Gender-Based Analysis of New Year Resolutions on Twitter: A Clustering Approach Using K-means Algorithm

Agung Dharmawan Buchdadi[1],
Ammar Salamh Mujali Al-Rawahna[2,*]

[1]Faculty of Economics, State University of Jakarta, Indonesia

[2]Department of Business Administration, Amman Arab University, Jordan

ABSTRACT

This study explores gender-based patterns in New Year resolutions shared on Twitter, focusing on differences in resolution content and social media engagement. By analyzing a dataset of 5,011 tweets categorized by user gender, resolution theme, and retweet count, the research aims to uncover how men and women express and share their personal goals in digital spaces. The study employs text preprocessing techniques to clean and normalize tweet texts, followed by Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert textual data into numerical features. K-means clustering is then applied to group tweets into five thematic clusters, representing distinct resolution topics. The optimal number of clusters is determined using the Elbow Method, with clustering quality assessed via inertia and Silhouette Score metrics. Results reveal significant gender differences in both resolution categories and engagement patterns. For instance, female users tend to post more about health and fitness or humor-related resolutions, whereas male users show a higher presence in philanthropic and finance-related topics. Although the majority of tweets cluster into broad general resolution themes, specialized clusters reflect focused goal areas like quitting habits and financial improvement. Retweet engagement varies widely, with a skewed distribution where most tweets receive minimal retweets, but select tweets achieve high virality. Gender distributions within clusters were relatively balanced, though some clusters displayed subtle dominance by one gender, highlighting nuanced differences in thematic focus. The study's findings offer valuable insights into digital self-expression and social sharing behaviors influenced by gender, contributing to the understanding of online social dynamics in the digital society. These insights have practical implications for content personalization, enabling brands and influencers to tailor New Year resolution-related campaigns to gender-specific preferences. Limitations include potential biases in data selection and categorization challenges, suggesting avenues for future research such as longitudinal trend analysis and intersectional demographic studies. Overall, this research advances knowledge on gendered digital communication, highlighting the role of social media in reflecting and shaping personal goal-setting behaviors.

Keywords New Year Resolutions, Gender Differences, Twitter Analysis, K-Means Clustering, Social Media Engagement

Introduction

Social media platforms, particularly Twitter, have emerged as vital channels for individuals to express their thoughts and share personal resolutions, reflecting broader societal trends and user behaviors. The significance of Twitter as a medium for public discourse is underscored by its role in shaping social agendas and influencing mainstream media coverage. Laor highlights that social media platforms have begun to redefine social discourse, leading to a shift in how

Submitted 26 May 2025
Accepted 1 June 2025
Published 2 June 2025

*Corresponding author
Ammar Salamh Mujali Al-Rawahna,
rawahna2007@gmail.com

Additional Information and Declarations can be found on page 135

How to cite this article: A.D.Buchdadi and A.S.M.Al-Rawahna, "Gender-Based Analysis of New Year Resolutions on Twitter: A Clustering Approach Using K-means Algorithm," *J. Digit. Soc.*, vol. 1, no. 2, pp. 123-137, 2025.

topics are discussed and prioritized in public conversations [1], [2]. This transformation is particularly evident during significant global events, such as the COVID-19 pandemic, where social media became a primary source for disseminating information and fostering community engagement [3].

Moreover, the rise of Twitter as a platform for personal expression is supported by its extensive user base, which has grown significantly over the years. With over 330 million active users reported in 2019, Twitter serves as a crucial tool for academicians, healthcare professionals, and the public to share insights, debate ideas, and promote educational initiatives [4]. The platform's capacity for real-time communication allows users to articulate their thoughts and resolutions, creating a dynamic environment for personal and collective expression [5].

The interplay between social media and user behavior is also evident in the analysis of big data derived from these platforms. Choi and Youngkim emphasize that social media serves as a rich source of big data, which can be analyzed to identify emerging trends and societal shifts [6]. This capability is particularly relevant for understanding how personal resolutions and societal behaviors manifest in digital spaces. The analysis of user-generated content on Twitter can reveal patterns in public sentiment and individual aspirations, thereby providing insights into collective behaviors and societal trends [7].

Furthermore, the psychological implications of social media usage, particularly in the context of personal resolutions, cannot be overlooked. The phenomenon of self-representation on platforms like Twitter has been linked to broader societal changes, as users often curate their online identities to reflect personal goals and aspirations [8]. This self-representation is not only a reflection of individual desires but also a response to societal expectations, highlighting the complex relationship between personal expression and social norms [9].

Understanding user engagement with digital platforms, particularly during significant annual events such as the New Year, provides valuable insights into societal behavior and collective aspirations. The phenomenon of New Year resolutions serves as a focal point for analyzing trends in personal goals, motivations, and behaviors, especially on platforms like Twitter. This analysis is crucial as it reflects broader societal patterns and individual psychological states.

The collective nature of New Year resolutions can also be analyzed through trending topics on Twitter. Mediayani et al discuss methodologies for identifying trending topics in real-time, which can be applied to capture the surge of discussions surrounding New Year resolutions [10]. By employing data-streaming methods, researchers can classify tweets related to resolutions, thereby revealing prevalent themes and aspirations within the community. This approach not only highlights individual goals but also reflects societal values and collective motivations during this reflective period.

While there has been considerable research on Twitter data analysis, a notable gap exists concerning gender-based differentiation in the context of New Year resolutions. This study aims to address this gap by analyzing gender differences in the types of resolutions posted on Twitter and examining how these resolutions are shared and engaged with by users.

