# Multi-Model NLP Analysis for the 2024 U.S. Election

**We agree that all members have contributed to this project in an approximately equal manner**

Alex Bell ek22528@bristol.ac.uk
George Corfield ne22902@bristol.ac.uk
Anss Hameed pg21003@bristol.ac.uk
Orlando Luscombe sv22482@bristol.ac.uk
Louie Sinadjan bn22907@bristol.ac.uk
Archie Soares-Mullen gv22078@bristol.ac.uk

***Abstract*—This study applies a multi-model NLP approach over 2.8 million tweets related to the 2024 U.S. presidential election; integrating BERTopic for topic modelling, BERTweet for sentiment and irony detection, RoBERTa for occupational inference, and M3Inference for demographic estimation. As a result, multiple dimensions of the political discourse are captured. The model outputs are validated against external indicators, including mainstream news headlines, market movements, and economic sentiment indices. Key findings reveal a persistent negative bias in tweets, nuanced demographic and occupational patterns in discourse participation, and limited alignment between Twitter topics and traditional media coverage. Overall, this work demonstrated the ability of multiple NLP methods to extract nuanced insights and explores the strengths and weaknesses of using a combination of NLP models to analyse the discourse around the 2024 U.S. election.**

## I. Introduction

The 21st century has witnessed the meteoric rise of social media as an influential force in everyday life, with over 5.17 billion users worldwide maintaining social media accounts [1]. Increasingly, people are turning to these platforms as their primary source of political news. A 2024 Pew Research study [2] found that one in five Americans preferred to use social media to keep up with politics. Of these digital platforms, Twitter is the most influential, with 59% of users saying they use the site to keep up with politics [3]. On the 27th of October 2022, Elon Musk purchased Twitter and reinstated previously banned accounts, including that of former President Donald Trump [4], setting the stage for the 2024 U.S. presidential election in which $1.35 billion would be spent on social media campaigns [5].

Natural Language Processing (NLP) models have been widely used to analyse political polarisation with 154 papers published between 2010 & 2022 [6]. For example, Bidirectional Encoder Representations from Transformers (BERT) has been applied to Twitter datasets to correctly predict the outcome of the 2020 U.S. Presidential Election [7] and the VADER sentiment model was used with Random Forest and Decision Trees to predict the outcomes of Indian Gujarat Assembly Elections 2022 [8]. These papers highlight the effectiveness of deep learning methods for political analysis. To ensure this paper adds value to the literature, a combination of multiple different NLP models were used, which allowed for multiple dimensions of analysis: demographics, topics, occupations and sentiment.

BERT [9] is the foundational model for all except the M3Inference model [10] used in this paper. The development of BERT by Google in 2018 offered significant improvements on NLP tasks. Improving on state of the art results on the MultiNLI, GLUE and SQuAD benchmarks. Traditional language modelling approaches such as Latent Dirichlet Allocation (LDA) rely on word frequency and co-occurrence. In contrast, BERT considers contextual information of a word, providing a 786 dimensional word embedding.

A publicly available dataset from the USC Election Integrity Initiative [11] is used for the analysis performed in this paper. The dataset consisted of 47 million tweets collected between the 1st of May and 30th of November 2024. These tweets were scraped by targeting keyword queries related to political figures, events, and election themes.

The modelling objective of this paper is to validate a combination of NLP methods against real world indicators. Such as news headlines, market dynamics and economic indicators. The NLP methods include topic modelling, sentiment analysis and demographic inference. Topic modelling was performed with BERTopic [12], thematically clustering tweets. BERTweet [13] enabled sentiment classification and irony detection. M3Inference is used to understand the demographics of Twitter users and finally, RoBERTa [14] was used for user occupation inference.

## II. Data Preparation

Each tweet in the dataset was linked to a user profile, potentially exposing personally identifiable information (PII). Therefore, before analysis, cell suppression was applied to all fields containing PII, such as profile URLs and full names. This approach aligns with the principles of "Purpose limitation" and "Data minimisation" under the UK GDPR framework [15]. Since tweet IDs and usernames are considered quasi-identifiers, only tweet IDs were kept to minimise the risk of re-identification.

Given the 280 character limit on tweets [16], users often rely on informal language and emojis to convey meaning

and sentiment. To make sure this nuance wasn't lost when trying to categorise topics or gauge sentiment, tweets were pre-processed before being used by the NLP models. Expressions that were extended versions of words (e.g. noooo and nooooooo) were truncated and encoded as the same token to prevent unnecessary variations. The Pysentimiento [17] opinion mining toolkit was used to translate emojis into word descriptions and hashtags were separated from the # symbol and treated as meaningful keywords; tweets are often constructed so that hashtags are replaced by their equivalent words (#Election becomes election).

The size of the dataset required extensive computation to process. Therefore, to accelerate model inference and training time, tweets that did not meet certain criteria were removed.

| **Filter Applied** | **Number of Tweets** |
|---|---|
| Less than 5 likes removed | 40,890,000 |
| Comments removed | 1,410,000 |
| Non-English tweets removed | 1,880,000 |
| **Final tweets used** | **2,820,000** |

TABLE I
SUMMARY OF TWEETS REMOVED DURING THE DATA CLEANING PROCESS FOR THE 2024 DATASET

Table I shows that after applying the filters, approximately 6% of the dataset remained. Previous research [18] has performed similar filtering, removing tweets with less than 5 likes, as they are more likely to be spam or bot-generated content. Furthermore, comments risked introducing sentiment directed toward other users rather than political candidates or issues, potentially distorting the overall sentiment. Therefore, they were also removed.

The original dataset contained a significant minority of tweets in languages such as Spanish, French, and German. To reduce the size of the dataset and focus the analysis on predominantly U.S. voters, non-English tweets were removed. Tuned BERT models are capable of processing other languages, but in doing so have reduced performance [19]. The decision was made to prioritise performance in this paper. This limits tweets originating from non-U.S. users. However, it is acknowledged that this filtering may have excluded relevant tweets, particularly within the Hispanic community; Spanish being a first language for approximately 13% of Americans [20].

