The Effect of Platform Amplification on Social Media Posts

ADEYE Abiola, BRUNO Etienne
Distributed Information Systems Laboratory (LSIR)
École Polytechnique Fédérale de Lausanne
June 2023

Abstract
This study aims to investigate and perform a cross-cultural analysis of various Twitter datasets encompassing French celebrities, French politicians, US politicians, US celebrities, and a random sample set. The objective is to identify key factors that influence high engagement rates, quantified through likes, quotes, retweets, and impression counts. The steps undertaken include data acquisition from Twitter, followed by comprehensive data exploration and visualisation. A bespoke pipeline was developed to facilitate regression analysis and to identify the most relevant features that contribute to heightened engagement. The findings from this study will shed light on the efficacy of platform amplification in social media posts, particularly in varying cultural and public figure contexts. The study underscores the importance of understanding the unique characteristics of different audiences and stakeholders, promoting tailored and effective communication strategies in the ever-evolving landscape of social media.

Contents

1 Introduction

In the present digital era, social media sites function not only as hubs for entertainment and social engagement but also as vast, dynamic repositories of knowledge. A myriad of public figures, including politicians and celebrities, command significant attention on Twitter, resulting in their tweets becoming substantial drivers of public discourse. These tweets, often subject to numerous likes, retweets, displays, or quotations, deploy significant influence. Recognizing the potential power and impact of these profiles is therefore paramount.

The primary objective of this research paper is to unearth key elements that stimulate high engagement rates, measured via metrics such as likes, quotes, retweets, and impression counts on four datasets that span French and American celebrities and politicians. The research

methodology involves data acquisition from Twitter, comprehensive data exploration, and subsequent visualization. A custom-built analytical pipeline is utilized to perform regression analysis and identify features that lead to enhanced engagement levels. It is later compared to more advanced models.

Guiding this study is the specific research question: "Which quantifiable attributes of a tweet, and the associated Twitter profile, most significantly correlate with high user engagement rates within the contexts of French and American public figures?"

The anticipated findings from this research are set to offer valuable insights into the effectiveness of platform amplification in social media posts, particularly within varied cultural and public figure contexts.

2 Related Work

Review of pertinent literature that contributed to our understanding and approach

2.1 The Half-Life of a Tweet

The paper "The Half-Life of a Tweet" by Jürgen Pfeffer, Daniel Matter, and Anahit Sargsyan[2] from the Technical University of Munich delves into the lifespan and dissemination process of a tweet by tracking its impression count over time. They use the following concepts:

- **Impression Count:** The impression count is a metric provided by Twitter that shows how often a particular tweet has been displayed to Twitter users at the time of data collection. This metric is crucial for understanding the reach and impact of a tweet.
- **Half-Life of a Tweet:** The authors introduce the concept of the half-life of a tweet, which is the time it takes for a tweet to accumulate half of its total impressions. They found that the median half-life of a tweet is approximately 80 minutes. This means that, on average, a tweet will receive half of its total impressions within the first 80 minutes of being posted.
- **Peak of Impressions:** The authors also analyzed the peak of impressions per second, which is the moment when a tweet receives the most impressions in a single second. They found that, on average, this peak occurs 72 seconds after a tweet is posted. After 24 hours, about 95% of all tweets do not receive a significant number of impressions, indicating that the relevance of a tweet decreases rapidly after it is posted.
- **Outlook and Ethical Considerations:** The authors discuss potential future research questions, such as how the half-life of a tweet might affect the strategies of businesses and influencers on Twitter, and how the concept of information half-life could be applied to other social media platforms.
- **Future Research:** The authors suggest that future research could focus on identifying the factors that drive view counts and half-life. They also mention the need to study the temporal interplay of times series with the number of views, the number of followers of the tweet senders, the tweet content, and possibly connected images and websites.

The suggested future research areas are directly related to the objectives of our study. Our research has centered on identifying key factors that impact Twitter engagement. The proposed areas for future research, align with our aim of discerning the mechanisms that lead to high engagement on the platform.

2.2 Just Another Day on Twitter

The research paper titled "Just Another Day on Twitter: A Complete 24 Hours of Twitter Data" by Juergen Pfeffer, Daniel Matter, and Kokil Jaidka[1] provides a comprehensive analysis of Twitter data collected over a single day in 2022. The paper presents a detailed exploration of various aspects of Twitter usage, including user demographics, language use, media attachments, geo-tagging, and the prevalence of bot accounts.

The authors found that the active accounts on the day of data collection had a mean of 2,123 followers, with six accounts having more than 100 million followers. The study also revealed that the 374 million tweets collected had been retweeted 401 billion times, indicating significant parts of historic Twitter get retweeted daily. The language analysis showed that 15 languages made up 92.5% of all tweets.

The study also analyzed the content on Twitter, finding that a large proportion of tweets referred to entertainment, which together comprised about 30% of tweets. Moreover, 79.2% of all tweets refer to other tweets, i.e. they are retweets or quotes of or replies to other tweets. Consequently, 20.8% of the tweets in the dataset are original tweets. The tweets with references are of the following types: 50.7% retweets, 4.3% quotes, and 24.2% replies.

The authors use several metrics to measure engagement. These include the number of followers for active accounts, the number of retweets for the collected tweets, and the bot scores for unique accounts. The bot scores indicate the likelihood of an account exhibiting bot-like behavior. The authors also analyze the content of tweets, including the language used and the categories of content covered by the top hashtags. These metrics provide a comprehensive view of engagement on Twitter, encompassing not only the number of interactions but also the nature of the content and the authenticity of the users involved.

In our research, we adopt similar metrics as those used in the aforementioned studies, employing them as features for our regression analysis and inputs for our other models. Furthermore, we extend our analysis to the meta-data of the tweet. This approach allows us to gain a more nuanced understanding of engagement on Twitter, as it takes into account not only the volume of interactions but also the nature of the content and the context of the tweets such as the time of the day when it was published or the number of hashtags used in the tweets.

By integrating these comprehensive metrics into our models, we aim to provide a more holistic and accurate analysis of Twitter engagement. This approach will enable us to get a better understanding of the dynamics of social media interactions and their implications for public discourse, strategic communication, and social media management.

3 Data Acquisition

Data for this research was derived from the public Twitter API, employing the command-line tool Twarc for extraction.

