	Average compound sentiment score
Negative Sentiment	9.38%	0.0737	0.0755	0.0507	0.0198
Positive Sentiment	16.35%	0.0835	0.075	0.0563	
Neutral Sentiment	74.27%	0.7427	0.816	0.2520	

Note: sample size n=80; positive sentiment: (compound score >= 0.05); neutral sentiment: (compound score > -0.05) and (compound score < 0.05); negative sentiment: (compound score <= -0.05)

THEME DISTRIBUTION

Based on the retrieved trigrams from the subreddits, several themes were identified. The distribution of these themes is as follows.

A significant proportion of the trigrams were related to Politics and Legislation, accounting for 26.32% of the total. The Technology and Surveillance theme constituted 12.28% of the trigrams. Trigrams related to Social Issues and Controversies made up 10.53% of the total. The Health and Medicine theme accounted for 8.77% of the trigrams. Trigrams in the Environment and Energy theme constituted 7.02% of the total. The remaining 35.08% of the trigrams were distributed among various other themes such as education, religion, business, entertainment, military, and international relations. Table 4 summarizes data results retrieved for theme distribution with examples.

TABLE 4. Distribution of themes associated with the “controversial” token across three subreddits

Theme	Percentage in the sample	Examples of retrieved trigrams
Politics and Legislation	26.32%	controversial election-law case controversial legislation against controversial Georgia voting controversial copyright rules controversial tax reform controversial judicial overhaul controversial Brexit protocol controversial immigration policy controversial gun-control measures controversial finance regulations
Technology and Surveillance	12.28%	controversial facial recognition controversial facial-recognition technology controversial Clearview AI controversial mass surveillance controversial spyware technology controversial data privacy regulation controversial internet censorship controversial AI-driven policing controversial biometric tracking controversial drone surveillance
Social Issues and Controversies	10.53%	controversial virginity tests controversial cuties movie controversial arrest of controversial asylum detention controversial BBC documentary controversial gender-neutral bathrooms

		controversial racial profiling controversial pay-gap debate controversial police brutality controversial maternity bill
Health and Medicine	8.77%	controversial new drug controversial opioid treatment controversial third-dose vaccination controversial vaccine bills controversial Covid rules controversial stemcell research controversial assisted suicide controversial mental health controversial abortion restrictions controversial GMO food
Environment and Energy	7.02%	controversial Farallon islands controversial pebble mine controversial wolf cull controversial mining project controversial drilling operations controversial pipeline construction controversial NP plant controversial deforestation policy controversial fracking regulations controversial climate-change policy
Other	35.08%	controversial Mohammed cartoons controversial scene depicting controversial pastor Tony controversial military program controversial university admissions controversial religious conversion controversial corporate tax controversial casting choice controversial military intervention controversial diplomatic relations controversial curriculum changes controversial celebrity endorsement controversial trade deal controversial standardized testing controversial CEO compensation

DISCUSSION

INTERPRETATION OF SENTIMENT ANALYSIS RESULTS

Study results provide a comprehensive understanding of the sentiment associated with controversial topics in news-related social media discourse. By selecting r/news and r/worldnews, the study was able to carefully examine the general sentiment surrounding controversial subjects in popular news discussion forums. In contrast, turning to r/upliftingnews as a control pool of data proved useful, as study results have confirmed our expectation that it would yield different sentiment results compared to the other two subreddits.

The sentiment analysis results indicate that post titles containing the term “controversial” in r/news and r/worldnews are predominantly neutral, with 80.01% and 81.38% of the titles having neutral sentiment, respectively. However, the average compound sentiment scores were negative for both r/news (-0.1661 <= -0.05) and r/worldnews (-0.1228 <= -0.05), suggesting that there is still a negative undertone in the discussions of controversial topics in these subreddits.

Importantly, the predominance of neutral sentiment in the linked text articles suggests that the majority of the articles are fact-based and present information without strong emotional connotations. This observation can offer valuable clues as to how controversial topics are discussed in different subreddits, indicating that authors tend to present information in a more neutral manner. However, as is shown in study results, the overall compound score, which captures positive, negative, and neutral sentiment, can still be negative even if 70 to 80% of the dataset is neutral. This is because the compound score is a measure of the overall sentiment across all articles in a subreddit. In our case, although the majority of the articles may have neutral sentiment, the remaining negative sentiment appeared strong enough to ultimately influence the overall compound score to a statistically significant extent. Thus, in the case of r/news and r/worldnews, the negative compound scores suggest that the negative sentiment in the remaining articles is strong enough to offset the neutral sentiment, resulting in a negative overall compound score. This finding demonstrates that there is a negative undertone in the discussions of controversial topics in these subreddits, despite the high percentage of neutral sentiment articles.

