American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 4 ISSUE 1 (2025)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 4 Issue 1, Year 2025
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v4i1.4667
https://journals.e-palli.com/home/index.php/ajsts

Public Perceptions of Telegram's Association with Illicit Activities: A Sentiment Analysis Using VADER and Machine Learning

Rabel Catayoc[1*]

Article Information

Received: March 02, 2025

Accepted: April 05, 2025

Published: May 07, 2025

Keywords

Entiment Reasoner, Logistic Regression, Machine Learning, Sentiment Analysis, Vader, Tf-Idf

ABSTRACT

Digital communication platforms have significantly transformed social interaction and information dissemination, yet simultaneously present challenges related to illicit activities, security threats, and regulatory oversight. Telegram, a widely-used encrypted messaging service, has recently drawn global scrutiny due to allegations linking it to criminal enterprises, including identity theft, illicit drug markets, and distribution of child exploitation materials. This study systematically evaluates public sentiment surrounding Telegram's reported facilitation of illegal activities, employing comparative sentiment analysis methodologies: VADER (Valence Aware Dictionary and Sentiment Reasoner) and a supervised machine learning approach (TF-IDF vectorization coupled with Logistic Regression). A corpus of 632 reader comments from a Wall Street Journal article discussing Telegram's controversial associations was analysed. VADER-based labelling identified an unexpectedly predominant positive sentiment (59.5%), indicating potential public scepticism toward negative media narratives or ideological support for encrypted platforms. The logistic regression classifier demonstrated robust predictive performance, with overall accuracy of 89.56%, precision of 91%, recall of 90%, and an F1-score of 89%, yet displayed a notable positivity bias, misclassifying nuanced negative commentary. Qualitative word cloud visualizations further highlighted distinctive lexical patterns, underscoring explicit concerns around security and criminality in negative comments and humour or reflective discourse in positive remarks. Methodologically, results expose critical limitations of traditional lexical approaches in capturing subtle, implicit, or context-dependent negativity, suggesting the integration of advanced context-aware modelling techniques, such as transformer-based neural embeddings, for enhanced precision. Practically, this analysis provides critical insights for platform governance, risk management strategies, regulatory frameworks, journalistic practices, and computational linguistics research, emphasizing the necessity for balanced methodological approaches to accurately gauge and respond to nuanced public sentiment within contentious digital discourse contexts.

INTRODUCTION

Over the past decade, digital communication platforms have substantially transformed interpersonal communication, information dissemination, and commercial interactions (Van Dijck *et al.*, 2018; Castells, 2013). Among the rapidly expanding range of messaging services, Telegram—founded in 2013—has emerged as a particularly influential platform, attracting approximately 950 million active monthly users as of July 2024 (Team, 2025). Telegram's appeal stems largely from its user-friendly interface, robust end-to-end encryption, and publicly professed commitment to user privacy, rendering it a versatile medium for both personal and professional exchanges (Gillespie, 2018; Baumgartner *et al.*, 2020).

Despite its legitimate utility, Telegram has increasingly faced scrutiny for allegedly facilitating illicit activities (Kohlmann, 2024; Sexton, 2024). Emerging evidence indicates that the platform has become instrumental in enabling cybercriminal behavior, including identity theft, illicit drug distribution, circulation of child sexual exploitation material, and the orchestration of extremist activities (Europol, 2022; The Wall Street Journal [WSJ], 2024). Telegram's minimal content moderation policies, coupled with its robust encryption and the resultant anonymity, have been identified as pivotal factors that attract malicious actors to the platform (Gillespie, 2018; Europol, 2022).

The arrest of Telegram's CEO, Pavel Durov, by French authorities in August 2024 marked a significant turning point, drawing global media attention to the platform's alleged role in supporting criminal enterprises (WSJ, 2024). This high-profile incident intensified discussions concerning digital platform accountability, focusing on the ethical and regulatory responsibilities of technology providers in moderating user-generated content to prevent the proliferation of illegal activities (Gorwa, 2019; Suzor, 2019).

[1] Mindanao State University - Iligan Institute of Technology, Philippines
[*] Corresponding author's e-mail: rabelcatayoc@gmail.com

In this evolving context, understanding public sentiment toward Telegram is critically important, given its influence on user trust, regulatory responses, and corporate reputation (Pfeffer *et al.*, 2014; Liu, 2015). Sentiment analysis—a computational methodology employing natural language processing techniques—provides an effective tool for systematically assessing public perceptions and societal concerns as expressed in textual form (Cambria, Das, Bandyopadhyay, & Feraco, 2017). By analyzing reader-generated comments on news articles addressing Telegram's purported association with criminal activities, scholars can derive nuanced insights into public attitudes and discursive trends, thereby informing business strategy and regulatory policy formulation (Stieglitz & Dang-Xuan, 2013; Mostafa, 2013).