Given the potential for regional variances in subsequent observations, the data was segmented into distinct subsets. This study emphasized two categories of individuals typically boasting a high follower count and interaction rate on Twitter: politicians and celebrities. Recognizing the potential for cultural influences on engagement, these groups were further separated into French and American subsets. This bifurcation allowed for an in-depth exploration of regional differences that could potentially impact Twitter engagement patterns.

This required the preliminary task of manually curating lists of the most influential French and American celebrities and politicians. These lists served as the basis for extracting the relevant tweets from the Twitter API. The collected data, all curated between January 6th, 2023, and March 31st, 2023, underwent a standardization process to limit the number of tweets per user to a maximum of 1,000 in each subset.

To ascertain if the trends observed for specific users or regions were also discernible in the broader Twitter population, a collection of over 200,000 random tweets was also aggregated.

Data extracted from the Twitter API encompasses more than the tweets themselves, providing rich metadata about the users and their respective tweets. This includes but is not limited to, user location, beneficial for geographical and demographic analyses; the timestamp of each tweet, valuable for identifying temporal patterns; and the follower count of each user, offering insights into their respective influence and outreach.

The size of each dataset varied, ranging from several hundred megabytes to approximately 22

gigabytes. Given the considerable size and complexity of these datasets, we opted to employ Apache Spark for data exploration and processing.

4 Data Exploration

The primary objective of this project is to develop a predictive model that can accurately estimate engagement levels based on certain features. These features, which will be discussed in detail in the subsequent section, are carefully selected based on their potential influence on a tweet's visibility and engagement.

To begin our exploration, we have chosen to focus on the engagement and views of the tweets within our datasets. Engagement, in the context of this project, is defined as the sum of retweets, replies, likes, and quotes a tweet receives. Views, on the other hand, are quantified by the number of impressions a tweet makes. An impression is counted each time a tweet appears in a user's timeline, regardless of whether the user interacts with it or not.

Our analysis spans four distinct datasets. The diversity of these datasets allows us to account for a variety of factors and conditions, thereby enhancing the generalizability of our model.

To visualize the relationship between the variables of interest, we have generated two types of plots:

- Impressions as a function of followers: This plot aims to illustrate the relationship between the number of followers a Twitter account has and the number of impressions its tweets receive. The underlying hypothesis is that accounts with a larger follower base are likely to generate more impressions due to their wider reach.

- Engagement as a function of impression count: This plot seeks to elucidate the relationship between the number of impressions a tweet receives and the level of engagement it garners. The premise here is that tweets with higher impression counts have a greater likelihood of receiving more engagement, given the increased exposure.

The primary motivation behind creating these plots is to visually inspect and understand any clear trends or relationships between these variables.

4.1 Impression Counts and Followers

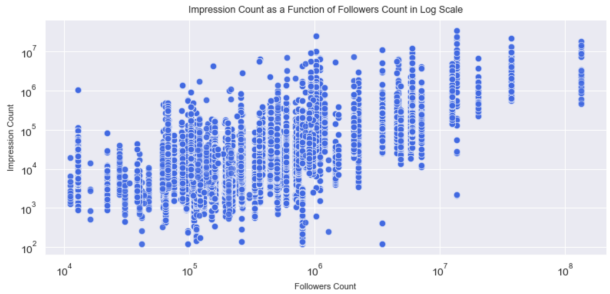

Figure 1: Log-Log Scatter Plot of Impression Count as a function of followers for the American Politicians dataset.

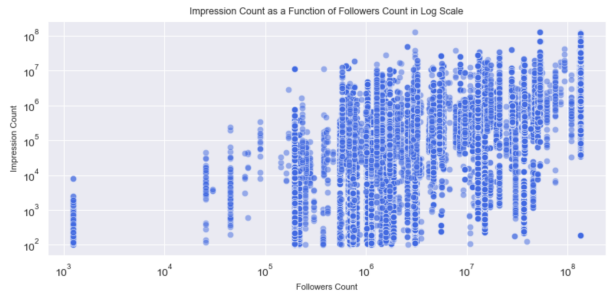

Figure 2: Log-Log Scatter Plot of Impression Count as a function of followers for the American Celebrities dataset.

From an examination of Figures 1 and 2, it becomes apparent that there exists a log-linear relationship between the count of followers and the count of impressions. However, it is important to note the presence of significant volatility in this relationship. This implies that even for a user with a consistent follower count, the impression count can vary greatly. Similarly, not all tweets from a given user achieve the same level of viral-

ity. This variability is visually represented by the vertical bars observed in the plots.

This pattern is not exclusive to the datasets represented in these figures; similar trends can be observed across other datasets as well. Detailed visualizations for these additional datasets have been included in the appendix for further reference and analysis.

4.2 Engagement and Impression Counts

To gain a deeper understanding of the relationship between various components of engagement and views, we have generated a series of plots for each dataset. In this section, we will focus on the plots for the American politicians' dataset, while the corresponding plots for the remaining three datasets can be found in the appendix.

This implies that the relationship between retweets and impressions is multiplicative rather than additive. In other words, a certain percentage increase in impressions is associated with a similar percentage increase in retweets. This suggests that as a tweet reaches a wider audience, the likelihood of it being retweeted increases proportionally, rather than at a fixed rate. This pattern indicates a positive feedback loop where more impressions lead to more retweets, which in turn can lead to even more impressions. This insight could be particularly valuable for understanding and predicting the virality of tweets.

We have generated analogous plots for the remaining components of our defined engagement metric, specifically Likes, Replies, and Quotes. These scatter plots also exhibit a multiplicative relationship when viewed on a log-log scale.

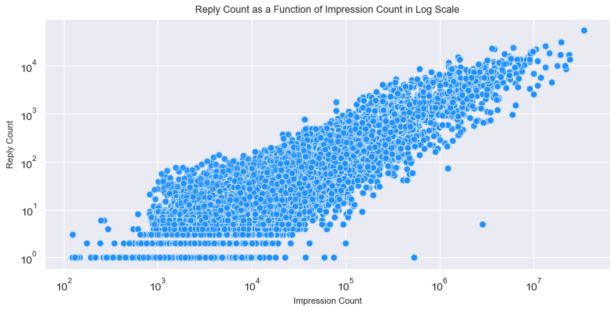

Figure 4: Log-Log Scatter Plot of Replies as a function of Impression Count for the American Politicians dataset.

In order to generate plots of Figures 3, 4, 5 and 6 it was necessary to implement a filtering process to exclude outliers. Specifically, we removed tweets with an impression count of less than 100 for the American datasets, and those with less than 50 impressions for the French celebrities and politicians datasets. This decision was driven by the high degree of variance observed in these tweets, which made them less reliable for our analysis.