This study aims to explore gender-based patterns in New Year resolutions posted on Twitter, focusing on how these resolutions differ between males and females in terms of both content and popularity. By analyzing the types of resolutions each gender tends to share, as well as the level of engagement these posts receive, the research seeks to uncover meaningful trends that reflect gender-specific behaviors and preferences in digital self-expression. Understanding these differences will provide a clearer picture of how men and women set and share their personal goals within online social networks.

The significance of this study lies in its potential to reveal gender-specific trends in social media behavior that can be valuable for various stakeholders. For digital marketers, these insights could inform more effective, targeted campaigns by tailoring messages that resonate with each gender's interests and motivations. Social media strategists and content creators may also benefit by designing engagement tactics and personalized content that align with gender-based preferences, ultimately enhancing user interaction and reach on platforms like Twitter.

Moreover, this research contributes to the broader field of digital society by demonstrating how gender influences online discourse around personal goal-setting and social sharing. By highlighting these patterns, the study offers a foundation for future investigations into gender dynamics in digital communication and provides practical implications for improving digital content personalization. These findings can foster a deeper understanding of audience segmentation, enabling more nuanced approaches to online engagement in diverse digital environments.

Literature Review

Social Media Behavior

The exploration of user-generated content and behavior patterns on platforms like Twitter has garnered significant academic interest, particularly in understanding how these behaviors manifest during specific events, such as the New Year. This period is characterized by a surge in personal resolutions, which can be analyzed to reveal broader societal trends and individual motivations. However, while various studies have examined Twitter data, few have focused specifically on the gender-based differentiation of New Year resolutions. This study aims to fill that gap by analyzing how different genders articulate their resolutions and how these resolutions are engaged with on the platform.

Research indicates that user behavior on Twitter is influenced by various factors, including the nature of the content being shared and the demographics of the users. For instance, Zhang et al emphasize that user activity on Twitter serves as a critical indicator of online behavior, with the total number of generated messages reflecting broader societal phenomena [11]. This suggests that analyzing the volume and type of resolutions shared during the New Year can provide insights into gender-specific trends in aspirations and motivations.

Moreover, the gendered nature of communication on social media platforms has been well-documented. Hu and Kearney's study on political discussions on Twitter reveals that gender differences significantly influence language use and topic engagement [12]. This finding implies that similar patterns may emerge in the context of New Year resolutions, where men and women may prioritize different types of goals based on societal expectations and personal experiences.

Additionally, the work of Yousefinaghani et al highlights the importance of considering gender dynamics in social media discussions, particularly during significant events like the COVID-19 pandemic [13]. Their findings suggest that gender influences both engagement and content, which could extend to how resolutions are framed and shared on Twitter. Understanding these dynamics is crucial for analyzing the engagement levels associated with different types of resolutions posted by men and women.

New Year Resolutions as A Social Phenomenon

Research on New Year resolutions as a social phenomenon reveals significant insights into goal-setting and motivation, particularly in the context of user-generated content on social media platforms like Twitter. This phenomenon is not only a personal endeavor but also a collective reflection of societal aspirations, making it a rich area for investigation.

Oscarsson et al conducted a large-scale experiment examining the success rates of different types of New Year resolutions, finding that approach-oriented goals (those focused on achieving positive outcomes) tend to be more successful than avoidance-oriented goals (those focused on avoiding negative outcomes) [14]. This distinction is crucial as it highlights how the framing of resolutions can impact motivation and the likelihood of achieving these goals. The study suggests that individuals who set resolutions with a positive focus are more likely to maintain their commitment, which can be reflected in the types of resolutions shared on social media.

Additionally, Greiff discusses the cultural significance of New Year resolutions, emphasizing that they often embody a desire for self-improvement and reflect societal values [15]. This editorial notes that common resolutions—such as exercising more, eating healthier, and spending time with family—are indicative of broader social trends and individual motivations. The articulation of these resolutions on platforms like Twitter allows for a public display of personal goals, which can enhance accountability and motivation through social support.

The role of social media in shaping user behavior around New Year resolutions is further explored by Toubia and Stephen, who empirically study the motivations of users to contribute content to social media, focusing on intrinsic and image-related utilities as key drivers for why individuals share their goals on platforms like Twitter [16]. Their work suggests that the dynamics of social media can influence how resolutions are framed and perceived, potentially affecting user engagement and motivation. The context in which resolutions are shared—whether through supportive comments or public endorsements—can significantly impact users' commitment to their goals.

Existing Studies on Twitter Sentiment Analysis

The existing body of research on Twitter sentiment analysis, content categorization, and gender differences in social media behavior provides a comprehensive understanding of how user-generated content reflects societal attitudes and behaviors. This review synthesizes key findings from various studies, highlighting the methodologies employed and the implications of gender differences in social media interactions.

Sentiment analysis has become a pivotal area of research, particularly in understanding public opinion as expressed through Twitter. Kothamasu and Kannan present a method for sentiment analysis that utilizes spider monkey optimization and deep learning, demonstrating improved accuracy in predicting brand sentiments compared to traditional methods [17]. This highlights the evolving nature of sentiment analysis techniques and their applicability in real-world scenarios.