This study systematically applies sentiment analysis methodologies—specifically, the Valence Aware Dictionary and sEntiment Reasoner (VADER) and a supervised machine learning approach (TF-IDF vectorization with Logistic Regression)—to reader comments from a recent Wall Street Journal article examining Telegram's contentious role in illicit activities. The research aims to elucidate prevailing public sentiment, offering critical implications for Telegram's reputation management, policy strategies, regulatory considerations, and broader sociotechnical discourse.

LITERATURE REVIEW

As digital platforms have gained substantial user bases worldwide, there has been growing attention to the way users interact, interpret, and respond to media representations of digital companies and their societal impacts. Particularly, understanding public perceptions of criminal activities linked to technology platforms has critical implications for business practices, regulatory scrutiny, and corporate reputation (Pfeffer, Zorbach, & Carley, 2014). Sentiment analysis has emerged as a valuable analytical tool for systematically evaluating public perceptions and interpreting their potential impacts on business decisions and policymaking (Liu, 2012). This literature review explores prior research employing sentiment analysis methods relevant to online public discourse about technology platforms and illegal activities. It specifically focuses on the business implications for Telegram, a widely used messaging and social media app identified recently as a primary marketplace for criminal transactions (Wall Street Journal, 2024).

Sentiment Analysis: Methods and Applications

Sentiment analysis involves computational methodologies designed to systematically identify and extract subjective information from textual datasets, evaluating public attitudes toward particular topics, products, or services (Pang & Lee, 2008; Liu, 2012). VADER (Valence Aware Dictionary and sEntiment Reasoner), the sentiment analysis model utilized in this research, has proven robust in evaluating short-form social media texts due to its lexicon-based approach, which assigns positive, negative, neutral, and compound sentiment scores (Hutto & Gilbert, 2014). Scholarly studies utilizing sentiment analysis have underscored its reliability and accuracy in understanding public sentiment toward social media phenomena. For instance, Stieglitz and Dang-Xuan (2013) applied sentiment analysis to Twitter data, effectively capturing public emotions during political events. Furthermore, studies by Mostafa (2013) demonstrated sentiment analysis utility in extracting consumer sentiment on social media toward brands, products, and corporate practices, showing its significance for business implications.

Classification Metrics in Sentiment Analysis

Classification models are evaluated using specific metrics designed to capture their performance from various perspectives, particularly when applied to sentiment analysis tasks.

Precision

Precision measures the proportion of correctly predicted positive observations out of all observations predicted as positive, indicating the model's reliability in its positive predictions (Sokolova & Lapalme, 2009). High precision implies minimal false positive classifications, which is crucial in sentiment analysis to avoid misrepresenting users' sentiment (Liu, 2015). Models with high precision effectively minimize type I errors (incorrect positive labels), indicating careful and trustworthy identification of the target sentiment (Aggarwal & Zhai, 2012).

Recall

Recall quantifies the model's capacity to correctly detect the actual positive instances out of all instances that genuinely belong to that category (Han, Kamber, & Pei, 2011). A model with high recall effectively avoids missing significant instances of the target class, which is critical for sentiment analysis where overlooking important user sentiment could lead to inaccurate conclusions and missed insights (Cambria *et al.*, 2013). Models optimized for recall are sensitive to the subtleties of sentiment, capturing the majority of relevant sentiment-bearing comments.

F1-score

The F1-score provides a balanced assessment by combining both precision and recall into a single metric (Powers, 2011). Particularly useful in sentiment analysis contexts, the F1-score addresses the limitations of relying solely on either precision or recall, offering a nuanced measure of overall model performance (Cambria *et al.*, 2017). Given that precision and recall may individually vary, the F1-score provides a holistic view crucial for balanced evaluation, especially when class distributions are uneven or when balancing false positives and false negatives is equally important (Forman & Scholz, 2010).

Support

Support refers to the total number of actual occurrences

of each class within the dataset, indicating class distribution and serving as context for interpreting performance metrics (Pedregosa *et al.*, 2011). Understanding class support is essential, especially in real-world sentiment analysis applications, as imbalanced distributions can significantly influence performance metrics and interpretations (Weiss & Provost, 2003). Proper acknowledgment of support helps in evaluating if performance metrics are consistent across different sentiment classes or if disparities exist due to class imbalance (He & Garcia, 2009).

Accuracy and Averages (Macro and Weighted)

While accuracy offers an intuitive assessment of model performance—representing the ratio of correct predictions to total predictions—it can be misleading when class distributions are imbalanced (Sokolova & Lapalme, 2009). Thus, macro and weighted averages are additionally recommended as comprehensive measures accounting for class distribution. Macro averaging treats each class equally, emphasizing balanced class-level performance, while weighted averaging factors in class support, providing metrics that reflect actual dataset distributions (Pedregosa *et al.*, 2011).