Moreover, tweets with such low impression counts are less likely to be associated with public figures such as politicians or celebrities, who typically have a much larger audience reach. By excluding these outliers, we aimed to ensure that our analysis was focused on the tweets that were most representative of the engagement patterns of the public figures in our datasets. This filtering process helped to enhance the accuracy and relevance of our findings, providing a clearer picture of the relationship between various engagement metrics and impression counts.

Figure 3: Log-Log Scatter Plot of Retweet as a function of Impression Count for the American Politicians dataset.

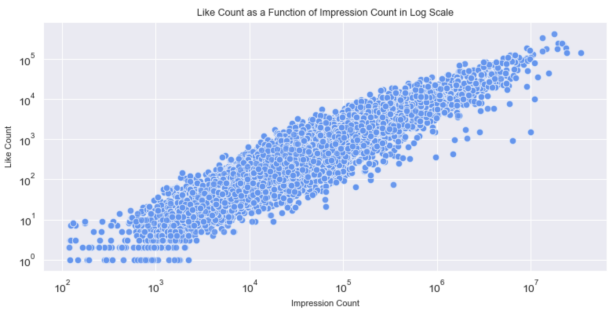

Figure 5: Log-Log Scatter Plot of Likes as a function of Impression Count for the American Politicians dataset.

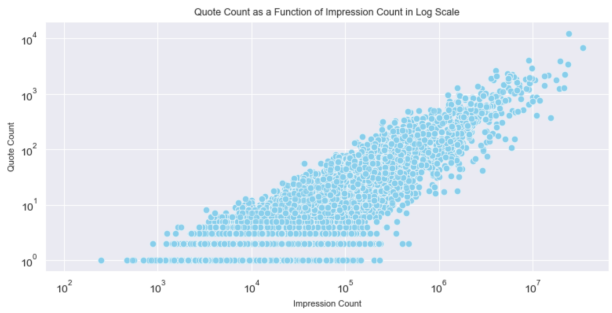

Figure 6: Log-Log Scatter Plot of Quotes as a function of Impression Count for the American Politicians dataset.

5 Data Processing

Our study employed a multi-tiered approach for parsing and processing our datasets.

5.1 Feature Extraction

The first step involved the extraction of key data components from each tweet. This encompassed details such as the tweet text, the time of creation, the number of accompanying hashtags, mentions count, external URLs, media items, followers count and the context annotations, if any.

5.2 Zero Shot Learning for Context Annotation

In the second stage of our analysis, we focused on the categorization of context annotations linked to each tweet. The intention was to assign each tweet to pre-existing clusters of annotations, thereby facilitating a thematic grouping of discussions. A portion of the tweets presented a unique challenge. Some did not contain context annotations that matched any of the established clusters, potentially due to the unique or emerging topics they addressed. To handle these instances, we turned to a method known as zero-shot classification.

Zero-shot classification is an algorithm that can make sensible predictions for inputs it has not explicitly seen during training. Essentially, it can understand and classify data it hasn't encountered before, making it an excellent tool for handling unique or unexpected data inputs. Here, we used zero-shot learning to assign context annotations to these 'anomalous' tweets, ensuring that they were still associated with one or more of the most prevalent annotation clusters.

We performed this zero-shot classification by leveraging the semantic similarity between the tweet content and the themes represented by the existing clusters. If a tweet's content was semantically closer to a certain cluster, even if it did not contain the exact context annotation, the algorithm would assign it to that cluster. This process ensured that we left no data ungrouped and could make the most of the information available, even in the absence of directly matching context annotations. Each dataset contained its own set of context annotations, each of which was represented as a feature.

5.3 Text Analysis

Subsequently, we moved on to the feature generation phase. Here, we produced additional features derived from the tweet text and other metadata. The generated features were sentiment score (obtained through sentiment analysis of the tweet text), tweet length as well as the hour that the tweet was posted.

Lastly, we utilized one-hot encoding on the context annotations. By transforming these categorical data into a binary format, machine learning algorithms can treat each unique context annotation as a distinct, independent feature.

At the end of this comprehensive process, we consolidated all the extracted and generated features into a structured data frame for each tweet. The finalized data frame included all of the above-encoded features.

6 Regression Analysis

Having gained a comprehensive understanding of our data distribution, we proceeded to clean, process, and store the data in easily manageable data frames. This prepared us to implement our initial model. The primary objective of this model is to predict the impression count of a given tweet based on a set of features. As previously described, we have, for each dataset, a data frame of tweets and their features. These features range from specific characteristics to dummy variables for the most frequent context annotations.

6.1 Strategies to a better model

Our analytical pipeline first regresses the impression count on these features. However, this raw regression did not yield efficient results for any of the datasets, as evidenced by a low R^2 value of around 0.15, high Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), and a significant number of features that were not statistically significant at the 5% level. Consequently, our goal shifted towards improving this regression model. To achieve this, we applied several strategies:

- **Dataset Filtering:** We filtered the dataset to eliminate tweets with an impression count below a certain threshold (defaulted to 100). These tweets were considered outliers, as it is unlikely for politicians and celebrities, who are the focus of our study, to have such a low impression count.
- **Log Transformation:** We explored the application of a logarithmic transformation to all variables. This decision was informed by our data exploration phase, where we observed a more coherent relationship between impression count and engagement metrics when both were presented on a log scale. Therefore, we decided to apply a logarithmic transformation to all variables to examine if this would lead to a better fit in the regression model.
- **Feature Selection:** In our final regression model, we applied both strategies described above and additionally retained only those variables that were statistically significant at the 5% level. This approach aimed to streamline our model by focusing on the most impactful features.

6.2 Logarithmic regression

The decision to apply a logarithmic transformation to all variables in our dataset is rooted in several key considerations (in addition to the previous observational study):

- Addressing Skewness: In many real-world datasets, especially those related to social media metrics, the data can be heavily skewed. This means that there are a few values that can distort the overall distribution of the data. A logarithmic transformation can help to reduce this skewness and make the data more symmetric, which is a desirable property when building regression models.
- Scaling Down Large Values: Logarithmic transformation has the effect of scaling down large values more than smaller ones. This can be particularly useful in our case where the range of values for metrics like impressions and retweets can be quite large. By scaling down these values, we can prevent them from unduly influencing the model.
- Linearizing Relationships: Another important reason for applying a logarithmic transformation is that it can help to linearize relationships between variables. By applying a logarithmic transformation, we can convert the multiplicative relationship uncovered in previous sections into an additive one, which can be better captured by a linear regression model.