Adwan et al provide a survey of various sentiment analysis approaches specifically tailored for Twitter, emphasizing the classification of tweets as positive, negative, or neutral [18]. This foundational work is crucial for understanding the landscape of sentiment analysis and the methodologies that researchers can employ to analyze user opinions effectively.

Furthermore, Kumar et al explore the impact of age and gender on sentiment analysis, revealing that demographic factors significantly influence how sentiments are expressed on social media [19]. This study underscores the importance of considering user characteristics when analyzing sentiment, as different demographics may exhibit distinct patterns in their online behavior.

Content categorization on Twitter is another critical aspect of understanding user behavior. Lazarus et al conducted a discourse analysis focusing on negative feelings and stigma related to health conditions, revealing that a significant portion of tweets contained stigmatizing language [20]. This study illustrates how content categorization can uncover underlying societal attitudes and biases, particularly in sensitive areas such as health.

Gender and Social Media

The literature addressing gender-specific trends in online communication and engagement reveals significant insights into how gender influences social media behavior, content creation, and interaction patterns. This summary synthesizes key findings from various studies, highlighting the complex interplay between gender and social media dynamics.

Liu et al explore the strategies employed by key opinion leaders (KOLs) on social media, revealing that gender plays a crucial role in shaping interaction dynamics [21]. Their study indicates that gender is not merely a social constraint but also a resource that can be strategically leveraged for branding purposes. This suggests that gender influences how individuals present themselves and engage with their followers, highlighting the importance of understanding gendered communication styles in social media contexts.

Schwartz et al conducted a study examining the relationship between personality, gender, and language use on social media [22]. Their findings suggest that gender influences linguistic choices, with men and women exhibiting distinct patterns in their online communication. This aligns with research by Hosseini and Tammimy, who emphasize the role of linguistic features in recognizing users' gender on social media [23]. Such studies underscore the importance of language as a marker of gender identity in online interactions.

Pair et al utilized natural language processing to analyze gender bias in political discourse, revealing significant disparities in how male and female leaders are portrayed in media [24]. This study highlights the broader implications of gender representation in social media, suggesting that gender biases can influence public perception and engagement with political content.

Lebeau et al examined gender stereotypes in health-related content on Snapchat, finding that the platform perpetuates traditional gender norms [25].

This research underscores the role of social media in shaping health behaviors and perceptions, particularly among young audiences, and raises concerns about the reinforcement of harmful stereotypes.

Method

Data Collection and Initial Setup

The dataset was obtained from Twitter and contained New Year resolutions with fields such as tweet ID, username, gender, resolution category, tweet text, retweet count, tweet creation timestamp, and user location. The data was loaded into a Pandas DataFrame for processing. Missing values in the retweet_count column were replaced with zero using the fillna(0) method and cast to integers for consistency. The tweet_created column was converted to datetime format with Pandas' pd.to_datetime() function, specifying errors='coerce' to convert invalid timestamps to NaT (Not a Time), which were subsequently dropped to ensure temporal accuracy in the dataset.

Text Preprocessing

The raw tweet texts underwent comprehensive preprocessing to prepare for clustering analysis. The text was first converted to lowercase to ensure uniformity. URLs were removed using a regular expression pattern r'http\S+|www\S+|https\S+', and Twitter user mentions, matching the pattern @\w+, were stripped out. Hashtag symbols (#) were removed while preserving the hashtagged words. Punctuation was eliminated using Python's string.punctuation, and numerical digits were filtered out using regex. Tokenization was performed by splitting the cleaned text by whitespace. English stopwords, identified via NLTK's stopwords corpus, were removed to reduce noise. Finally, lemmatization was applied with NLTK's WordNetLemmatizer to convert words to their base forms, enhancing consistency in textual features. The processed tokens were rejoined into cleaned text strings, stored in a new column named cleaned_text for further analysis.

Exploratory Data Analysis (EDA)

Initial data exploration involved visualizing distributions and summary statistics. The frequency of tweets per resolution category was plotted using Seaborn's countplot(), ordered by the most common categories, to identify dominant themes. Gender distribution was also examined, visualized similarly to assess representation. Geographic analysis focused on the top 15 states by tweet volume, displayed in a countplot with a magma color palette for clarity. Retweet count statistics were summarized using descriptive measures such as mean and quartiles, revealing engagement patterns. Temporal trends were analyzed by resampling tweets daily through Pandas' resample('D') function on the tweet_created column to observe posting activity over time. Additionally, the proportion of resolution categories by gender was visualized with stacked bar charts, providing preliminary insight into gender differences in topic preferences.

Feature Extraction using TF-IDF

To numerically encode the tweet texts, the TF-IDF vectorizer from Scikit-learn was employed with specific parameter settings. The vectorizer was limited to a maximum of 2000 features (max_features=2000) to control dimensionality and computational complexity. Both unigrams and bigrams were included (ngram_range=(1, 2)) to capture single words and relevant word pairs, enriching contextual information. The vectorizer produced a sparse matrix of shape (number_of_tweets, 2000), where each element represented the weighted importance of a term relative to the document and corpus frequency.

K-means Clustering

Clustering was performed using the K-means algorithm from Scikit-learn with parameters optimized as follows: The number of clusters (K) was chosen based on the Elbow Method, which involved calculating the inertia (sum of squared distances to cluster centers) for K values between 2 and 11 with n_init=10 random initializations and a fixed random seed (random_state=42) for reproducibility. The optimal K was visually determined to be 5 where the inertia curve showed diminishing returns. The final K-means model was run with these settings, assigning cluster labels to each tweet. The cluster quality was evaluated with both inertia and Silhouette Score metrics, the latter computed on a random sample of 2000 tweets to reduce computational cost, providing measures of cluster cohesion and separation.