In contrast, r/upliftingnews, which focuses on positive news stories, had a noticeably different sentiment profile. Although the majority of titles (74.27%) still had a neutral sentiment, the average compound sentiment score was positive (0.0198 >= 0.05). This finding indicates that controversial topics in r/upliftingnews are framed in a more positive light compared to r/news and r/worldnews, thus highlighting the variation in the framing of controversial topics between different subreddits.

While the average compound sentiment score gives a general indication of the overall sentiment across the dataset, the mean, median, and standard deviation (SD) offer additional insights into the distribution and variability of sentiment scores within each subreddit. This information can help better understand the nuances of sentiment analysis results.

The mean provides an average value for the sentiment scores, showing the central tendency of sentiment in the dataset. This value provides an overall assessment of the sentiment for each category, which is useful when comparing the sentiment profiles across the subreddits. For example, the mean scores in the negative sentiment category for r/news (0.0881) and r/worldnews (0.0745) are higher than those of r/upliftingnews (0.0737), indicating that the negative sentiment is more pronounced in r/news and r/worldnews. Thus, by comparing the means across different subreddits, we can better understand how sentiment differs between them. However, the mean can be sensitive to extreme values or outliers, which is why we also looked at the median.

The median score, which represents the middle value in the distribution of sentiment scores, provides a measure of central tendency less susceptible to the influence of extreme values. Comparing the median scores across the subreddits can help identify potential differences in the overall sentiment distribution. For instance, the median negative sentiment score for r/news (0.08) is higher than that of r/worldnews (0.073) and r/upliftingnews (0.0755), suggesting that the negative sentiment is generally more concentrated in r/news. Thus, by comparing the median sentiment scores, we were able to gain a clearer picture of the “typical” sentiment associated with controversial topics in each subreddit, without the influence of extreme values.

The standard deviation (SD) measures the dispersion or variability of the sentiment scores around the mean. A higher SD indicates a more diverse range of sentiment scores, while a lower SD suggests a more consistent sentiment distribution. In our analysis, the SD scores for negative sentiment in r/news (0.0569) and r/worldnews (0.0543) are higher than that of r/upliftingnews (0.0507), indicating that the negative sentiment scores in r/news and r/worldnews are more dispersed compared to r/upliftingnews. Thus, by comparing the standard deviations across the subreddits, we were able to understand the degree of sentiment variability in discussions of controversial topics within each subreddit.

To recap, while the average compound sentiment score provides an overall view of the sentiment in each subreddit, the mean, median, and SD offered valuable insights into the distribution and variability of sentiment scores. These values, when considered together, indicate that r/news and r/worldnews have a more pronounced negative sentiment and a more dispersed sentiment distribution compared to r/upliftingnews. Understanding these values is crucial when interpreting the sentiment analysis results, as they uncover the overall sentiment landscape of controversial topics in news-related social media discourse. These additional metrics helped us better understand the nuances of sentiment in each subreddit and draw more comprehensive conclusions about how controversial topics are discussed across different online communities.

Overall, study results for sentiment analysis corroborate our previously stated hypothesis suggesting that r/news and r/worldnews would yield different sentiment results compared to r/upliftingnews, as the latter subreddit emphasizes uplifting stories by design.

INTERPRETATION OF THEME DISTRIBUTION RESULTS

The theme distribution analysis provides a deeper understanding of the types of controversial topics discussed across the three subreddits. The most predominant theme associated with the term "controversial" in the post titles across the three subreddits was Politics and Legislation (26.32%), followed by Technology and Surveillance (12.28%), Social Issues and Controversies (10.53%), Health and Medicine (8.77%), and Environment and Energy (7.02%). The remaining themes collectively accounted for 35.08% of the trigrams. This distribution highlights the significant role that politics and legislation play in shaping controversial discussions on social media platforms.

The theme distribution analysis not only provides a deeper appreciation of the types of controversial topics discussed across the three subreddits, but also showcases the various aspects of society that are prone to controversy and debate. As mentioned earlier, Politics and Legislation emerged as the most predominant theme (26.32%), reflecting the highly polarized and contested nature of political discourse in the contemporary social media landscape. The prominence of this theme suggests that users are actively engaging with and discussing policy-related matters, which can significantly impact their lives and society at large.