These benefits can enhance the performance and interpretability of our regression model, making it a more effective tool for understanding the factors influencing tweet impressions.

6.3 Results

In the aforementioned pipeline, we ensured the robustness of our model by evaluating the R-Squared (R^2). It measures how well a model fits the data. by quantifying the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 is calculated as the ratio of the explained sum of squares to the total sum of squares, and it is commonly used in regression analysis to evaluate the goodness of fit of a model. For each dataset, we found that the model that incorporated both the logarithmic transformation and the filtering of low impression counts consistently performed better, as indicated by higher R^2 values close to 0.6 for each dataset. This suggests that these strategies were effective in improving the fit of our model to the data, thereby enhancing its predictive accuracy and reliability.

6.4 Interpretation of the coefficients in OLS

In a traditional linear model, the magnitude of the coefficient reflects the degree to which the dependent variable's average (impression counts) shifts in response to a unit change in the independent variable, assuming all other variables in the model remain static. This assumption of other variables being static is paramount as it provides an indication of the intensity of the relationship between a specific feature and the target variable.

An examination of the coefficients depicted in Figures 7, 26, 27, and 28 provides insights into the impact of various metrics on the engagement across all datasets. Prominently, the number of followers and the incorporation of media in tweets display positive coefficients, signaling a strong association with an enhancement in impressions. In contrast, some features either mitigate engagement or exert no discernible influence on it. Remarkably, the mention count registers a highly negative coefficient across all four regression models, implying an inverse correlation between the use of mentions and the tweet's impression count. Equally, the counts of hashtags and the sentiment of the tweet barely affect engagement, as evidenced by their near-zero coefficients. These results suggest that the excessive inclusion of hashtags or the sentiment derived from a tweet does not significantly influence the overall impression count across the datasets that we observed.

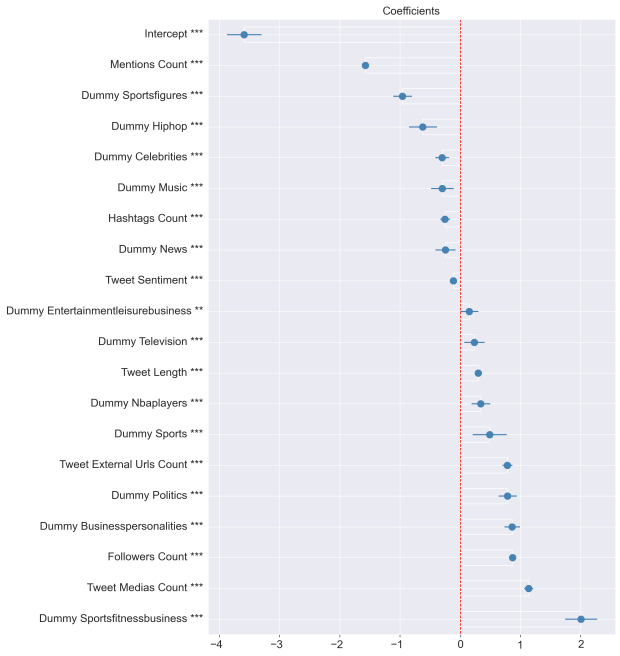

Figure 7: Coefficient plot of the regression using log-variables and cleaned dataset for the American celebrities.

Regarding the contextual annotations, it proves challenging to draw any concrete conclusions across all four datasets. Nevertheless, discernible patterns emerge within specific datasets, where certain annotations appear to have a statistically significant positive or negative impact.

The interpretation of the coefficients in our model demands careful consideration. While the observed coefficients and their significance levels - expressed by their p-values - offer valuable insights, they should be interpreted within the context of their original, untransformed scales. It is essential to understand that a higher or lower coefficient does not necessarily imply a higher or lower importance of a given feature. Moreover, due to our use of a log-log scale, the coefficients are not immediately readible in terms of absolute changes. Instead, they should be understood in

terms of elasticity. Specifically, a one-unit change in the independent variable (on a log scale) corresponds to a percentage change in the dependent variable. For example, if the coefficient of a predictor is 0.1, a 1% increase in the predictor leads to an approximate 0.1% increase in the response variable, all else being equal.

7 Other methods

In extending our analysis beyond the earlier logistic regression methodology, we decided to investigate a variety of alternative regression techniques to enhance the results derived from the preceding section.

Our investigation included the subsequent regression approaches:

1. Support Vector Machine Regression (SVM): This is a non-parametric approach that leverages the principle of structural risk minimization, providing efficient generalization performance. Its potential advantage lies in its ability to handle high dimensional spaces and manage non-linear relationships effectively, which may be relevant to our study.

2. Random Forest Regression: An ensemble learning method that utilizes multiple decision trees, where the final prediction is the average of individual tree predictions. The strength of this technique lies in its capacity for mitigating overfitting, handling outliers, and providing insights into feature importance, which could enrich our understanding of the underlying relationships.

3. Boosted Decision Trees Regression: A method that incrementally builds decision trees in a sequence designed to correct the predecessors' errors. Boosting may potentially enhance performance, specifically in scenarios where complex interactions between variables exist, given its unique ability to target and adjust errors iteratively.

4. Neural Network Regression: A flexible method that uses artificial neurons arranged in layers, trained to minimize the error between the predicted and actual value. This approach might be interesting due to its exceptional capacity for modeling complex, non-linear relationships and capacity to learn from large volumes of data, which could be highly beneficial for our complex dataset.

As for the testing framework: We partitioned each one of our datasets into a training set and a test set, with a ratio of 80 to 20, respectively. We executed each benchmark on three separate data splits, subsequently taking the average to compute the Mean Squared Error (MSE), thereby ensuring a robust evaluation of our model performance.

The results of the regression techniques applied to different datasets can be visualized in the following table:

MSE Comparison	American Celebrities	French Celebrities	American Politicians	French Politicians
Log-Linear	2.61	1.5	1.35	1.75
SVM	2.65	1.58	1.29	1.78
Boosted Decision Tree	1.83	1.09	0.95	1.49
Random Forest	1.4	1.1	0.12	1.4
Neural Network	2.14	1.54	1.37	1.7

Table 1: Performance of different regression methods across various categories

In summarizing the findings, the Random Forest Regression method delivered the most promising results across all categories, specifically exhibiting significant performance when applied to American politicians' data. The SVM and Neural Network methods, despite their comprehensive capabilities, did not surpass the efficiency of the Random Forest in our specific datasets.