## III. Data Exploration

The data set used for this analysis was originally compiled as part of a research paper exploring public discourse on Twitter during the 2024 U.S. presidential election [11]. Subsequently, the dataset was made publicly available on GitHub and updated with additional fields to provide more extensive information for each post. The data was collected using a scraper that captured posts using keywords adapted to match time-specific events, ensuring that the content remained relevant. The dataset includes 47 million posts spanning from the 1st May to the 30th November 2024, which encompasses the whole election campaign. The posts are distributed across 47 parts, each containing a million posts split into chunks of 50,000 tweets from similar time periods. Demonstrated in data preparation, the data set was filtered to 2.4 million tweets spanning 30 weeks.

The dataset contains 31 features for each tweet, offering comprehensive metadata such as timestamps, attached media, used hashtags, and engagement metrics such as likes, retweets, and view counts. Engagement metrics provided valuable insight into the influence and reach of election-related content. The dataset encompasses a diverse range of accounts, including political figures, celebrities and organisations, resulting in a wide distribution of engagement. The most liked tweets received more than 100,000 likes and reached audiences that exceeded 300,000 users. These metrics enable analysis of the importance of social media in shaping and amplifying political opinions in response to events.

Initial exploration of the prepared dataset revealed significant patterns in tweet volume. The analysis demonstrated that tweets were distributed throughout the 30 week election campaign period, with notable concentration peaks shown in Figure 1. These peaks align strongly with events such as the presidential debates in June and September, as well as smaller events such as primaries and caucuses across the country.

The preliminary analysis identified quite dominant themes through both the frequency of hashtags and keywords. The most prevalent keywords remained the same over time, such as "maga", "trump" and "biden" appearing consistently throughout the campaign. Some keywords evolved over time and spiked based on events such as "jdvance" or "harris" appearing more after July, aligning strongly with affairs at the time. More complex models can use this information to analyse how tweet topics align with major events and news headlines, as well as to understand the sentiment of these posts.

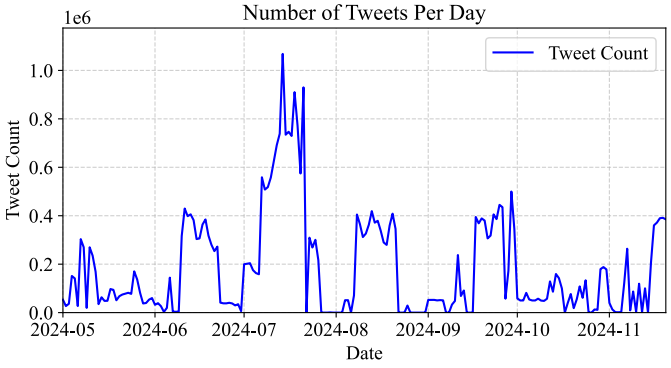


Fig. 1. Graph of the Total Number of Tweets in the Dataset Per Day

## IV. Data Modelling

Extracting insights from a text corpus this large required the use of the University's supercomputer, Blue Crystal Phase 4. Even with this computational leverage, the time constraints of the dataset required the use of filtering (as discussed in Section II). A set of transformer based NLP models were used,

the functionalities and justifications for each these models are detailed below.

*BERTopic*

BERTopic is a Python library developed by Maarten Grootendorst [12] that enables flexible development of topic modelling algorithms. It is separated into six stages, embedding generation, dimensionality reduction, clustering, a vectoriser, a weighting scheme, and optional representation fine-tuning. Although BERT could generate the necessary text embeddings, each sentence would have a different dimension embedding, leading to a high computational overhead when clustering. SBERT (Sentence-BERT) [21] has a constant output dimensionality, allowing cosine similarity metrics, significantly improving computational efficiency and is therefore used as the embedding model. The embeddings undergo dimensionality reduction using the UMAP [22] algorithm. Subsequently, these embeddings are clustered with the HDBSCAN algorithm [23]. Finally, topic representations (labels) are determined with the vectoriser and cTF-IDF algorithm, which highlights the most representative terms within each cluster and creates topic labels.
Figure 2 details the distribution of the topics assigned when run on the USC dataset. This distribution highlights a technicality of the BERTopic library: tweets are only assigned to a cluster if they are sufficiently close to the cluster (using a dynamic distance threshold to maintain cluster stability [23]). This leaves an outlier topic (topic -1) which in this case contains the majority of tweets (17.5%). Although having such a large proportion of the tweets labelled as outliers limits the topic class size, BERTopic's available outlier reduction method was not performed. This is because topic assignment accuracy was preferred over the sample size for later analysis. For example, for the Bitcoin price analysis V-E, a large number of incorrect topic assignments may obscure the pattern, preventing identification. A subset of the topics are explored in more detail in Figures 4 - 6 of the findings section .

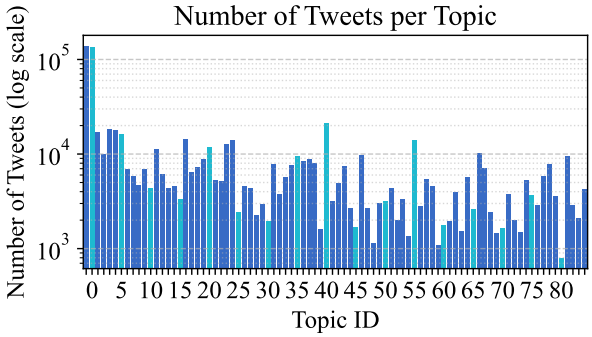


Fig. 2. Number of Tweets in Each Topic

*RoBERTa*

RoBERTa (Robustly Optimised BERT Approach) is a variant of BERT that improves performance with a refined pre-training procedure [14]. Retaining BERT's transformer architecture, RoBERTa introduces dynamic masking, larger batch sizes, longer input sequences and removes the Next Sentence Prediction objective. For this project, RoBERTa was implemented with the HuggingFace `roberta-large-mnli` model within a zero-shot classification pipeline to infer the occupation of Twitter users from their profile descriptions. Profile descriptions were preprocessed and normalised before being passed through the model with the hypothesis template *"This person works as a "*, performing both single-label (most probable occupation) and multi-label (all plausible occupations above a threshold) classification.