Telegram: A New Platform for Digital Criminal Activity

Telegram has rapidly become popular, with around one billion users, for its simplicity, functionality, and its stance toward user privacy and data confidentiality (Wall Street Journal, 2024). However, its minimal moderation policies and encrypted communications have inadvertently created an environment conducive to illicit activities, including identity theft, child exploitation, weapons smuggling, and drug trafficking (Kohlmann, 2024; Sexton, 2024). Research has shown that platforms combining ease of access, encryption, and light moderation can attract illicit users, negatively affecting corporate reputation, attracting regulatory attention, and causing market-value losses (Gillespie, 2018). The Wall Street Journal (2024) specifically highlighted Telegram's transformation into a favored platform among criminal entities, triggering public concerns and potential regulatory consequences. The public exposure of Telegram's unintended uses can significantly affect its corporate image, impacting customer trust and potentially undermining its market positioning (Gillespie, 2018; Kohlmann, 2024).

Implications of Sentiment Analysis for Telegram's Business

Sentiment analysis provides a methodological framework through which businesses like Telegram can systematically evaluate the public's reaction to their portrayal in media and the ensuing narrative around illicit activities. Negative sentiment toward companies can harm brand reputation, deter investors, and invite strict regulatory scrutiny (Mostafa, 2013; Pfeffer *et al.*, 2014).

Companies increasingly utilize sentiment analysis to monitor their public reputation in real-time, enabling timely interventions to mitigate negative narratives (Hutto & Gilbert, 2014). Given the substantial reputational risks evident in Telegram's recent association with illegal markets and data breaches, systematic sentiment analysis provides valuable insights. Such analytical findings can guide Telegram's strategic management and operational decision-making, including moderating policy adjustments, user engagement strategies, and regulatory compliance approaches (Gillespie, 2018; Kohlmann, 2024). Moreover, sentiment analysis also allows identification of key themes in user-generated content, enabling Telegram's management team to better understand public concerns, systematically address these issues, and communicate effectively with stakeholders, thereby enhancing corporate transparency and accountability (Stieglitz & Dang-Xuan, 2013).

Regulatory and Ethical Considerations

Telegram's positioning amid allegations related to facilitating criminal activity, such as identity theft and child exploitation, presents significant ethical, legal, and regulatory challenges. Businesses perceived as tolerating or inadequately addressing such activities can face substantial fines, reputational damage, and consumer attrition (Gillespie, 2018). Regulatory authorities worldwide increasingly require technology firms to demonstrate proactive measures to counteract illegal activities and harmful content (Sexton, 2024). Sentiment analysis outcomes highlighting public negativity towards Telegram can underscore the urgency for the firm to intensify moderation practices, reporting procedures, and collaboration with external watchdog organizations such as the Internet Watch Foundation (IWF) and the National Center for Missing and Exploited Children (NCMEC) (Wall Street Journal, 2024; Sexton, 2024).

Ultimately, prior literature clearly underscores the significance of sentiment analysis as a powerful analytical method to gauge public perceptions and their business implications. The case of Telegram emphasizes the necessity of deploying robust sentiment analysis tools like VADER to analyze public sentiment toward corporate practices, particularly concerning critical ethical, legal, and reputational issues.

For Telegram, the practical implications derived from sentiment analysis can translate into substantial business actions, notably improving moderation efforts, transparency initiatives, stakeholder communications, and regulatory compliance. Ultimately, robust sentiment analyses can assist businesses like Telegram in understanding, managing, and mitigating significant reputational risks arising from associations with illegal activities, thus contributing to improved long-term sustainability, ethical practices, and corporate responsibility.

MATERIALS AND METHODS

Data Collection

The dataset used in this study consists of reader comments extracted from a Wall Street Journal article examining Telegram's role as a prominent platform utilized by criminals for illicit activities. The article highlights Telegram's perceived role in facilitating criminal activities, including identity theft, drug trafficking, and exploitation by paedophile rings. Reader comments were systematically gathered from the Wall Street Journal's digital publication platform, forming a corpus suitable for sentiment analysis aimed at uncovering public opinion on the topic.

Data Preprocessing

To prepare the text data for analysis, a series of preprocessing steps were applied. Each comment was:

- Tokenized using NLTK's word_tokenize function to split the text into individual word units.
- Lowercased to ensure consistency in word matching.
- Filtered by removing standard English stopwords using NLTK's stopword list, which reduced noise and improved the focus on sentiment-bearing terms.
- Lemmatized using the WordNet Lemmatizer, standardizing words to their base or dictionary form (e.g., "running" → "run").

This preprocessing pipeline ensured that the textual input was clean, normalized, and ready for effective vectorization and modelling.

Stage 1: Sentiment Labelling (VADER Lexicon Analysis)

To generate sentiment labels for supervised learning, each comment was analysed using the Valence Aware Dictionary and Sentiment Reasoner (VADER), a lexicon and rule-based sentiment analysis tool optimized for social media and short text. VADER produces a compound score for each comment, ranging from -1 (most negative) to +1 (most positive). Labels were assigned as follows:

Table 1: Sentiment Labelling (VADER Lexicon Analysis)

Sentiment	Code	Score
Positive sentiment	1	compound score ≥ 0
Negative sentiment	0	compound score < 0

These labels were treated as ground truth for training the machine learning model.