8 Discussion

8.1 Limitations

Despite the extensive methodology and diverse models utilized in this study, several notable limitations should be recognized:

1. **Annotation Dependence**: The efficacy of the conducted analysis hinges on the precision and quality of context annotations. Incorrect or ambiguous annotations can compromise the accuracy of our models. The reliance on semantic similarity in the zero-shot learning technique to assign annotations to 'outlying' tweets also has its pitfalls, as it may not consistently encapsulate the actual context.

2. **Linear Assumption of Regression Models**: The regression models adopted in this research inherently presume a log-linear interrelationship between the predictors and the dependent variable. This log-linear assumption might not always align with reality, particularly in scenarios with intricate interactions between several variables, which may not be adequately captured by these models.

3. **Feature Selection Constraints**: Despite the inclusion of several pivotal features like sentiment score, tweet length, time of posting, and context annotations, numerous potential features may be omitted. These might include specific phrasing, tweet subject, or prevailing sociopolitical events at the time of posting, which could all influence tweet impressions.

4. **Bias in Twitter Data**: The demographic distribution of Twitter users does not mirror that of the general population, which could confine the applicability of our findings. Furthermore, the extrapolation of broader trends based on tweets from celebrities and politicians may not fully encapsulate the patterns of ordinary individuals or businesses.

5. **Limitations of Logarithmic Transformation**: Our use of log transformation, although beneficial in certain respects, could distort the relationships for certain variables and may not always be the most suitable transformation method depending on the distribution of data.

These limitations do not invalidate our findings but serve to provide a comprehensive understanding of the study's scope and pave the way for further explorations in this domain.

8.2 Possible Improvements

While our models have exhibited significant potential in predicting tweet impressions, several areas for improvement could enhance the accuracy and robustness of our findings.

1. **Time-Series Analysis**: The dynamics of Twitter engagement are likely impacted by various time-sensitive factors. Significant local or global occurrences like elections, holidays, sports games, or notable news items can significantly stimulate activity and interaction on Twitter. The timing of a tweet about such events can substantially influence its impression count. For instance, an election-related tweet will likely attract more impressions if shared on or around the election day as compared to being posted weeks or months before or after the event. In our research, we incorporated the continuous time of the day as one of the variables. However, it might be beneficial to conduct a more comprehensive time-series analysis. This would take into account not just the time of the day but also the wider context, including ongoing trends, event timelines, and changing user behaviors over time. Conducting such an analysis may reveal valuable patterns and offer deeper insights into the factors that drive tweet impressions, thereby enhancing the precision of our predictive models.

2. **Sentiment Analysis**: The emotion conveyed within a tweet could potentially influence its impression count. As such, augmenting our model to include a more sophisticated sentiment analysis could yield further insights. Our current approach utilizes the VADER sentiment analysis tool, which is rule-based. While VADER provides a good starting point for sentiment analysis, there may be benefits to exploring more advanced techniques. For instance, we could consider using the Bidirectional Encoder Representations from Transformers (BERT). BERT has the advantage of being able to understand the context of words in sentences, making its sentiment analysis potentially more nuanced and accurate. We may be able to gain deeper insights into the influence of sentiment on tweet impressions and thereby enhance the predictive capability of our models.

3. **User Behavior Modeling**: Finally, deeper modeling of user behavior could improve predictions. For example, clustering users based on their followers, tweet frequency, or other factors could lead to more nuanced models that consider the behaviors of specific user groups.

In conclusion, our models represent a significant step in understanding the complex dynamics influencing tweet impressions. However, there are numerous potential enhancements and areas for further investigation. By exploring these opportunities, we could develop even more precise and nuanced models, providing a deeper understanding of the factors influencing tweet impressions.

9 Conclusion

In the course of this research project, we undertook a comprehensive analysis of cross-cultural engagement on Twitter. We acquired and processed an extensive dataset, which included Twitter data from both American and French celebrities and politicians. Utilizing a custom-built analytical pipeline, we conducted exhaustive data exploration, followed by visualization and regression analysis. The primary objective of these efforts was to identify key variables that significantly influence high engagement rates.

Our regression models revealed that applying a logarithmic transformation to all variables, coupled with stringent filtering of tweets with low impression counts, consistently led to superior model performance across all datasets. This was evidenced by higher R^2 values. In addition to these techniques, we explored alternative regression approaches, with Random Forest Regression emerging as the most promising method across all categories.

While our findings underpin the relevance of several key features, such as the number of followers, media use, and some context annotations, the study also illuminates the richness and complexity of social media data. Despite the inherent limitations, our study unveils an innovative framework for investigating tweet impressions, providing a foundation for future research in this domain.

10 Appendix

Impression Counts and Followers Scatter Plots

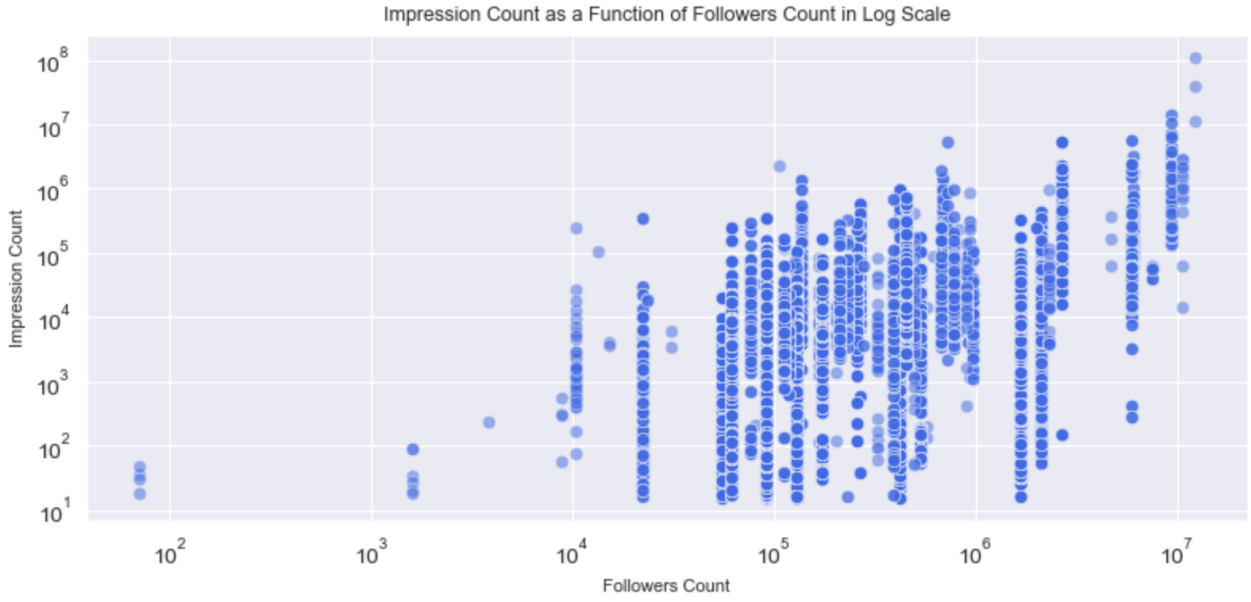

Figure 8: Log-Log Scatter Plot of Impression Count as a function of followers for the French Politicians dataset.