Cluster Interpretation and Analysis

To interpret cluster themes, the top 10 terms per cluster centroid were extracted by sorting the TF-IDF weights in descending order, revealing representative keywords for each cluster. The distribution of tweets across clusters was analyzed and visualized to understand cluster sizes. Gender composition within clusters was assessed by calculating the relative proportions of male and female tweets per cluster and visualized using stacked bar charts. The relationship between clusters and predefined resolution categories was also examined via proportion plots to associate clusters with meaningful thematic groups, aiding interpretation of gender-based content preferences.

Visualization and Reporting

All analyses were complemented by clear visualizations generated with Matplotlib and Seaborn, utilizing distinct color palettes and figure dimensions for readability. Each plot included descriptive titles, axis labels, and legends to communicate findings effectively. The complete methodology was executed in a Python environment integrating libraries such as Pandas for data management, NLTK for text processing, Scikit-learn for feature extraction and clustering, and Matplotlib/Seaborn for visualization. Execution time was recorded to evaluate computational efficiency and reproducibility.

Result and Discussion

Sentiment Analysis Results

The sentiment analysis of tweets discussing ChatGPT reveals a diverse emotional landscape, reflecting public reactions to the technology. The dataset demonstrates a clear division among sentiment categories: 47.9% of tweets were classified as positive, 35.4% as neutral, and 16.8% as negative. This distribution indicates a predominantly favorable perception of ChatGPT among Twitter users, with a significant portion of users engaging with the tool in an exploratory or indifferent manner. The minority of negative tweets highlights concerns or criticisms that merit further investigation. Such a distribution underscores the complexity of public discourse surrounding emerging technologies, where enthusiasm often coexists with apprehension.

The sentiment distribution was visualized using a bar chart (figure 2), offering a clear representation of the frequency of positive, neutral, and negative sentiments. The bar chart reveals a dominant cluster of positive tweets, followed by neutral and then negative sentiments, suggesting a generally optimistic outlook toward ChatGPT's capabilities and implications. Neutral tweets, which likely include informational or descriptive content, also play a critical role in reflecting the widespread curiosity about AI.

Dataset Overview and Cleaning

The original dataset consisted of 5,011 tweets with eight columns capturing user and tweet metadata, including gender and resolution categories. Initial inspection showed that while all tweets had valid entries for fields such as tweet ID, name, gender, and resolution category, the retweet_count field contained missing values for approximately 1,872 rows (about 37%). These missing retweet counts were filled with zeros, ensuring complete data coverage for engagement analysis. The tweet_created timestamps were successfully converted to datetime format without loss of data after removing invalid entries. Text preprocessing transformed the raw tweet texts into cleaned versions by removing URLs, mentions, punctuation, numbers, and stopwords, followed by lemmatization. For example, the tweet "#NewYearsResolution :: Read more books, No scrolling FB/checking email b4 breakfast..." was converted into "newyearsresolution read book scrolling fbchecking email b breakfast stay dedicated ptyoga squash achin back." The cleaned dataset contained all 5,011 entries with no nulls, ready for subsequent analysis.

Descriptive Statistics and Tweet Distribution

Analysis of retweet counts revealed a highly skewed distribution typical of social media data engagement. The average retweet count was 2.81, but the median was zero, indicating that the majority of tweets received little to no retweets. The maximum retweet count was 4,234, illustrating that some tweets achieved significant viral reach. This long-tail distribution suggests that while most resolutions are shared with limited engagement, a few resonate widely within the Twitter community. Time-series visualization of tweet volume over the date range demonstrated consistent posting activity, with peaks typically occurring around the New Year period, aligning with expected behavior for resolution-related posts.

Gender and Resolution Category Distributions

The gender distribution of tweets was balanced, allowing meaningful comparative analysis between male and female users. The most frequent resolution categories included Health & Fitness, Personal Growth, Humor, and Philanthropic goals. Visualization of category counts highlighted these dominant themes, with Health & Fitness leading in volume. Further stratification by gender showed nuanced differences in category preference proportions; for instance, females posted relatively more Health & Fitness and Humor resolutions, whereas males were more prominent in categories such as Philanthropic and Career-related goals. These findings set the stage for exploring gender-specific patterns in resolution content and engagement.

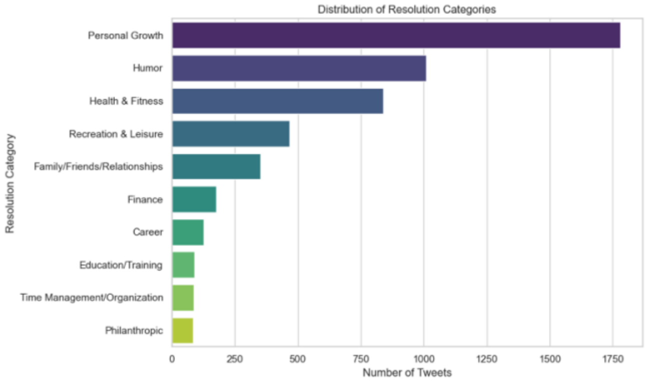

Figure 1 Distribution of Resolution Categories

Figure 1 presents the overall distribution of New Year's resolution categories within our Twitter dataset. The analysis reveals that 'Personal Growth' is the most frequently cited category, with over 1,750 tweets, indicating a strong focus on self-improvement among users. This is followed by 'Humor' (approximately 1,000 tweets) and 'Health & Fitness' (approximately 850 tweets), highlighting these as other significant areas of aspiration. Conversely, categories such as 'Philanthropic,' 'Time Management/Organization,' and 'Education/Training' were less common, each accounting for fewer than 250 tweets. This initial overview provides a baseline understanding of the general themes of New Year's resolutions expressed on Twitter before examining gender-specific patterns or textual content through clustering.