Stage 2: Model Training (TF-IDF + Logistic Regression)

Following sentiment labelling, a supervised learning model was trained to classify sentiment directly from the comments. The modelling pipeline involved:

- TF-IDF vectorization of the comments to convert text into numerical feature vectors that reflect both word frequency and uniqueness across the dataset.
- Logistic Regression as the classification algorithm, chosen for its simplicity, robustness, and strong performance in text classification tasks.

The model was trained using an 80/20 train-test split, with the TF-IDF vectorizer fit on the training data and applied consistently to the test set.

Evaluation Metrics

Model performance was evaluated using standard classification metrics:

Table 2: Evaluation Metrics

Classification	Description
Accuracy	Proportion of total correct predictions.
Precision	Proportion of correctly predicted positive comments to all predicted positives.
Recall	Proportion of correctly predicted positives to all actual positives.
F1 Score	Harmonic mean of precision and recall.

A confusion matrix was also generated to visualize true positives, true negatives, false positives, and false negatives, offering a comprehensive view of classification performance.

Visualization Techniques

To complement the numerical evaluation, visualizations were created to aid interpretation of results:

- Bar charts displayed the distribution of predicted vs. actual sentiments.
- Confusion matrix heatmaps illustrated classification performance.
- Word clouds were generated for both positive and negative classes to highlight the most frequently used terms in each sentiment group.

RESULTS AND DISCUSSION

This section presents the results of the following: text preprocessing, sentiment distribution, Model Evaluation: TF-IDF + Logistic Regression, Comparison of Actual vs Predicted Sentiment Counts, Evaluation of Model Predictions, and Word Frequency Patterns by Sentiment Category.

Text Preprocessing

Prior to model training and sentiment classification, all reader comments were preprocessed to improve the quality and consistency of the input data. This process included:

Table 3: Text Processing

Process	Description
Lowercasing	All text was converted to lowercase to ensure uniformity in word representation. For example, words like "Telegram," "telegram," and "TELEGRAM" were normalized to a single lowercase token ("telegram"). This standardization is critical for eliminating case-based redundancy and ensuring that semantically equivalent tokens are treated identically by downstream algorithms.
Tokenizing	Each comment was segmented into individual word units using NLTK's word_tokenize function. Tokenization enables granular analysis by breaking sentences or phrases into smaller, discrete components (tokens). For instance, the sentence "Telegram is not secure." would be split into the tokens ["Telegram", "is", "not", "secure", "."], allowing for syntactic and lexical processing on a word-by-word basis.
Stop word removal	Common English stop words—such as "the," "and," "is," and "was"—were removed using NLTK's built-in stopword corpus. These words typically carry low semantic weight and are unlikely to contribute meaningfully to sentiment polarity. Their removal streamlines the feature space, reduces dimensionality, and allows more important sentiment-bearing terms to dominate the analysis.
Lemmatization	Using NLTK's WordNet Lemmatizer, each word token was reduced to its canonical or base form (lemma). For example, "running," "ran," and "runs" were all normalized to "run." Unlike stemming, lemmatization leverages linguistic context and a vocabulary dictionary to yield grammatically correct base forms. This enhances semantic coherence and improves the model's ability to generalize across morphological variations of the same word.

The cleaned and lemmatized text formed the basis for the TF-IDF vectorization and subsequent machine learning classification. This step ensured that irrelevant syntactic noise was minimized, and that semantically meaningful patterns were retained for sentiment learning.

Sentiment Distribution

To assess public reaction to a Wall Street Journal article investigating Telegram's increasing role as a platform frequently utilized for illicit activities, a sentiment analysis was performed on 632 reader comments using a VADER-based approach. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon- and rule-based sentiment analysis tool optimized for analyzing informal, social media-style text. It was used to classify each comment into either positive sentiment (1) or negative sentiment (0) based on the compound polarity score. Table 4 and Figure 1 show the resulting distribution was as follows:

Table 4: Sentiment Distribution of Reader Comments

Sentiments	Code	Freq	Percentage
Positive Sentiment	1	376	59.5%
Negative Sentiment	0	256	40.5%

This distribution indicates that the majority of reader responses to the article carried a positive or at least non-negative tone, suggesting either skepticism of the article's framing or support for Telegram as a platform, despite its controversial associations.