The effective categorisation of Twitter users by occupation provides a powerful tool for analysing the discourse surrounding the 2024 U.S. election, particularly when combined with results from other NLP models. Identifying the professions of users offers an important context for interpreting attitudes toward specific topics, allowing a more holistic analysis of election discourse. Furthermore, incorporating occupational demographics improves the study's ability to validate findings against external contextual indicators by allowing deeper insight into which groups are driving particular narratives.

Figure 3 displays the box plots of the 10 top occupations inferred from the user bios by RoBERTa. Notably, the high confidence scores for journalists (median $> 0.98$), politicians and podcasters demonstrate RoBERTA's ability to identify politically engaged professions, this is likely due to distinctive terminology in their Twitter bios. The wider confidence distributions for freelancers and retired individuals suggest greater variability in how these users self-identify, potentially reflecting diverse engagement patterns with election discourse that doesn't align with their professional identity. The confidence patterns validate the methodology of using zero-shot classification with RoBERTa for occupational inference of the electorate on social media during a highly contentious election cycle.

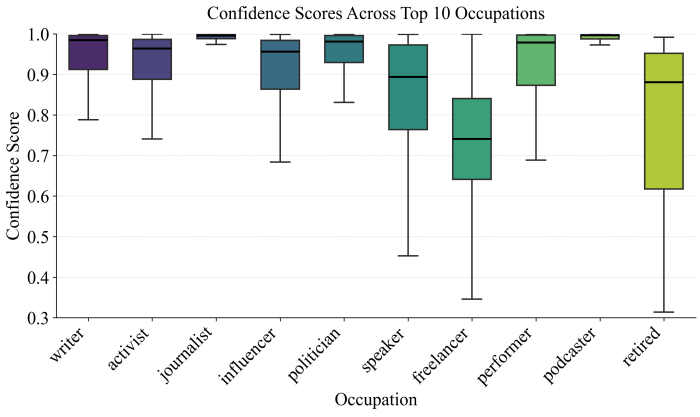


Fig. 3. Confidence score distributions across top 10 most common occupations in Twitter user bios extracted by RoBERTa

*BERTweet*

BERTweet is the first large-scale language model pre-trained specifically on English-language tweets. It uses the same architecture as BERT-base and uses the same pre-training procedure as RoBERTa. Trained on 850 million tweets, BERTweet is designed to handle the unique characteristics of Twitter data, including short length, informal grammar and irregular vocabulary [13]. As a result of the focus of the training data, BERTweet outperforms the other more general purpose models BERT, ELECTRA, RoBERTa and RoBERTuito on SemEval 2017 task 4 *'Sentiment Analysis in Twitter'* [24] with a mean Macro $F_1$ Score of 72.0. Further owing to the model's tweet specific training data, irony detection using BERTweet radically outperformed the other English language irony detectors, with a mean Macro $F_1$ Score of 80.8 on the SemEval 2018 task 3 *'Irony Detection in English Tweets'* [25]. The social media opinion mining toolkit Pysentimiento [17] was used as a wrapper for the BERTweet model to conduct sentiment analysis and irony detection.

Tweets were preprocessed as detailed in Section II and then inference was run on the cleaned tweets for both sentiment analysis and irony detection. Sentiment analysis is a classification task designed to determine the overall tone of a text. The outputs are split into 3 classes: positive (POS), neutral (NEU) and negative (NEG). Additionally, irony detection performs binary classification on text to determine whether the meaning of the text should be taken literally or if some alternative meaning is implied. Irony detection is of particular importance for social media data as posts are often written in a satirical tone [26].

Sentiment on its own provides very little insight into the discourse around the election. For a more meaningful analysis, it was combined with topic modelling to get representative plots of attitudes towards particular hotbed election issues. Additionally, inferred occupation provided the opportunity for analysis on subsets of the electorate. Since sarcasm and irony are often used to 'vent frustration' [26], the combination of irony labels with inferred occupation and topic labels can identify who and what constitutes the most emotionally charged political discourse on Twitter.

*M3Inference*

Following sentiment and irony analysis, M3Inference was employed to infer demographic characteristics of Twitter users. M3Inference is a demographic prediction model designed to infer age, gender, and organisation status from Twitter profiles [10]. While the original M3Inference model was developed for multilingual and multimodal input, including user profile pictures and names across 32 languages. As prefaced in Section II, this study looks solely at English tweets and does not utilise profile pictures, in adherence to ethical guidelines.

M3Inference integrates a DenseNet architecture for image analysis and a bi-directional LSTM for text processing, employing modality dropout to maintain robust predictions even when one of the inputs is missing. Co-training methods are used, where confident predictions from one modality are used to train the other, enhancing overall model performance. Model validation on crowdsourced datasets across multiple languages has shown M3Inference to outperform traditional tools such as Face++ and Microsoft's Face API for gender and age classification, and to outperform text-only demographic models like GenderPerformance and Demographer.

M3Inference's textual capabilities were utilised to classify user demographics into broad categories: age ($\leq$18, 19–29, 30–39, and 40+), gender (male, female), and organisation status (person, organisation). The detection of organisations is particularly valuable, allowing non-human accounts to be filtered from analysis focused on electorate behaviours. It is important to note that increasing the confidence threshold from $80\%$ to $95\%$ resulted in a 69.16% reduction in the number of tweets classified, highlighting the trade-off between prediction confidence and available sample size.

While the model was designed with ethical practices such as limited reliance on PPI, this trade-off could theoretically impact classification accuracy. Nevertheless, M3Inference's strong performance validates its use in large-scale social sensing tasks, enabling scalable, reasonably representative population estimates from Twitter data. Furthermore, the inferred demographics enable post-stratification, allowing demographic biases in the Twitter sample to be corrected and offering more meaningful insights when analysing sentiment, irony, and political discourse across different segments of the electorate.

Although the four defined age brackets offer a practical balance between granularity and classification confidence, it is worth considering whether these bins are sufficiently fine-grained to fully capture nuanced political behaviours across different generations. Future work might consider alternative categorisations if more detailed analysis is required.

## V. Findings

This section evaluates each model's findings either individually or in conjunction with others. Sub-sections V-C, V-E and V-F use real world indicators or data sources to validate the findings.