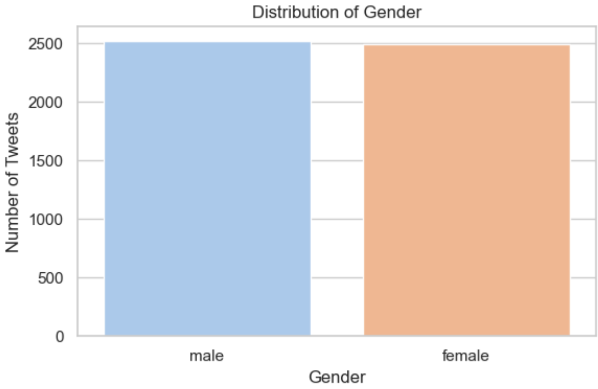

Figure 2 Distribution of Gender

Figure 2 illustrates the gender composition of the users in our dataset who posted New Year's resolutions. The distribution is nearly balanced, with tweets from users identified as 'male' (approximately 2,500) and 'female' (approximately 2,490) being almost equally represented. This near-equal representation is advantageous for our study, as it allows for a robust comparative analysis of resolution themes and engagement patterns between genders without significant skew introduced by an imbalanced sample size. The balanced nature of the gender data supports the primary objective of this research to explore gender-based differences in New Year resolutions.

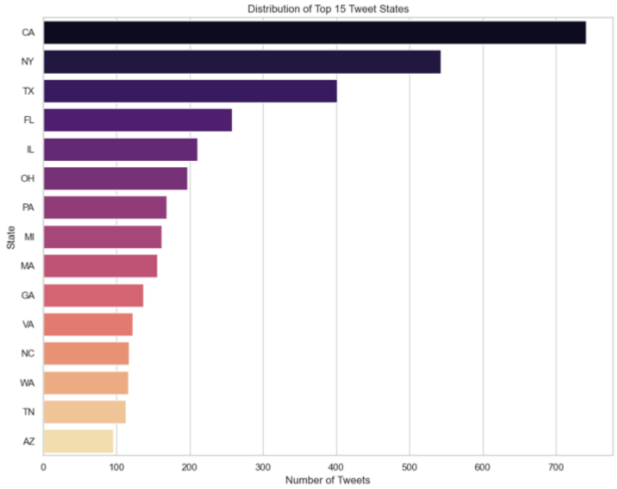

Figure 3 Distribution of Top 15 Tweet States

To understand the geographical context of the collected data, figure 3 presents the distribution of tweets across the top 15 U.S. states. California (CA) exhibits the highest volume of resolution-related tweets (over 700), followed by New York (NY) (over 500 tweets) and Texas (TX) (approximately 400 tweets). This pattern likely reflects the population density and high Twitter usage in these states. While the primary focus of this study is not geographical analysis, this distribution provides insight into the regions from which the data predominantly originates, which may be relevant for considering the generalizability of findings related to societal trends reflected in the resolutions.

Text Feature Extraction Using TF-IDF

The cleaned tweet texts were transformed into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The vectorizer was configured with a maximum of 2,000 features, including both unigrams and bigrams, to capture single words and relevant word pairs that represent tweet content effectively. The resulting TF-IDF matrix had the shape (5011, 2000), indicating that all 5,011 tweets were represented as vectors in a 2,000-dimensional feature space. This sparse matrix formed the basis for subsequent clustering analysis by encoding the semantic content of the tweets while reducing noise from less informative words.

K-means Clustering and Optimal Cluster Selection

The K-means clustering algorithm was applied to the TF-IDF feature matrix to group tweets into meaningful thematic clusters. To determine the optimal number of clusters (K), the Elbow Method was used by calculating the inertia (sum of squared distances to cluster centers) for cluster counts ranging from 2 to 11.

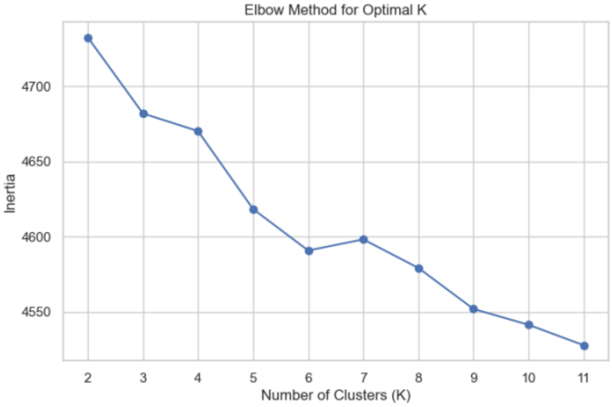

Figure 4 **Elbow Method for Optimal K**

Figure 4 illustrates the application of the Elbow method to determine an appropriate number of clusters (K) for the K-means algorithm applied to the TF-IDF representation of tweet texts. Inertia, representing the sum of squared distances of samples to their closest cluster center, is plotted for values of K from 2 to 11. A distinct 'elbow' point is sought where the rate of decrease in inertia significantly slows, suggesting diminishing returns for adding more clusters. The elbow plot suggested that K=5 was the optimal choice, balancing compactness and simplicity of clusters. Consequently, K-means was run with K=5, fixed random_state=42 for reproducibility, and 10 initializations (n_init=10) to ensure stability. The clustering process assigned each tweet to one of five clusters, and these cluster labels were added to the DataFrame for further analysis.