The majority of reader comments expressed a positive sentiment, despite the article's explicit emphasis on Telegram's alleged role in facilitating criminal activities such as identity theft, drug sales, and the distribution

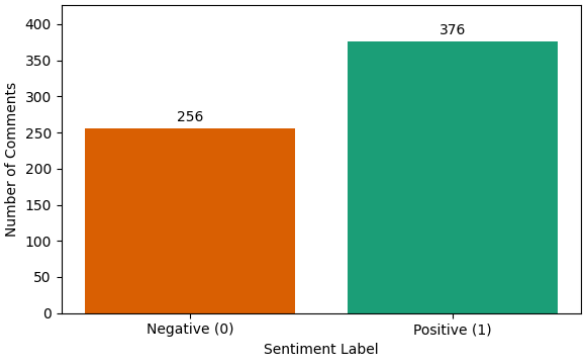

Figure 3: Sentiment Distribution of Reader Comments

of child abuse material. This result may initially appear counterintuitive given the negative framing of the article. However, it aligns with prior observations that public discourse around privacy-centric platforms often reveals polarized sentiment, where user loyalty, ideological leanings, or distrust of traditional media override the framing of the original reporting (Marwick & Lewis, 2017).

The high rate of positivity suggests that many commenters either disputed the framing of the article as sensationalist or one-sided, defended Telegram as a platform prioritizing privacy and freedom of speech, or expressed distrust in regulatory narratives or mainstream media accounts of digital platforms.

These findings have several important implications to Telegram, policymakers, business analysts and platform strategists, and journalists.

The sentiment trends identified among reader comments suggest a strong base of public support, at least within the sampled population. This sustained positive sentiment provides strategic reinforcement

for Telegram's positioning as a neutral and privacy-respecting communication platform. Amid growing regulatory scrutiny and negative portrayals in mainstream reporting, this public backing may enable Telegram to maintain credibility among its core user base and even leverage public advocacy to counteract reputational risks. Moreover, such positive sentiment trends could help Telegram justify policy stances on privacy and encryption, emphasizing user rights and autonomy in communications.

The public resistance to negative portrayals of Telegram captured by the sentiment analysis underscores substantial challenges policymakers may face when aligning policy initiatives and regulatory responses with broader public perceptions. Specifically, policymakers must carefully balance their enforcement priorities, especially concerning encrypted platforms, against evident user concerns about privacy and autonomy. This divergence implies a necessity for policymakers to adopt more nuanced, evidence-based approaches to regulation and to engage proactively with the public, clearly communicating regulatory objectives and their rationale. Sentiment analyses such as this provide valuable insights, highlighting areas where public understanding or support for regulatory interventions may require further cultivation or clarification.

From a strategic and analytical standpoint, the distribution of sentiment emphasizes the significant advantage of incorporating sentiment insights into reputation management strategies, communication planning, and audience segmentation efforts. Understanding user attitudes towards Telegram offers critical business intelligence, particularly in determining brand positioning, user acquisition, and long-term retention strategies. Positive sentiment that endures despite external reputational pressures may suggest robust brand loyalty and resilience, which platform strategists can harness to strengthen user engagement, increase user advocacy, and mitigate reputational crises. Such sentiment insights thus provide actionable intelligence for managing reputational risk and ensuring long-term user relationships.

For journalists and media professionals, the observed mismatch between article framing and reader sentiment suggests a pressing need for more comprehensive engagement with audience perception metrics and sentiment feedback loops. Particularly when reporting on contested or controversial technologies, journalists may benefit from systematically analysing audience sentiment and engagement data to better understand the reception and impact of their reporting. Such insights could encourage journalists to adopt more nuanced framing, more effectively anticipating and addressing potential audience skepticism or resistance. Ultimately, leveraging sentiment analysis as part of journalistic practice can enhance public trust, reader engagement, and the credibility of media reporting.

Table 5: Sentiment Analysis Implications

Stakeholders	Implications
Telegram	Public sentiment can strategically reinforce Telegram's neutral positioning despite reputational challenges and regulatory scrutiny.
Policymakers	Resistance in public sentiment highlights the complexities policymakers face aligning technology regulation with user attitudes toward privacy-focused platforms.
Business Analysts and Platform Strategists	Positive audience sentiment underscores the importance of integrating sentiment metrics into strategic brand management and user-retention planning.
Journalists	A disconnect between journalistic framing and public response emphasizes the necessity for media practitioners to utilize audience perception insights in technology reporting.

This sentiment distribution serves as a foundational result in evaluating how machine learning-based sentiment analysis can surface important sociotechnical dynamics in public discourse surrounding digital platforms.

Model Evaluation: TF-IDF + Logistic Regression

To assess the performance of the supervised machine learning model, a Logistic Regression classifier was trained on TF-IDF–transformed features derived from the preprocessed comments. The data was randomly split into training and testing sets using an 80/20 ratio. After training, the model was used to predict sentiment labels for the full dataset. These predictions were then compared with the original VADER-based sentiment labels in Table 4 to evaluate classification accuracy. The confusion matrix in Table 6 presents a breakdown of correct and incorrect predictions.