### A. BERTopic Results

Figures 4 - 6 show a subset of the topics identified by BERTopic.

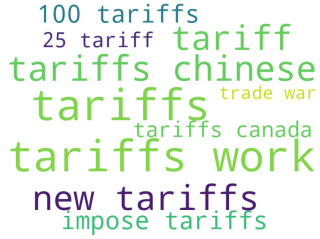


Fig. 6. Topic 22 - Tariffs

Topic 22, shown in figure 6, shows the model identifies specific policy issues and is able to contextually separate *trade war* from *war* from topic 0 - (militarised) war.


Fig. 4. Topic 0 - War

Topic 0 was the largest non-outlier topic, containing 17.5% of all processed tweets. These tweets generally relate to war, and given the ongoing conflicts in Ukraine and Gaza, has strong word weightings for these countries and their leaders (Putin and Netanyahu).


Fig. 5. Topic 2 - UK Politics

Topic 2 includes UK political parties and figures. Although the dataset covers the UK 2024 election period (4th of July 2024) this topic was somewhat unexpected, especially given that the tweet scraping was based of American political hashtags and themes. This highlights the lack of geospatial fields in the data, which prevented the isolation of geographically American based tweets.

This set of topics demonstrates BERTopic has successfully isolated meaningful topics and understands context over simple word frequency and co-occurrence. Another notable topic is the "Bucks County" topic (40). This topic was the third largest (with 2.5% of the total tweets). Bucks County (of Pennsylvania) is important in the context of the U.S. election and was dubbed by the Independent 'the swingiest of all swing counties in the swingiest of all swing states' [27]. Swing states often attract large campaign attention, as they often determine the national election result. Pennsylvania broke records in 2024 with $1.2 billion in political advertisement spending [28], the largest recorded for any state and 12% of the national spend.

### *B. Negative Bias in Sentiment*

Using the sentiment analysis results from BERTweet [13] inference, the changes in positive and negative attitudes were plotted over the whole dataset. These sentiment trends were later incorporated into the results from other models to aid with analysis. The content and tone of the social media landscape are highly volatile, with radically different volume, topic, and polarity of posts day to day. Therefore, to see the general trends over the noise, the overall proportion of positive/negative/neutral tweets in a single day was calculated and smoothed using the EMA (exponential moving average). The EMA acts to distinguish long term trends from the high inter-day variance.

Inspection of Figure 7 shows that negative tweets were considerably more prevalent across the 2024 election. Polarisation has been studied on tweets during the 2021 election in Germany using more classical techniques for sentiment analysis. Similar results were found, with more than 70% of tweets negative [29] and only 19.8% positive. Social media polarisation in politics bridges cultural and linguistic gaps, with positive sentiment generally staying between 10 and 15% and the more prevalent negative sentiment making up around half of all election tweets for this project. One exception appears to be towards the actual election day itself, highlighted by the dashed black line on the graph. A temporary reduction in negative sentiment is observed.

This spike in the proportion of positive election sentiment highlights an unusual phenomenon; negative sentiment is more volatile, while positive sentiment stays relatively stable until the election, making up almost 30% of tweets. In the following days, there is a quick return to the status quo, with both positive and negative sentiment returning to levels prior to the election. Two possible explanations for such a phenomenon are; 1 - the true shift in sentiment is realised by an overwhelming upswing in victory tweets from supporters, or 2 - that the change in language after the election win may include more words that are associated with positive sentiment (win, celebrate, victory etc.) and cause a statistical bias towards positive classification.

Investigating the latter, inference was conducted using the BERTweet model on the phrase *'Donald Trump wins the U.S. General Election'*, returning positive sentiment with a 0.836 confidence score even though such a statement is purely informative and as such should be perceived as neutral. This highlights some of the potential limitations of BERT models for sentiment analysis on political tweets, since there may be a high volume of similarly informative neutral statements focusing on the election result tweeted by journalists and news outlets that would have been classified as positive. The responsibility of each of these two factors for the spike in positive/negative sentiment is unclear; however, this still marks an interesting observation owing to both its transience and significant effect size.

| Week Start | Title | Assigned Topic | Top Tweet Topic |
|---|---|---|---|
| 2024-06-01 | "Republican voters have one last chance to rebuke Trump" [31] | -1:Outlier topic | 55:Free speech |
| 2024-06-08 | "Washington 'lunatics' seek regime change in Russia – former Trump rival" [32] | 38:American democracy threat | 2:UK Politics |
| 2024-06-15 | "Former KKK Leader David Duke Joins Anti-Israel Protesters, Says He's 'Saving Our Country From Jewish Supremacy' [33] | 4:Black community | 40:Bucks County |

TABLE II
SUBSET OF HEADLINES USED FOR TOPIC COMPARISON

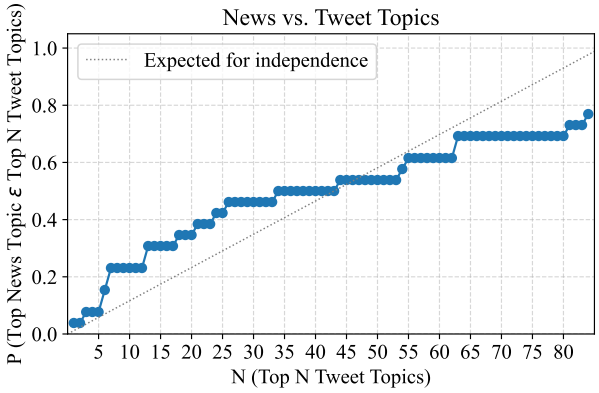


Fig. 8. Graph to Show the Probability that the Top News Topic was in the Nth Most Prevalent Topic of the Twitter Data