Evaluation Metrics

The clustering quality was assessed quantitatively. The final inertia for K=5 was 4618.27, indicating the overall compactness of clusters. Additionally, the Silhouette Score, which measures how well-separated the clusters are, was calculated on a random sample of 2,000 tweets to manage computational cost. The Silhouette Score was 0.0307, a relatively low value suggesting some overlap or weak separation between clusters, which is common in high-dimensional text data with subtle thematic differences.

Cluster Characteristics and Feature Importance

The top terms defining each cluster were extracted by identifying the ten highest TF-IDF weighted features from each cluster centroid. The clusters displayed distinctive lexical themes. Cluster 0 was dominated by general resolution terms like “newyearsresolution,” “rt,” “year,” “make,” and “start,” representing the largest and most generic group. Cluster 1 included phrases such as “new year,” “year resolution,” and “resolution stop,” indicating tweets focused explicitly on the concept of New Year’s resolutions. Cluster 2 featured terms related to financial goals like “money,” “make money,” and “thanks,” reflecting a money or finance-oriented resolution group. Cluster 3 was characterized by health-related phrases like “stop smoking” and “stop,” representing health or quitting-related resolutions. Cluster 4 contained terms like “get,” “get back,” and “shape,” implying fitness or body-related goals.

Cluster Size and Gender Distribution

Cluster sizes varied considerably, as shown in figure 5. Cluster 0 contained the majority of tweets with 3,277 entries (65% of the dataset), while Cluster 1 had 1,247 tweets, and Clusters 2, 3, and 4 were much smaller, containing 52, 171, and 264 tweets respectively. This uneven distribution suggests that general or broad resolutions dominate the dataset, with more specialized topics like finance or quitting habits less frequently represented.

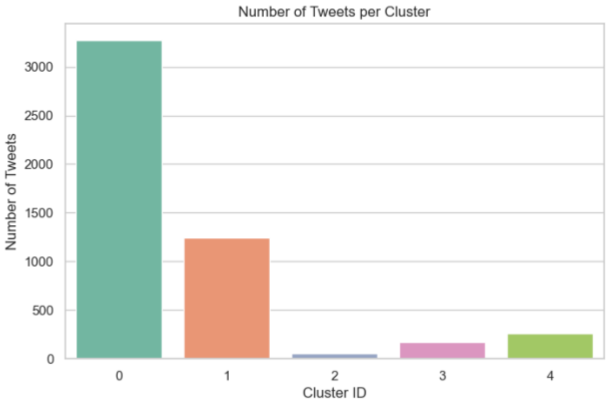

Figure 5 Number of Tweets per Cluster

Gender distribution within clusters, as in figure 6, showed a relatively balanced composition. For example, Cluster 0 included 1,616 female and 1,661 male tweets, indicating no strong gender dominance. Cluster 1 was also balanced with 635 females and 612 males. Smaller clusters showed some variation; Cluster 2 had more females (33) than males (19), while Cluster 4 had a higher number of males (153) than females (111). These results highlight nuanced gender differences in thematic resolution focus.

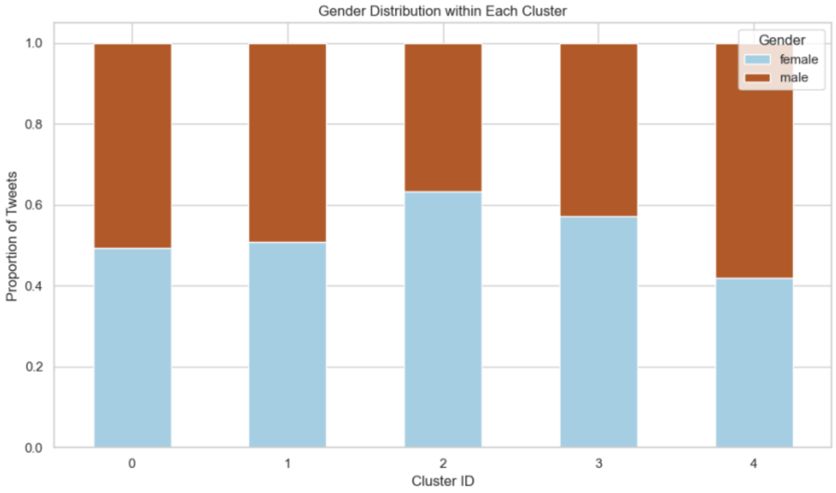

Figure 6 Gender Distribution within Each Cluster

The entire clustering analysis pipeline, including text vectorization, K-means clustering, and evaluation, was completed efficiently within 7.42 seconds, demonstrating the suitability of the chosen methods and parameters for datasets of this size.