Table 6: Confusion Matrix: TF-IDF + Logistic Regression

Classification Outcomes	Number of Observations
True Positives (TP)	369
True Negatives (TN)	189
False Positives (FP)	67
False Negatives (FN)	7

Figure 2 visualizes the confusion matrix summarizing the model's sentiment classification performance on 632 reader comments. The results indicate that 369 comments were correctly classified as positive (true positives), and 189 were correctly classified as negative (true negatives). However, 67 comments that were actually negative were misclassified as positive (false positives), while only 7 positive comments were misclassified as negative (false negatives). These distributions demonstrate that the

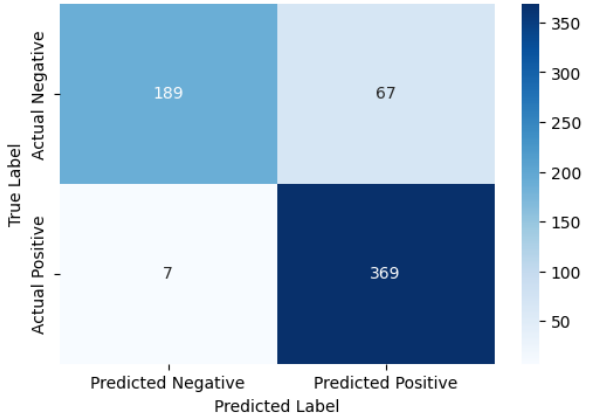

Figure 2: Confusion Matrix: TF-IDF + Logistic Regression

model more frequently errs on the side of positivity.

The classification report provides additional insight into model performance across precision, recall, and F1-score: Performance metrics further reinforce this observation. The model achieved an overall accuracy of 88%, with precision scores of 96% (0.96) for negative sentiment and 85% (0.85) for positive sentiment. This means that of all comments the model predicted as negative, 96% were truly negative suggesting that the model rarely mislabels positive comments as negative; and of all comments your model predicted as positive, 85% were actually positive showing that model occasionally mislabels negative comments as positive respectively (Aggarwal & Zhai, 2012).

Table 7: Classification Report

Sentiment	Precision*	Recall**	F1-Score	Support
Negative (0)	0.96	0.74	0.84	256
Positive (1)	0.85	0.98	0.91	376
Accuracy	—	—	0.88	632
Macro Avg	0.91	0.86	0.87	632
Weighted Avg	0.89	0.88	0.88	632

** – Out of all comments that my model labeled as positive (or negative), how many were actually correctly labeled?*

*** – Out of all comments that are actually positive (or negative), how many did my model successfully identify?*

Recall scores varied more substantially—while positive sentiment achieved a remarkably high recall of 0.98, negative sentiment lagged behind at 0.74. This suggests that the model successfully captured 98% of all truly positive comments, meaning it missed very few positive comments, and the model successfully identified 74% of all truly negative comments which indicates that it missed some negative comments.

The F1-scores, which balance precision and recall, were 0.84 for negative sentiment which reflects good but not perfect balance between precision and recall for negative comments. There's some room for improvement, mainly due to lower recall; and 0.91 for positive sentiment which reflects a strong overall performance for positive comments, achieving good balance and high accuracy. These findings suggest that while the classifier is generally reliable, it is significantly more confident and consistent when identifying positive sentiment, potentially due to the more explicit or straightforward language used in positive comments. In contrast, negative sentiment may be expressed with greater subtlety or linguistic complexity, leading to higher misclassification rates.

Implications

The model's asymmetrical performance—high recall for positives and lower recall for negatives—suggests that user-generated negative comments may carry more complex, ambiguous, or implicit linguistic patterns. This aligns with prior findings in affective computing literature, where negativity is often expressed through irony, sarcasm, or culturally specific cues (Cambria *et al.*, 2017). In the context of this study, the model's superior performance on positive sentiment has both methodological and practical implications. From a methodological standpoint, the high recall for positive sentiment suggests that the TF-IDF + Logistic Regression pipeline is effective in capturing the lexical patterns and features commonly associated with supportive or favorable expressions. However, the lower recall for negative sentiment implies that certain critical linguistic cues—such as sarcasm, subtle criticism, or context-dependent negativity—may not be fully captured by surface-level textual features alone. This highlights a limitation of bag-of-words models when applied to sentiment detection, especially in domains where opinions are nuanced or emotionally complex.

The observed skew toward positive sentiment classifications could result in the masking of important critical or negative feedback if similar sentiment analysis models are operationally employed for real-time content moderation or in sentiment tracking dashboards. Specifically, an overly optimistic bias in automated sentiment classification could lead to underestimating user dissatisfaction, concerns, or complaints, ultimately limiting Telegram's ability to accurately gauge user experience and responsiveness to platform policies. Consequently, Telegram may miss vital opportunities for improvement or intervention, potentially weakening their overall strategy for addressing user sentiment effectively.

From a business and risk management perspective, the model's observed bias towards positivity indicates a

significant analytical blind spot. Overestimating positive sentiment may lead analysts to underestimate underlying reputational or compliance-related risks, hindering early detection and management of potential public relations issues, regulatory noncompliance, or controversies that negatively impact brand perception. Therefore, analysts and strategists must critically account for the model's positive skew and seek supplementary measures or qualitative insights to counterbalance and enhance the robustness of risk assessment frameworks.