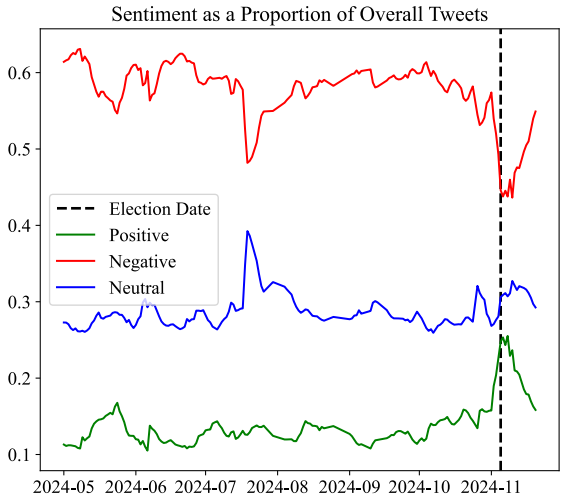


Fig. 7. Sentiment Over the Whole Dataset Smoothed with EMA, Plotted as a Proportion of Total Tweets on a Single Day

### C. News Headline Topic/Tweet Topic Comparison

As another method of validation, the top headlines for each week between 2024/06/01 - 2024/11/30 were scraped using TheNewsAPI [30]. Each headline was assigned a topic based on the highest cosine similarity between the headline's embeddings and the topic representations. A sample of this news data and its assigned topics are presented in Table II.
To compare the topics assigned to the tweets of each week, and the topics assigned to the headlines; the probability that the Nth most common tweet topic (of a week) was the same as the topic assigned to the news headline (of a week) is plotted in Figure 8. The dotted line in this figure shows the expected distribution if there was no association between the topic assigned to the tweet and the topic assigned to the news headline. The data points are shifted to the left of this line, suggesting that the top news topic (of a week) is more likely to occur in the most prevalent tweet topics (of a week). However, using the chi-squared test to conduct a hypothesis test ($H_0$: The top news topic and top tweet topic are independent) returned a p-value of 0.1641 ($<$0.05) therefore the null hypothesis cannot be rejected, showing that the top news topic and tweet topics are not dependent. This demonstrates that the discourse from media outlets and discourse on Twitter are distinctly different.

### D. Occupation Topic Dependency

Combining the occupational inference from *RoBERTa* and topics from BERTweet allowed greater understanding of the distribution of political discourse across different occupations. Figure 9 contains the top 10 occupations and their 8 joint most common topics. The clearest insight from this figure is that all these occupations had a strong bias towards tweeting about topic 0 (war), this supports the observation that the discourse on Twitter is highly distorted towards the most extreme topics of conversation, with the content delivery algorithms maximising the attention of users. For example, Milli et a demonstrated that Twitter's engagement-based algorithm significantly amplifies emotionally charged and partisan content, particularly expressing anger and outgroup hostility [34]. Next, it is seen that politicians had a tenancy to talk far more than other occupations about Bucks County (Pennsylvanian). This suggests that politicians may have catered their tweets to target the most important voters in the election, in the swing state, recognising that these votes will likely determine the overall result. Finally, it is noticed that influencers and freelancers are much more likely to tweet about crypto, this aligns with the trend in 'meme coins', which rely on social media hype to drive value. To understand the role of Twitter on these cryptocurrencies further analysis is performed in the following section C.

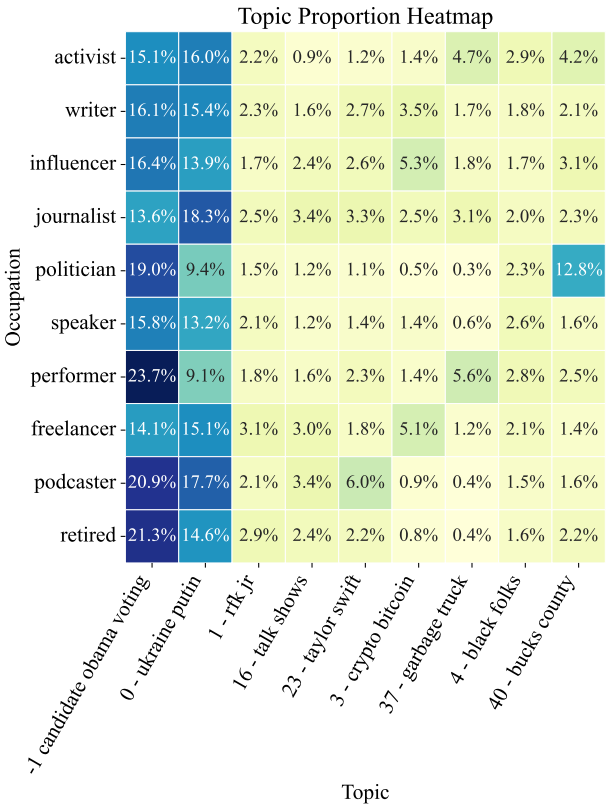


Fig. 9. Heatmap to Show the Percentage of a Given Occupations That Tweets Are Categorised Into Each Topic

### *E. Crypto Currency Sentiment Analysis*

Cryptocurrencies such as Bitcoin and Ethereum are emblematic of the information age and have become intertwined with political and financial discussion. This intersection was most clear with the release of 'Trump coin' and 'Melania coin', these cryptocurrencies were branded by many as a scam, drawing on patriotic and polarised voters to inflate the coins values. Both coins have slumped in value since their initial spike, which often occurs as the creators sell large stakes in the coin. This has not been shown directly in the case of Trump coin and Melania coin however, the anonymous nature of cryptocurrencies allow complex strategies to obfuscate such activity [35]. The BERTopic implementation returned a *crypto* topic, labelling 18,427 tweets as relating to cryptocurrencies. Merging these tweets with results from the sentiment analysis and historical bitcoin price data reveals a remarkable result. Plotting the positive, negative and neutral sentiment suggests a time delayed correlation 10. By time shifting the bitcoin price by up to 30 days find the strongest correlation with a 25 day time delay from inverse negative sentiment (as seen in Figure 11). The correlation with positive is at a 5 day delay. The Pearson correlation coefficient (PMCC) is used and the significance level is adjusted at different time shifts as the number of daily tweets ($n$ - number of samples) is variable. The correlation values go from statistically significant and positive to statistically significant and negative. This shows Twitter discourse is positively correlated in the short term but