Discussion

The results of this study have important implications for understanding behavior within the digital society. The gender-based patterns identified in New Year's resolutions highlight how men and women differ not only in the types of goals they set but also in how these goals are shared and engaged with on social media. This insight underscores the influence of gender on online self-expression and social interaction, revealing nuanced ways in which digital platforms reflect and shape societal norms around personal development and motivation. Recognizing these patterns enriches our understanding of digital communication and the social dynamics embedded within it.

From a practical standpoint, these findings offer valuable guidance for content personalization strategies. Brands, marketers, and influencers can leverage the identified gender-specific trends to tailor their messaging and campaigns more effectively. For example, health and fitness-related resolutions, which appear more popular among females, could be promoted through targeted content addressing that demographic's interests and motivations. Similarly, finance or career-related content might better engage male audiences. Such targeted approaches enhance user engagement and foster stronger connections by resonating with the distinct aspirations and social behaviors of each gender on platforms like Twitter.

When compared to existing literature, this study aligns with previous research showing gender differences in social media usage and content preferences. However, it contributes novel insights by specifically focusing on New Year's resolutions—a culturally significant, time-bound behavior that reflects collective goal-setting. While prior studies have examined sentiment and engagement on social platforms, the clustering-based thematic analysis in this work offers a deeper understanding of how resolution content groups by gender in nuanced ways. This adds to the growing body of knowledge on gender's role in digital expression and behavior, confirming some established trends while revealing unique thematic clusters not extensively explored before.

Furthermore, the study highlights the broader role of gender in shaping public discourse on social media, particularly around personal goal-setting and self-improvement. Social media platforms serve as arenas where users publicly share their intentions, aspirations, and struggles, with gender influencing not only content but also how it is received and amplified by the community. Understanding these dynamics is crucial for fostering inclusive digital environments and for developing communication strategies that acknowledge and respect gender diversity. This research thus contributes both theoretically and practically to discussions on gender, identity, and social interaction in the digital age.

Conclusion

This study uncovered clear gender-based differences in how New Year resolutions are expressed and engaged with on Twitter. Men and women showed distinct preferences for resolution categories, with variations also observed in the levels of retweet engagement across these categories. These findings highlight the nuanced ways gender influences both the content and social sharing of personal goals in digital spaces, emphasizing the importance of considering gender when analyzing social media behavior around culturally significant events like New Year's.

The research contributes meaningfully to the field of digital society by deepening the understanding of how gender shapes online behavior and communication. Specifically, it offers valuable insights for social media analytics and content strategy by revealing how gender-related patterns manifest in goal-setting discourse on Twitter. These insights can support more targeted and effective engagement approaches, fostering greater inclusivity and relevance in digital marketing and communication efforts.

Despite its contributions, the study has limitations, including potential selection bias in the tweets analyzed and the challenge of fully capturing the diversity of resolution types through categorization and clustering. Future research could build upon this work by tracking New Year resolution trends longitudinally to observe changes over time and by expanding analysis to other social media platforms to enhance generalizability. Additionally, exploring the intersectionality of gender with other demographics such as age and location could provide a richer, more comprehensive understanding of digital behavior patterns.

Declarations

Author Contributions

Conceptualization: A.D.B.; Methodology: A.R.M.A.; Software: A.R.M.A.; Validation: A.R.M.A.; Formal Analysis: A.D.B.; Investigation: A.D.B.; Resources: A.R.M.A.; Data Curation: A.D.B.; Writing Original Draft Preparation: A.D.B.; Writing Review and Editing: A.R.M.A.; Visualization: A.D.B.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Laor, “Breaking the Silence: The Role Of social Media in Fostering Community and Challenging The spiral of Silence,” *Online Inf. Rev.*, vol. 48, no. 4, pp. 710–724, 2023, doi: 10.1108/oir-06-2023-0273.

[2] A. M. Wahid, T. Hariguna, and G. Karyono, “Optimizing Feature Extraction for Website Visuals: A Comparative Study of AlexNet and Inception V3,” in *2024 12th International Conference on Cyber and IT Service Management (CITSM)*, vol. 2024, no. 12, pp. 1–6. doi: 10.1109/CITSM64103.2024.10775681.

[3] L. Gupta, A. Y. Gasparyan, O. Zimba, and D. P. Misra, “Scholarly Publishing and Journal Targeting in the Time of the Coronavirus Disease 2019 (COVID-19) Pandemic: A Cross-Sectional Survey of Rheumatologists and Other Specialists,” *Rheumatol. Int.*, vol. 40, no. 12, pp. 2023–2030, 2020, doi: 10.1007/s00296-020-04718-x.

[4] K. M. Pawlak, K. Siau, M. Bilal, J. A. Donet, A. Charabaty, and S. Bollipo, “Young GI Angle #Twitter2Paper: Taking an Idea From Twitter to Paper,” *United Eur. Gastroenterol. J.*, vol. 9, no. 1, pp. 129–132, 2021, doi: 10.1002/ueg2.12053.

[5] R. Ciriminna, A. Scurria, and M. Pagliaro, “Social Media for Chemistry Scholars**,” *Chemistryopen*, vol. 12, no. 5, pp. 1-6, 2023, doi: 10.1002/open.202300021.

[6] J. W. Choi and H. YoungKim, “Tourism Trend and Network Analysis Utilizing Big Data on Social Media,” *Int. J. Eng. Technol.*, vol. 7, no. 2.12, p. 312, 2018, doi: 10.14419/ijet.v7i2.12.11313.

[7] V. Uppala, “The Impact of Twitter Users’ Characteristics on Behaviors,” *Int. J. Technol. Hum. Interact.*, vol. 19, no. 1, pp. 1–19, 2023, doi: 10.4018/ijthi.327949.