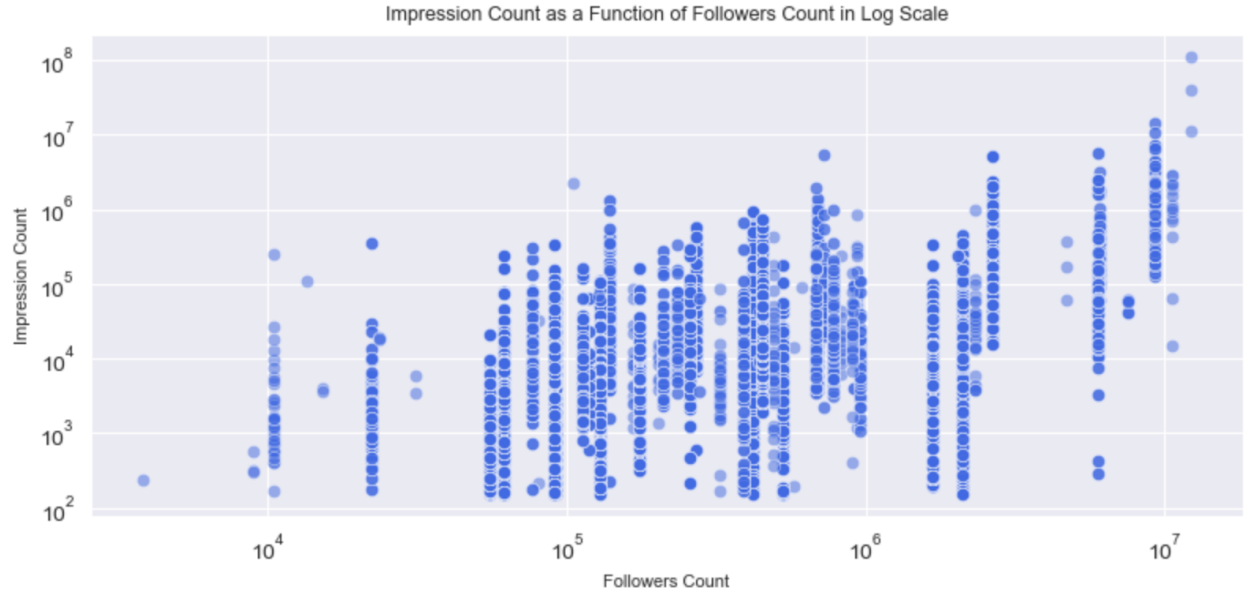

Figure 9: Log-Log Scatter Plot of Impression Count as a function of followers for the French Celebrities dataset.

Engagement and Impression Counts Scatter Plots - American Politicians

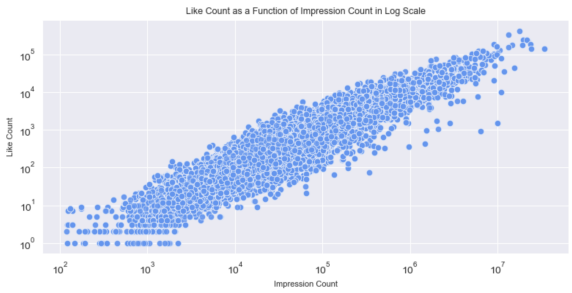

Figure 10: Log-Log Scatter Plot of Likes as a function of Impression Count for the American Politicians dataset.

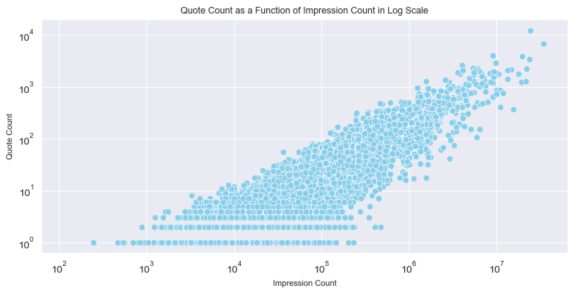

Figure 11: Log-Log Scatter Plot of Quotes as a function of Impression Count for the American Politicians dataset.

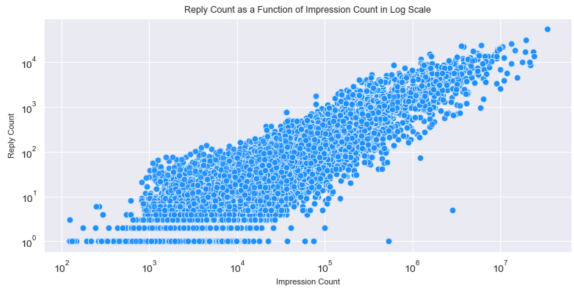

Figure 12: Log-Log Scatter Plot of Replies as a function of Impression Count for the American Politicians dataset.

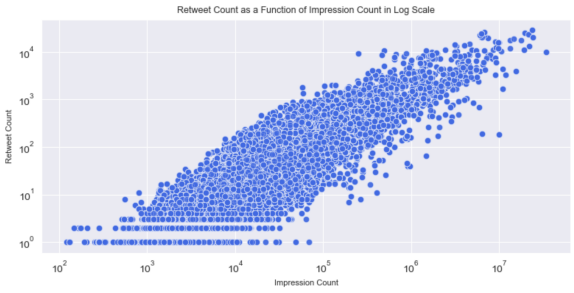

Figure 13: Log-Log Scatter Plot of Retweets as a function of Impression Count for the American Politicians dataset.