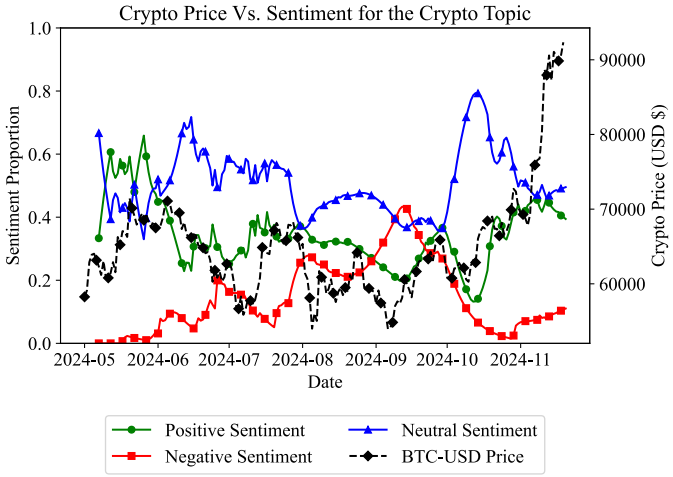


Fig. 10. Sentiment of the Crypto Topic Against Bitcoin price

negatively correlated in the long term. A causal conclusion cannot be drawn about the tweet sentiments effect on the Bitcoin price; however, this relationship offers an interesting avenue for future research.

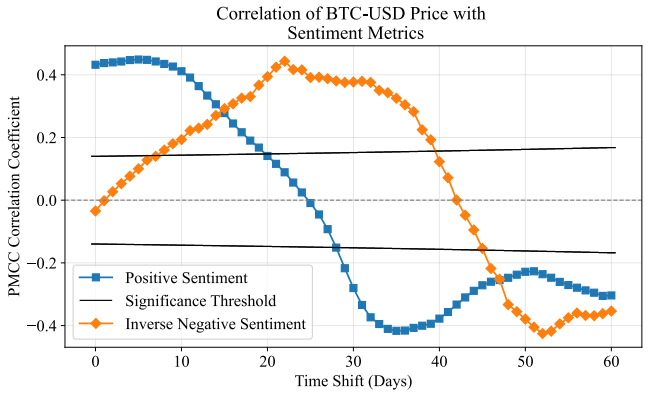


Fig. 11. PMCC Correlation Between Time Delayed Sentiment Metrics and Bitcoin Strike Price - The Significant Threshold Lines At (p=0.05), they are Non-linear as Number of Daily Samples Changes

### *F. Economic Indicator Validation*

One method used to evaluate the effectiveness of the combined NLP models was through comparison with external polling data. In this study, sentiment trends extracted from BERTweet related to the economy (Topic 5) were compared with the U.S. Consumer Confidence Index (CCI) [36]. The CCI is an economic indicator that quantifies consumer attitudes toward both current conditions and future expectations. It is constructed through monthly surveys of households using standardised questions about business conditions, employment prospects and household finances. Survey responses are aggregated and normalised to produce a single index value, where a score of 100 represents a neutral outlook, values above 100 indicate optimism, and values below 100 reflect pessimism.

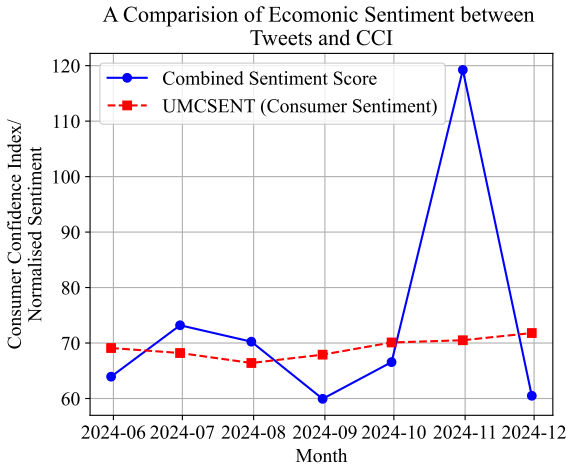


Fig. 12. Figure plotting the normalised sentiment against Consumer Confidence Index (CCI)

To compare twitter sentiment to CCI, a sentiment index was calculated using the formula: Sentiment Index $= (P - N) \times 100 + 100$, where $P$ represents the proportion of positive tweets and $N$ represents the proportion of negative tweets. This centred the sentiment index at 100, mirroring the CCI. Figure 12 compares the Twitter-derived sentiment index and the CCI across the election period. While both measures generally exhibit a negative attitude towards the economy, the CCI remains relatively stable with only minor fluctuations, whereas Twitter sentiment shows considerably greater volatility, including a notable spike in optimism around election day.

Overall, the CCI presents a slightly different narrative compared to the sentiment trends derived from Twitter. Sentiment around the economy remains consistently negative in the CCI data, but with relatively little variation over time (see figure 12 ). However, this does not undermine the ability of the combined NLP models to capture political discourse during the 2024 U.S. election, given the fundamental differences in how the data is collected. Twitter discourse reacts to events in real time, resulting in sharper fluctuations in sentiment, and analysis on occupation using RoBERTa showed a high proportion of politically active users, likely contributing to more extreme views than those reflected in traditional household surveys.

### G. Irony Analysis by Topic/Occupation

It is commonly understood that irony and sarcasm perform quite a unique social function. Sarcasm in particular (sarcasm and irony are seen as the same class in this work) is interpreted as signalling group membership to promote 'group solidarity' [26] that is frequently encountered in polarised political debate. The plot in Figure 13 details how the prevalence of irony differs depending on the topic in question. With over 60% of tweets being classified as ironic, topic 37 'calling supporters garbage', has the highest presence of irony out of all topics. This topic corresponds to the event of Joe Biden referring to Trump supporters as 'garbage' [37]; it is trivially clear that this topic sparked substantial outrage amongst twitter users and, given its one-sided nature, supports the notion of irony as a fundamental characteristic of emotionally charged, heavily polarised political discourse. The other two topics with high levels of irony (61, 80) correspond to 'trump tax tips' and 'insulin costs' (resp.) which are both very personal and relate to the hotly discussed political issues of low income inequality and the American healthcare system.