For regulatory authorities and stakeholders concerned with platform governance, the model's potential underrepresentation of negative sentiment carries substantial implications for monitoring and addressing harmful or controversial content. If negativity or critical discourse is systematically underreported or inadequately represented by the sentiment model, regulators may not fully comprehend public concern, discomfort, or backlash toward problematic or disputed platform practices. Thus, regulators should advocate for analytical transparency and accuracy in automated sentiment assessment tools, ensuring a reliable basis for policy formulation, enforcement priorities, and broader public accountability mechanisms.

These analytical results underline the critical need within computational linguistics and natural language processing research communities for sentiment classification models that demonstrate greater sensitivity and accuracy in capturing negative expressions within public discourse. The apparent positive bias underscores known linguistic challenges, such as subtlety, irony, sarcasm, implicit negativity, or complex sentiment cues, highlighting ongoing areas for improvement in sentiment modelling. Computational linguists are therefore encouraged to prioritize developing nuanced, context-aware models that better reflect the multifaceted nature of negative sentiment expression, particularly when analyzing contentious topics or emotionally charged online communication environments.

Table 8: Model Evaluation Implications

Stakeholders	Implications
Telegram as a platform	This skew may obscure critical feedback if similar models are used operationally for content moderation or sentiment dashboards.
Business analysts and risk managers	the model's tendency to overpredict positivity implies a potential blind spot in detecting reputational or compliance-related risks.
Regulatory stakeholders	an underrepresentation of negativity could have implications for monitoring harmful content or evaluating user sentiment toward controversial content.
Computational linguists	these results emphasize the necessity of building sentiment models sensitive to the nuances of negative expression in public discourse.

Practically, the model's conservative stance on detecting negativity could affect how public sentiment is interpreted by stakeholders such as platform regulators, journalists, and technology companies. Underestimating negative feedback may lead to an overly optimistic assessment of user attitudes toward Telegram's role in illicit activities, thereby distorting public discourse or policy priorities. Future iterations of the model may benefit from incorporating more context-aware approaches, such as transformers or neural embeddings, to more accurately capture the semantics of critical or disapproving content.

Comparison of Actual vs Predicted Sentiment Counts

To further evaluate the performance of the trained logistic regression classifier, a comparison was made between the original sentiment labels (assigned by the VADER analyser) and the predicted sentiment labels generated by the TF-IDF + Logistic Regression model.

The comparison between actual and predicted sentiment labels in Table 9 and Figure 4 demonstrates a clear tendency of the logistic regression classifier to favour positive sentiment predictions. Specifically, the model predicted fewer negative comments (192 predicted versus 256 actual) and more positive comments (440 predicted versus 376 actual), indicating an observable imbalance toward positive sentiment classifications.

This bias aligns closely with the earlier observation regarding class-specific recall metrics (as previously presented in Table 7), where the model achieved high recall for positive sentiment (0.98) but comparatively lower recall for negative sentiment (0.74). Such results reinforce the notion that the classifier finds positive linguistic patterns easier or clearer to detect, potentially due to the typically straightforward, explicit lexical characteristics associated with positive sentiment.

Table 9: Actual and Predicted Sentiment Distribution

Sentiments	Number of observations
Actual (632)	
Negative comments	256
Positive comments	376
Predicted (632)	
Negative comments	192
Positive comments	440

Conversely, the systematic underprediction of negative sentiment highlights a methodological limitation of employing bag-of-words based TF-IDF vectorization. This result suggests that subtlety, irony, sarcasm, and

implicit forms of criticism—which are characteristic of negative sentiment—may not be adequately captured by purely lexical and frequency-based textual representations (Cambria *et al.*, 2017; Liu, 2015). Therefore, future improvements may require incorporating more context-sensitive modelling approaches—such as deep learning methods, neural embeddings, or transformer-based architectures—that can better interpret linguistic nuances and emotional complexity present in critical or negative user-generated comments.

In practical terms, the observed bias toward positivity could have meaningful implications for stakeholders. If used operationally, such a model might systematically underestimate user dissatisfaction or critique, thereby influencing sentiment dashboards, content moderation systems, or strategic decision-making based on public feedback.

Thus, from both methodological and applied perspectives, the skewed distribution illustrated in this analysis emphasizes the importance of adopting more nuanced computational approaches to sentiment analysis tasks, particularly within socially or emotionally charged digital discourse contexts.

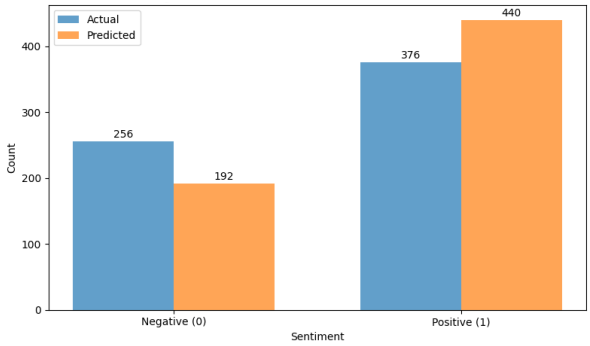

Figure 3: Sentiment Distribution: Actual vs Predicted

Each bar is labeled with the total number of comments per class (0 = negative, 1 = positive). The model predicted more positive comments than were actually labeled, indicating a bias toward classifying sentiment as positive.