Engagement and Impression Counts Scatter Plots - American Celebrities

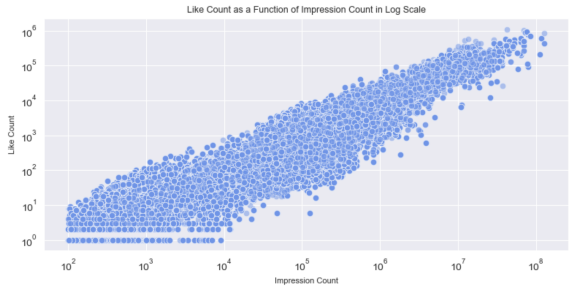

Figure 14: Log-Log Scatter Plot of Likes as a function of Impression Count for the American Celebrities dataset.

Figure 15: Log-Log Scatter Plot of Quotes as a function of Impression Count for the American Celebrities dataset.

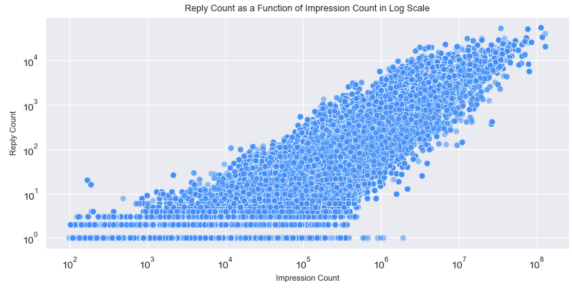

Figure 16: Log-Log Scatter Plot of Replies as a function of Impression Count for the American Celebrities dataset.

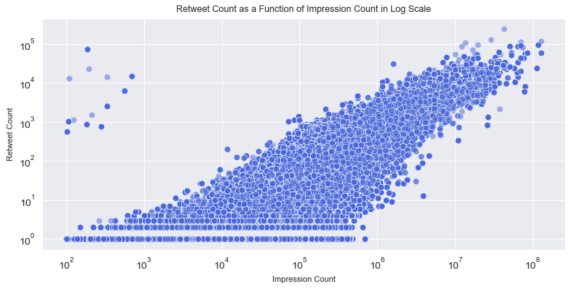

Figure 17: Log-Log Scatter Plot of Retweets as a function of Impression Count for the American Celebrities dataset.

Engagement and Impression Counts Scatter Plots - French Politicians

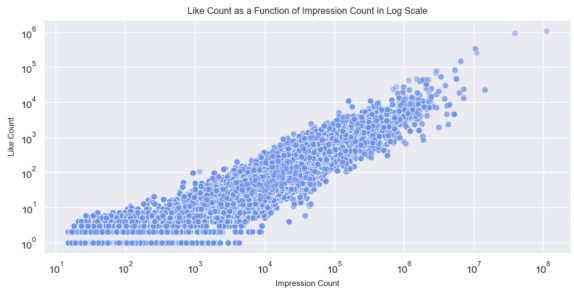

Figure 18: Log-Log Scatter Plot of Likes as a function of Impression Count for the French Politicians dataset.

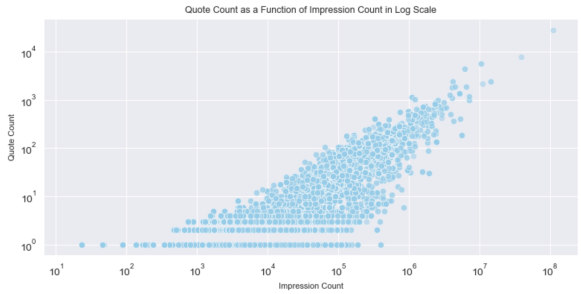

Figure 19: Log-Log Scatter Plot of Quotes as a function of Impression Count for the French Politicians dataset.

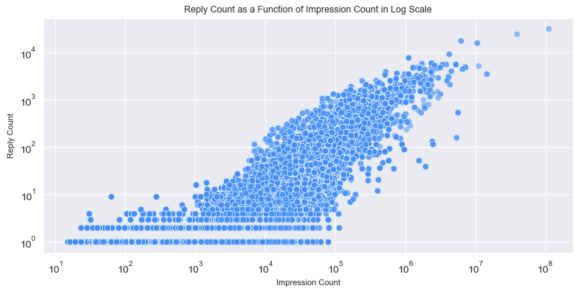

Figure 20: Log-Log Scatter Plot of Replies as a function of Impression Count for the French Politicians dataset.

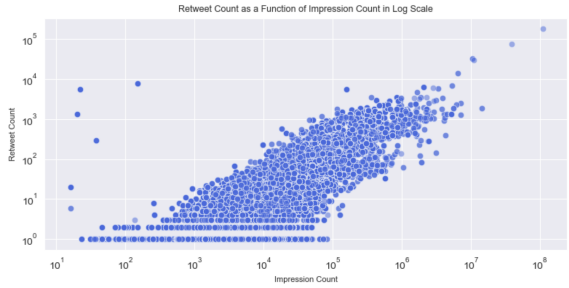

Figure 21: Log-Log Scatter Plot of Retweets as a function of Impression Count for the French Politicians dataset.

Engagement and Impression Counts Scatter Plots - French Celebrities

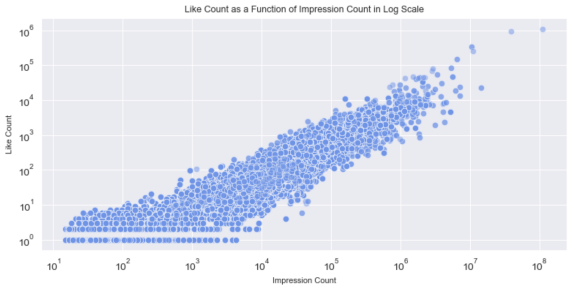

Figure 22: Log-Log Scatter Plot of Likes as a function of Impression Count for the French Celebrities dataset.

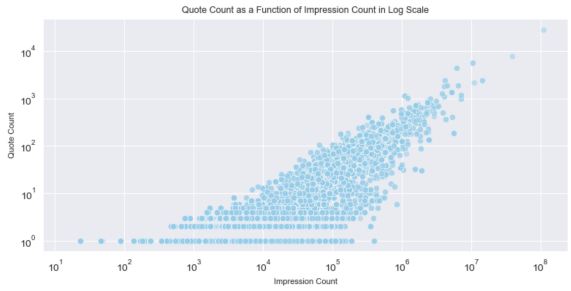

Figure 23: Log-Log Scatter Plot of Quotes as a function of Impression Count for the French Celebrities dataset.