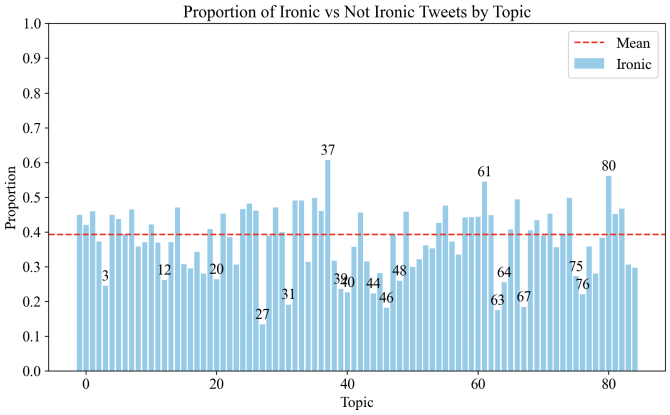


Fig. 13. Bar Chart Displaying the Proportion of Ironic Tweets by Topic, Topics of Most Significant Deviation are Labelled

Irony and occupation were separated in Figure 14 to identify how each occupation interacts with political issues on twitter. One obvious insight relates to the lower levels of irony expressed by journalists, this could be a result of their use of twitter as an extension of their journalistic platform; preferring to make professional comment and observation rather than informal discussion and debate. One of the recent trends ushered in by social media is the existence of politicians presenting themselves as 'accessible, relatable and authentic individuals' [38]. This is evidenced in Figure 14 with the occupation *politician* falling only very slightly below the mean for ironic tweets, suggesting only a slight difference in formality between the discourse of politicians and others, a trend successfully found with by the models in this work.

### H. Demographic Inference

The results of the age inference from M3Inference are displayed in Figure 15. These provide useful context when considering the other results. For example, in section V-F, consumer confidence index is analysed; which is found by sampling the whole working population. However, the normalised consumer confidence index derived from the tweets is sampled from this biased age distribution. This may contribute to the disparity between the two metrics. Both gender and organisational status were inferred at the 95% confidence level. This resulted in 63% of tweets labelled as male and 30% as female. Also, 80 % were labelled as people and 23% labelled as organisations. The remaining tweets were not within the 95% confidence threshold. Combining these results with other

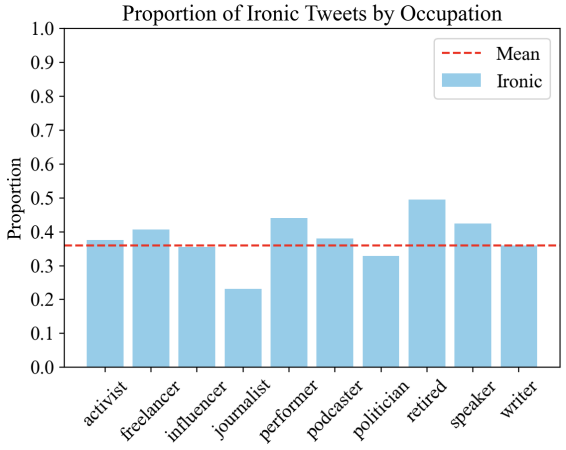


Fig. 14. Bar chart displaying the proportion of ironic tweets by most prevalent occupations

models would be beneficial but was out of the scope of the project.

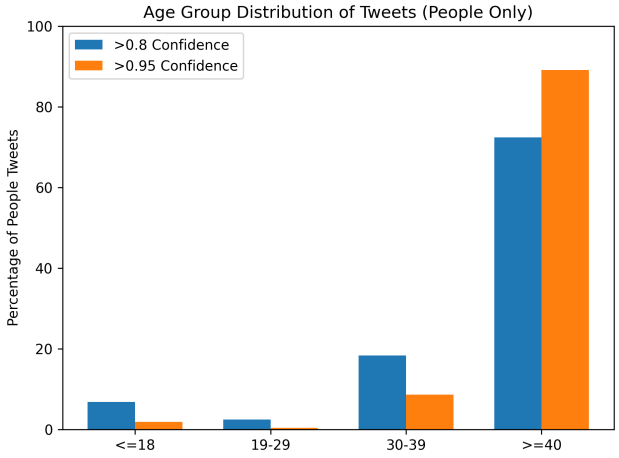


Fig. 15. Age distribution of Twitter users inferred using M3Inference, comparing demographic estimates at two confidence thresholds ($> 80\%$ and $> 95\%$).

## VI. Discussion and Conclusion

The analysis conducted in this paper aimed to validate the use of a combination of NLP methods against real world indicators. From the findings, this approach was broadly successful, identifying both relevant and nuanced trends. The combination of NLP models was particularly insightful for focused analysis, where patterns were magnified when amalgamating topics, occupations or events. Each model was selected for its specialised performance on the task in question. However, this approach limited the confidence associated with the results as the intersection of inferred characteristics compounds model inaccuracy.

Further limitations were introduced by uneven temporal sampling of the dataset, creating periods of reduced coverage, where events may be under-represented in the inference results. Also, the lack of geospatial fields in the dataset prevented regional analysis and required the assumption that all English tweets originated in the U.S. (exemplified by the UK politics topic 5). This reduces the accuracy of the findings when incorporating U.S. based indicators. It is important to note that since the acquisition of Twitter (Now X), API access has become highly restricted, X charging $42,000 per month for enterprise API access. The detrimental impact this change has had on research spanning 14 disciplines is discussed further in [39].

### A. Future Work

NLP models are integral to the findings of this paper. Future work should consider using the more resource-intensive LLMs, to capitalise on performance gains, reinforcing intersectional confidence levels (ie. when combining multiple probabilistic models). For example, BERTopic facilitates the integration of models such as llama [40] and Deepseek [41] for improved topic allocation. Due to the resource intensity and the size of the dataset, these larger models were not used.

Another avenue for future work is to explore datasets from different countries and languages, allowing greater demographic coverage. For example, if Spanish language tweets were modelled using language specific BERT adaptations such as RoBERTuito [17]; this would likely include more tweets from the 13% of the U.S. population that use Spanish as their first language [20]. This is important to allow more diverse topics and opinions to be identified, avoiding demographic blind spots.