The second most predominant theme, Technology and Surveillance (12.28%), highlights the growing concerns and debates surrounding privacy, data protection, and the ethical implications of technological advancements. This theme's prominence points to the increasing relevance of technology in everyday life and the need for a broader societal discourse on the responsible use and regulation of technology.

The Social Issues and Controversies theme (10.53%) underscores the wide range of societal concerns that spark discussions and debates on social media platforms. These issues may include race, gender, immigration, and other topics that are often deeply rooted in cultural and historical contexts. The prevalence of this theme suggests that social media platforms serve as crucial spaces for raising awareness and fostering dialogue around these critical issues.

The Health and Medicine theme (8.77%) demonstrates the significance of healthcare-related controversies in public discourse, especially considering the recent global health crisis and debates around vaccination, healthcare access, and the role of the pharmaceutical industry. This theme's prominence also points to the importance of accurate and reliable information dissemination regarding health and medicine to ensure informed decision-making by the public.

The Environment and Energy theme (7.02%) reflects the increasing awareness of environmental issues, climate change, and sustainable energy solutions. The presence of this

theme in the analysis puts a focus on the need for constructive discussions and collaborative efforts to address global environmental challenges and promote sustainable practices.

The theme distribution results highlight the interconnectedness of various themes and their potential to influence one another. For instance, Politics and Legislation may have direct implications on Technology and Surveillance, Health and Medicine, and Environment and Energy, as policies and regulations often govern these sectors. This interconnectedness emphasizes the importance of fostering interdisciplinary dialogues and collaborations in addressing complex, controversial issues.

Another valuable insight is the potential impact of current events and evolving societal priorities on the theme distribution. For example, during times of heightened political activity, such as election seasons, the prominence of Politics and Legislation may increase. Similarly, global health crises or environmental disasters may elevate the significance of Health and Medicine or Environment and Energy themes. Analyzing the theme distribution over time could reveal the ways in which public discourse evolves in response to changing global dynamics.

Lastly, the theme distribution analysis can also give exposure to the different types of controversial topics that may engage diverse demographics and user groups on social media platforms. Understanding the preferences and concerns of various user segments can help tailor communication strategies and create targeted interventions that address the specific needs and interests of these groups, fostering more inclusive and effective discussions around controversial topics.

Thus, the theme distribution analysis offers actionable insights into the various aspects of society that generate controversy and debate on social media platforms. These findings can inform future research on social media discourse and guide discussions and policies related to controversial topics in online communities, ensuring a more inclusive and constructive dialogue.

IMPLICATIONS OF STUDY FINDINGS

The findings of this study have several implications for social media research. First, they demonstrate the importance of conducting sentiment analysis to understand the emotional context of discussions around controversial topics. This knowledge can be beneficial for researchers aiming to study public opinion, online behavior, or the impact of controversial topics on user engagement.

Additionally, the identified themes provide valuable information for researchers interested in understanding the types of controversial issues that generate discussions on social media platforms. These observations can inform the development of future research questions and contribute to a more comprehensive understanding of online discourse surrounding controversial topics.

Study findings also have implications for language research, particularly in the field of natural language processing (NLP). The results showcase the ability of NLP techniques, such as sentiment analysis and trigram extraction, to expose the content and sentiment of online discussions. This information can be used to further the development of more advanced NLP models that are better equipped to understand and analyze complex human language.

Finally, the findings of this study have implications for media practice. By identifying the sentiment and themes associated with controversial topics on social media platforms, media professionals can make sense of how these topics are framed and discussed by users. This knowledge can be used to refine content strategies, develop more engaging and balanced news stories, and understand the factors that influence public opinion.

Thus, the sentiment analysis and theme distribution results of this study transparently depict the emotional context and themes associated with the evaluative adjective "controversial" in post titles across three subreddits. These findings have important implications for social media research, language research, and media practice, and can contribute to a more comprehensive understanding of online discourse surrounding controversial topics.

CONCLUSION

This study aimed to analyze the sentiment and themes associated with the term "controversial" in post titles across three subreddits: r/news, r/worldnews, and r/upliftingnews. The research questions focused on identifying and interpreting the sentiment distribution and the prevalence of themes in the context of controversial topics. To address these questions, a mixed-methods NLP approach was employed, combining Python-instrumented sentiment analysis using the VADER sentiment analysis library and a qualitative content analysis using n-grams to identify and categorize themes.