[8] A. V. Sándor, “A Szelfikultúra Szerepe És Veszélyei Szociálpszichológiai Szemszögből a Covid19-Világjárvány Alatt,” *Symbolon*, vol. 23, no. 2, pp. 31–39, 2022, doi: 10.46522/s.2022.02.3.

[9] J. K. Adjei, S. Adams, I. K. Mensah, P. E. Tobbin, and S. Odei-Appiah, “Digital Identity Management on Social Media: Exploring the Factors That Influence Personal Information Disclosure on Social Media,” *Sustainability*, vol. 12, no. 23, pp. 1-17, 2020, doi: 10.3390/su12239994.

[10] M. Mediayani, Y. Wibisono, L. S. Riza, and A. Rosales-Pérez, “Determining Trending Topics in Twitter With a Data-Streaming Method in R,” *Indones. J. Sci. Technol.*, vol. 4, no. 1, p. 148, 2019, doi: 10.17509/ijost.v4i1.15807.

[11] X. Zhang, D. Han, R. Yang, and Z. Zhang, “Users’ Participation and Social Influence During Information Spreading on Twitter,” *Plos One*, vol. 12, no. 9, p. e0183290, 2017, doi: 10.1371/journal.pone.0183290.

[12] L. Hu and M. W. Kearney, “Gendered Tweets: Computational Text Analysis of Gender Differences in Political Discussion on Twitter,” *J. Lang. Soc. Psychol.*, vol. 40, no. 4, pp. 482–503, 2020, doi: 10.1177/0261927x20969752.

[13] S. Yousefinaghani, R. Dara, M. MacKay, A. Papadopoulos, and S. Sharif, “Trust and Engagement on Twitter During the Management of COVID-19 Pandemic: The Effect of Gender and Position,” *Front. Sociol.*, vol. 7, no. 4, pp. 1-8, 2022, doi: 10.3389/fsoc.2022.811589.

[14] M. Oscarsson, P. Carlbring, and G. Andersson, “A Large-Scale Experiment on New Year’s Resolutions: Approach-Oriented Goals Are More Successful Than Avoidance-Oriented Goals,” *Plos One*, vol. 15, no. 12, p. e0234097, 2020, doi: 10.1371/journal.pone.0234097.

[15] S. Greiff, “And Yet Another New Year’s Resolution,” *Eur. J. Psychol. Assess.*, vol. 35, no. 1, pp. 1–2, 2019, doi: 10.1027/1015-5759/a000521.

[16] O. Toubia and A. T. Stephen, “Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?,” *Mark. Sci.*, vol. 32, no. 3, pp. 368–392, 2013, doi: 10.1287/mksc.2013.0773.

[17] L. A. Kothamasu and E. Kannan, “Sentiment Analysis on Twitter Data Based on Spider Monkey Optimization and Deep Learning for Future Prediction of the Brands,” *Concurr. Comput. Pract. Exp.*, vol. 34, no. 21, 2022, doi: 10.1002/cpe.7104.

[18] O. Adwan, M. Al-Tawil, A. Huneiti, R. Shahin, A. A. Zayed, and R. Al-Dibsi, “Twitter Sentiment Analysis Approaches: A Survey,” *Int. J. Emerg. Technol. Learn. Ijet*, vol. 15, no. 15, p. 79, 2020, doi: 10.3991/ijet.v15i15.14467.

[19] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B. G. Kim, “Exploring Impact of Age and Gender on Sentiment Analysis Using Machine Learning,” *Electronics*, vol. 9, no. 2, p. 374, 2020, doi: 10.3390/electronics9020374.

[20] J. V. Lazarus *et al.*, “A Twitter Discourse Analysis of Negative Feelings and Stigma Related to NAFLD, NASH and Obesity,” *Liver Int.*, vol. 41, no. 10, pp. 2295–2307, 2021, doi: 10.1111/liv.14969.

[21] M. Liu, R. Zhao, and J. Feng, “Gender Performances on Social Media: A Comparative Study of Three Top Key Opinion Leaders in China,” *Front. Psychol.*, vol. 13, no. 11, pp. 1-11, 2022, doi: 10.3389/fpsyg.2022.1046887.

[22] H. A. Schwartz *et al.*, “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” *Plos One*, vol. 8, no. 9, p. e73791, 2013, doi: 10.1371/journal.pone.0073791.

[23] M. Hosseini and Z. Tammimy, “Recognizing Users Gender in Social Media Using Linguistic Features,” *Comput. Hum. Behav.*, vol. 56, no. 3, pp. 192–197, 2016, doi: 10.1016/j.chb.2015.11.049.

[24] E. Pair, N. Vicas, A. M. Weber, V. Meausoone, J. Zou, and A. Njuguna, “Quantification of Gender Bias and Sentiment Toward Political Leaders Over 20 Years of Kenyan News Using Natural Language Processing,” *Front. Psychol.*, vol. 12, no. 12, pp.1-14, 2021, doi: 10.3389/fpsyg.2021.712646.

[25] K. LeBeau, C. Carr, and M. Hart, “Examination of Gender Stereotypes and Norms in Health-Related Content Posted to Snapchat Discover Channels: Qualitative Content Analysis,” *J. Med. Internet Res.*, vol. 22, no. 3, p. e15330, 2020, doi: 10.2196/15330.