An additional area for further work includes the detection of social media bots. Bots present a significant challenge when analysing political discourse, adding noise and bias to results. A previous study estimated that anywhere between 9-15% of the users on Twitter are bots [42]. Whilst this project attempts to remove the influence of bots by filtering out tweets with fewer than 5 likes, it is limited in its effectiveness. A less naive approach would be to use an NLP tool to filter for bots such as LMBot [43], this would add an additional dimension for intersectional model analysis.

## ACKNOWLEDGMENT

We would like to thank our supervisor Panagiotis Soustas for supporting us during this project and the unit directors Nirav Ajmeri and Seth Bullock.

## REFERENCES

[1] Statista, "Number of worldwide social network users from 2017 to 2027," 2024.
[2] Pew Research Center, "News platform fact sheet," September 2024.
[3] ——, "How americans navigate politics on tiktok, x, facebook and instagram," June 2024.
[4] BBC News, "Musk takes control of twitter and fires top executives," October 2022.
[5] Brennan Center for Justice. (2024) Online ad spending in the 2024 election topped $1.35 billion. Brennan Center for Justice.
[6] F. Yang, J. Groshek, and M. W. Ragas, "A scoping review on the use of natural language processing in research on political polarization: trends and research prospects," *Journal of Computational Social Science*, vol. 6, no. 2, pp. 707–739, 2022.
[7] R. Chandra and R. Saini, "Biden vs trump: Modeling us general elections using bert language model," *IEEE Access*, vol. 9, pp. 128 494–128 505, 2021.
[8] A. Unknown, "Sentiment analysis for predicting election outcomes in india," in *ITM Web of Conferences*, vol. 68, 2024, p. 03008.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018.
[10] Z. Wang, S. A. Hale, D. Adelani, P. A. Grabowicz, T. Hartmann, F. Flöck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. Association for Computing Machinery, Inc, 5 2019, pp. 2056–2067.
[11] A. Balasubramanian, V. Zou, H. Narayana, C. You, L. Luceri, and E. Ferrara, "A public dataset tracking social media discourse about the 2024 u.s. presidential election on twitter/x," vol. 1, 11 2024.
[12] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.
[13] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," 2020.
[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.
[15] Information Commissioner's Office, *Guide to the General Data Protection Regulation (GDPR)*, August 2018, version 1.0.248.
[16] X. Documentation, "Counting characters," n.d.
[17] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. V. Martínez, "pysentimiento: A python toolkit for opinion mining and social nlp tasks," 2023.
[18] A. Nourbakhsh, X. Liu, S. Shah, R. Fang, M. M. Ghassemi, and Q. Li, "Newsworthy rumor events: A case study of twitter," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 27–32.
[19] S. Rönnqvist, O. de Gibert, and J. Tiedemann, "Multilingual probing of bert representations," *arXiv preprint arXiv:1910.03806*, 2019.
[20] S. J. U. S. District, "Character counts program," n.d.
[21] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.
[22] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.
[23] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, ser. Lecture Notes in Computer Science, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., vol. 7819. Springer, Berlin, Heidelberg, 2013, pp. 160–172.
[24] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518.
[25] C. Van Hee, E. Lefever, and V. Hoste, "SemEval-2018 task 3: Irony detection in English tweets," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 39–50.
[26] M. Sykora, S. Elayan, and T. W. Jackson, "A qualitative analysis of sarcasm, irony and related #hashtags on twitter," *Big Data & Society*, vol. 7, no. 2, p. 2053951720972735, 2020.
[27] R. Hall. (2024, November) In a purple county in pennsylvania, residents say 'it's been insane' as they go out to vote.
[28] T. Dias. (2024, December) In review '24: Pennsylvania saw a record breaking $1.2b in election ads.
[29] Y. Zhang, B. Zhou, Y. Hu, and K. Zhai, "From individual expression to group polarization: A study on twitter's emotional diffusion patterns in the german election."
[30] The News API. (2025) The news api.
[31] S. D. Percio. (2024, June) Republican voters have one last chance to rebuke trump.
[32] R. News. (2024, June) Washington 'lunatics' seek regime change in russia – former trump rival.
[33] A. Staff. (2024, June) Former kkk leader david duke joins anti-israel protesters, says he's 'saving our country from jewish supremacy'.
[34] S. Milli, M. Carroll, Y. Wang, S. Pandey, S. Zhao, and A. D. Dragan, "Engagement, user satisfaction, and the amplification of divisive content on social media," *arXiv preprint arXiv:2305.16941v6*, Dec 2024.
[35] P. D. Huynh, S. H. Dau, H. Y. Tran, N. Huppert, H. Sun, J. Cervenjak, X. Li, and E. Viterbo, "Serial scammers and attack of the clones: How scammers coordinate multiple rug pulls on decentralized exchanges," *arXiv preprint arXiv:2412.10993*, 2024.
[36] F. R. B. of St. Louis, "University of michigan: Consumer sentiment (umcsent)," n.d.
[37] J. Fitzgerald. (2024) Biden tries to clarify 'garbage' comment after uproar.
[38] R. L. B. D. V. A. X. M. Manning, Nathan; Penfold-Mounce, "Politicians, celebrities and social media: a case of informalisation?" *Journal of Youth Studies*, vol. 20, no. 2, pp. 127–144, 2016.
[39] R. Murtfeldt, N. Alterman, I. Kahveci, and J. D. West, "Rip twitter api: A eulogy to its vast research contributions," 2024.
[40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, E. Rodriguez, E. Grave, A. Joulin, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[41] DeepSeek-AI, "Deepseek-v3 technical report," 2024.
[42] S. Cresci, M. L. Cava, F. Martino, and S. Tesconi, "Online human-bot interactions: Detection, estimation, and characterization," in *Proceedings of the AAAI Conference on Web and Social Media*, vol. 14, no. 1, 2020, pp. 56–67.
[43] Z. Cai, Z. Tan, Z. Lei, Z. Zhu, H. Wang, Q. Zheng, and M. Luo, "Lmbot: Distilling graph knowledge into language model for graph-less deployment in twitter bot detection," *arXiv preprint arXiv:2306.17408*, 2024.