The study found that although the majority of the linked news articles in r/news and r/worldnews had neutral sentiment, their overall compound sentiment scores were negative. In contrast, r/upliftingnews, used a control pool of data, presented a much more positive sentiment profile, supporting our hypothesis that sentiment results would differ between r/news and r/worldnews and r/upliftingnews. This finding suggests that the framing of controversial topics varies depending on the subreddit's focus and target audience. As highlighted in the research, high neutral sentiment scores bear record to the nature of discussions around controversial topics, while the compound scores reveal the overall sentiment trends and differences between subreddits.

The theme distribution analysis revealed that Politics and Legislation was the most predominant theme across the three subreddits, followed by Technology and Surveillance, Social Issues and Controversies, Health and Medicine, and Environment and Energy. This distribution highlights the various aspects of society that generate controversy and debate on social media platforms, registering the nature of public discourse on these topics.

The implications of these findings are manifold. First, understanding the sentiment and themes associated with controversial topics can help promote social media research, language research, and media practice by unravelling the way users engage with and discuss these issues. This knowledge can guide the development of more inclusive and constructive dialogues on social media platforms, fostering a healthier online environment.

Second, the findings have practical implications for policymakers and stakeholders by clarifying the areas of society that generate the most debate and controversy. This information can advance the development of policies and initiatives that address these contentious issues, ultimately expediating more informed and balanced discussions in the public sphere.

Lastly, this study contributes to the broader awareness of social media discourse and its impact on society. By analyzing the sentiment and themes associated with controversial topics, researchers can expose the dynamics of online discourse and the role that social media plays in shaping public opinion and discourse.

For future research, it would be beneficial to expand the scope of the study to include other social media platforms and online discussion forums, as well as different languages and cultural contexts. This would provide a more comprehensive understanding of the sentiment and themes associated with controversial topics across various online environments. Additionally, future research could explore the relationship between sentiment, language and user engagement, such as comments, upvotes, and shares, to determine how the emotional context of controversial topics influences audience behavior. Moreover, longitudinal studies could be

conducted to assess changes in sentiment and themes over time, offering a more time-sensitive take on the evolving nature of social media discourse on controversial topics.

REFERENCES

Agapova, E. A., & Grishechko, E. G. (2016). Censorship as a factor of information warfare. *Russian Linguistic Bulletin, 3*(7), 43-44. https://doi.org/10.18454/RULB.7.06

Agur, C., & Frisch, N. (2019). Digital disobedience and the limits of persuasion: Social media activism in Hong Kong's 2014 Umbrella Movement. *Social Media and Society, 5*(1). https://doi.org/10.1177/2056305119827002

Alkhammash, R. (2021). The social media framing of gender pay gap debate in American women's sport: A linguistic analysis of emotive language. *Training, Language and Culture, 5*(4), 22-35. https://doi.org/10.22363/2521-442X-2021-5-4-22-35

Ash, E., Xu, Y., Jenkins, A., & Kumanyika, C. (2019). Framing use of force: An analysis of news organizations' social media posts about police shootings. *Electronic News*, *13*(2), 93-107. https://doi.org/10.1177/1931243119850239

Azhari, A., & Fang, X. (2018). Social media applications framework: A lexical analysis of users online reviews. *International Journal of Services and Standards*, *12*(2), 140-162. https://doi.org/10.1504/IJSS.2018.091850

Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus*. London: A&C Black.

Belcastro, L., Branda, F., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2022). Analyzing voter behavior on social media during the 2020 US presidential election campaign. *Social Network Analysis and Mining, 12*(1). https://doi.org/10.1007/s13278-022-00913-9

Benrouba, F., & Boudour, R. (2023). Emotional sentiment analysis of social media content for mental health safety. *Social Network Analysis and Mining, 13*(1). https://doi.org/10.1007/s13278-022-01000-9

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205. https://doi.org/10.2139/ssrn.1528077

Chau, D., & Lee, C. (2021). "See you soon! ADD OIL AR!": Code-switching for face-work in edu-social Facebook groups. *Journal of Pragmatics, 184*, 18-28. https://doi.org/10.1016/j.pragma.2021.07.019