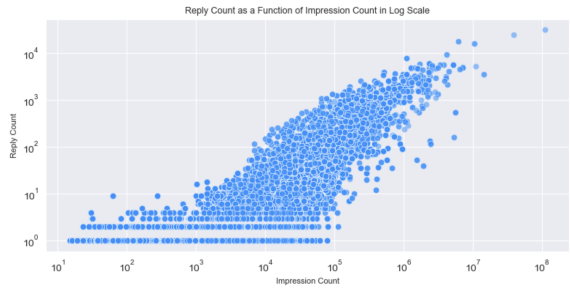

Figure 24: Log-Log Scatter Plot of Replies as a function of Impression Count for the French Celebrities dataset.

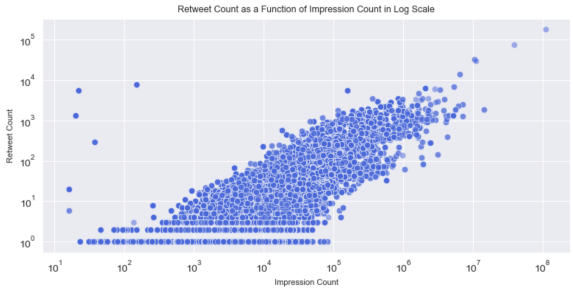

Figure 25: Log-Log Scatter Plot of Retweets as a function of Impression Count for the French Celebrities dataset.

Coefficients plots of logarithmic regressions

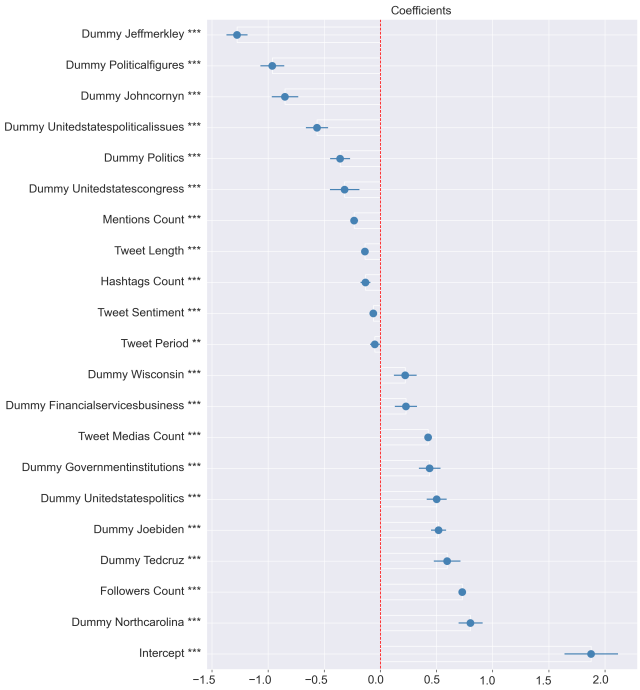

Figure 26: Coefficient plot of the regression using log-variables and cleaned dataset for the American politicians.

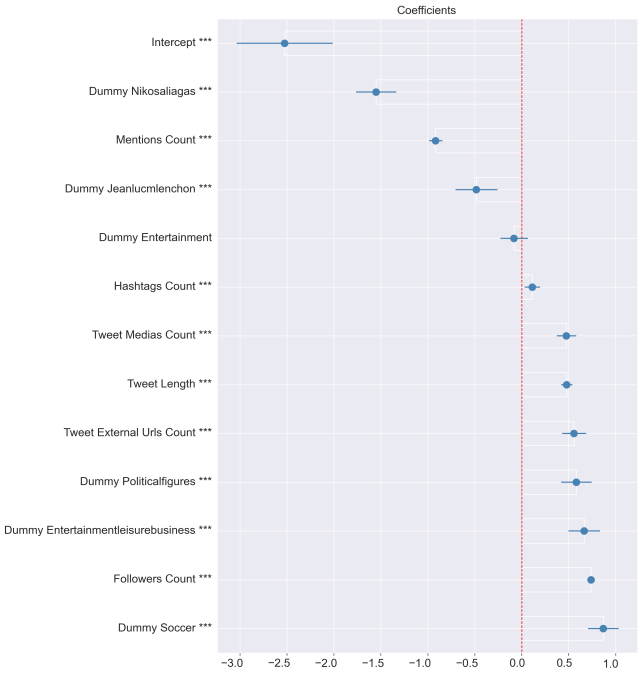

Figure 27: Coefficient plot of the regression using log-variables and cleaned dataset for the French celebrities.

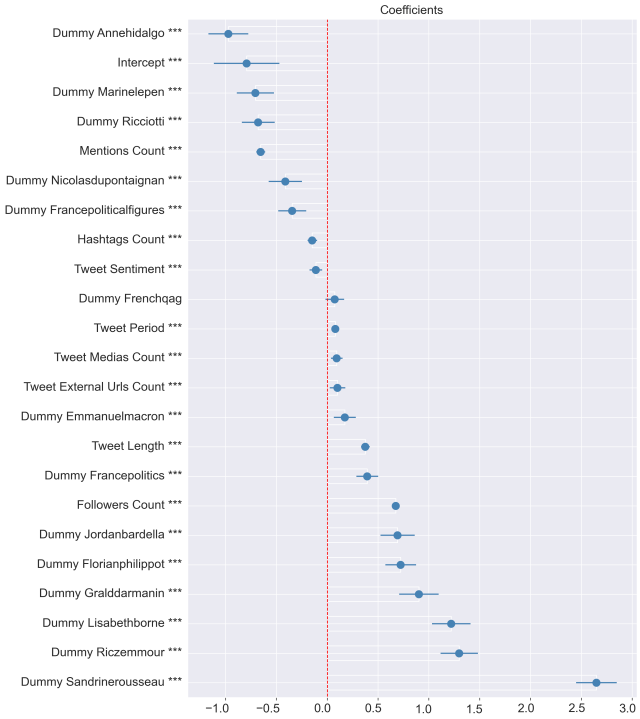

Figure 28: Coefficient plot of the regression using log-variables and cleaned dataset for the French politicians.

References

[1] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, Daniel M Romero, Jahna Otterbacher, Carsten Schwemmer, Kenneth Joseph, David Garcia, and Fred Morstatter. Just another day on twitter: A complete 24 hours of twitter data. *arxiv*, January 2023.

[2] Juergen Pfeffer, Daniel Matter, and Anahit Sargsyan. The half-life of a tweet. *arxiv*, February 2023.