Fernández-Luque, L., & Bau, T. (2015). Health and social media: Perfect storm of information. *Healthcare Informatics Research, 21*(2), 67-73. https://doi.org/10.4258/hir.2015.21.2.67

Grašič, T. (2022). The writing style in travel blogs. *Folia Linguistica et Litteraria*, *13*(41), 187-210. https://doi.org/10.31902/fll.41.2022.9

Johannessen, M. R. (2015, August 30 – September 2). Please like and share! A frame analysis of opinion articles in online news. In *Proceedings of the 7th IFIP 8.5 International Conference* (pp. 15-26). Thessaloniki: Springer International Publishing. https://doi.org/10.1007/978-3-319-22500-5_2

Kasperiuniene, J., & Zydziunaite, V. (2019). A systematic literature review on professional identity construction in social media. *Sage Open, 9*(1). https://doi.org/10.1177/2158244019828847

Kavitha, M., Naib, B. B., Mallikarjuna, B., Kavitha, R., & Srinivasan, R. (2022, April 28-29). Sentiment analysis using NLP and machine learning techniques on social media data. In *Proceedings of the 2nd International Conference on Advance Computing and Innovative Technologies in Engineering* (pp. 112-115). Greater Noida: IEEE. https://doi.org/10.1109/ICACITE53722.2022.9823708

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS One, 10*(12), e0144296. https://doi.org/10.1371/journal.pone.0144296

Lai, C. H., & Fu, J. S. (2021). Exploring the linkage between offline collaboration networks and online representational network diversity on social media. *Communication Monographs, 88*(1), 88-110. https://doi.org/10.1080/03637751.2020.1869797

López-Rabadán, P. (2021). Framing studies evolution in the social media era. Digital advancement and reorientation of the research agenda. *Social Sciences, 11*(1). https://doi.org/10.3390/socsci11010009

Miyake, K. (2007). How young Japanese express their emotions visually in mobile phone messages: A sociolinguistic analysis. *Japanese Studies, 27*(1), 53-72. https://doi.org/10.1080/10371390701268646

Mostafa, M. M. (2019). Clustering halal food consumers: A Twitter sentiment analysis. *International Journal of Market Research, 61*(3), 320-337. https://doi.org/10.1177/1470785318771451

Păvăloaia, V. D., Teodor, E. M., Fotache, D., & Danileț, M. (2019). Opinion mining on social media data: Sentiment analysis of user preferences. *Sustainability, 11*(16). https://doi.org/10.3390/su11164459

Schoenebeck, S., Lampe, C., & Triệu, P. (2023). Online harassment: Assessing harms and remedies. *Social Media and Society, 9*(1). https://doi.org/10.1177/20563051231157297

Shifman, L. (2013). *Memes in digital culture.* Cambridge, MA: MIT Press.

Sibul, V. V., Vetrinskaya, V. V., & Grishechko, E. G. (2019). Study of precedent text pragmatic function in modern economic discourse. In E. N. Malyuga (Ed.), *Functional approach to professional discourse exploration in linguistics* (pp. 131-163). Springer. https://doi.org/10.1007/978-981-32-9103-4_5

Splendiani, S., & Capriello, A. (2022). Crisis communication, social media and natural disasters: The use of Twitter by local governments during the 2016 Italian earthquake. *Corporate Communications, 27*(3), 509-526. https://doi.org/10.1108/CCIJ-03-2021-0036

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wu, G., & Pan, C. (2022). Audience engagement with news on Chinese social media: A discourse analysis of the People's Daily official account on WeChat. *Discourse & Communicatio*n, *16*(1), 129-145. https://doi.org/10.1177/17504813211026567

Zappavigna, M. (2018). *Searchable talk: Hashtags and social media metadiscourse*. New York: Bloomsbury Publishing.

Zubbir, N., Dass, L. C., & Ahmad, N. (2021). Analysis on adjective Suki and its co-occurrences in Japanese YouTube's comment. *Pertanika Journal of Social Sciences & Humanities, 29*(2), 1357-1374. https://doi.org/10.47836/pjssh.29.2.32

ABOUT THE AUTHOR

Elizaveta G. Grishechko, PhD in Linguistics, Senior Lecturer in the Foreign Languages Department, Faculty of Economics, RUDN University, Russia. Research interests cover the issues of cultural linguistics, corpus linguistics, sentiment analysis, natural language processing and the language of academic writing.