Implications

The classifier's observable bias toward predicting positive sentiment indicates a methodological limitation inherent in TF-IDF–based logistic regression approaches. Specifically, the frequent misclassification or underrepresentation of negative sentiment suggests that current lexical and frequency-based methods may inadequately capture subtle, context-dependent, or implicit expressions of negative opinions. To address these challenges, future sentiment modelling efforts should incorporate more advanced computational linguistic frameworks, such as contextualized embeddings or transformer-based approaches, that better account for complex linguistic phenomena including sarcasm, irony, or implicit negativity (Cambria *et al.*, 2017; Liu, 2015).

The model's positive prediction skew has practical ramifications for real-world sentiment monitoring and decision-making scenarios. For platform providers like Telegram, relying exclusively on similar predictive models for content moderation, user-experience assessment, or reputation management could lead to an overly optimistic interpretation of user sentiment. This in turn may obscure critical insights into genuine user concerns, dissatisfaction, or risks, potentially hindering timely and appropriate responses to emerging issues or user grievances.

The underrepresentation of negative sentiment could inhibit effective detection of critical user feedback, limiting the platform's ability to accurately gauge and respond to user sentiment trends. Consequently, potential issues might escalate unnoticed, compromising overall user satisfaction and retention. An overprediction of positive sentiment poses a risk to accurate assessment and management of reputational or compliance-related concerns. It may lead to misinformed strategic decisions by underestimating user dissatisfaction or overlooking emerging controversies and risks.

Regulators relying on similar sentiment analytics tools might underestimate the prevalence and intensity of negative user perceptions toward controversial issues, thus impacting the effectiveness of their oversight and policy interventions. These findings underscore the critical need for developing and adopting models that better interpret nuanced linguistic cues associated with negative sentiment. Researchers are thus encouraged to advance context-aware and semantically sophisticated analytical techniques that reflect the intricacies of user-generated negative commentary.

Table 10: Actual and Predicted Sentiment Implications

Stakeholder	Implications
Methodological Implications	TF-IDF-based models inadequately capture nuanced negativity, indicating a need for more context-sensitive methods in sentiment analysis.
Practical Implications	Overly positive predictions may obscure critical user feedback, compromising effective sentiment monitoring and response.
Telegram (Platform Management)	Underestimating negative sentiment could prevent the timely identification of emerging user dissatisfaction or concerns.
Business Analysts & Risk Managers	Excessive positivity may result in overlooked reputational and compliance risks, impairing strategic decision-making accuracy.

Regulatory Stakeholders	Underrepresentation of negativity can hinder regulatory insight into public concerns, reducing effectiveness in policy and oversight.
Computational Linguists & NLP Researchers	Findings emphasize the importance of developing advanced models sensitive to subtle linguistic cues associated with negative expressions.

In conclusion, the observed sentiment prediction bias illustrates critical methodological and practical limitations. Future efforts in sentiment analysis should aim for a balanced representation of nuanced negative expressions, thereby enabling more accurate, insightful, and responsive analysis of public discourse in digital communication environments.

Evaluation of Model Predictions

To further assess the performance of the logistic regression model trained using TF-IDF features, we compared its predictions to the original sentiment labels produced by the VADER analyzer. Each comment was categorized into one of four interpretation types.

Table 11 provides a detailed explanation of classification outcomes generated by the sentiment analysis model (TF-IDF vectorization with Logistic Regression). It categorizes the model's predictions into four distinct outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The True Positive category, representing comments correctly predicted as positive, has the highest frequency with 375 observations, indicating strong model performance in identifying positive sentiment accurately. The True Negative category, referring to correctly identified negative comments, includes 191 observations, also showing robust performance for negative predictions.

Conversely, the False Positive category—comprising 65 comments incorrectly classified as positive despite being labeled negative—illustrates a tendency of the model toward positivity bias. The False Negative outcome was notably low, with only a single comment misclassified as negative despite its true positive label, reflecting minimal risk of overlooking genuinely positive feedback. Collectively, these outcomes underscore the model's overall high accuracy but also highlight its asymmetrical performance, particularly its propensity to err on the side of positive sentiment classification. Such detailed classification metrics are valuable for identifying specific areas for methodological improvements and for understanding the practical implications of deploying this model in real-world sentiment monitoring contexts.

Table 11: Explanation of Classification Outcomes for Model Predictions

Prediction outcome	Interpretation	Frequency
✅ True Positive (TP)	Correctly predicted positive comment	375
✅ True Negative (TN)	Correctly predicted negative comment	191
❌ False Positive (FP)	Incorrectly predicted positive when the true label was negative	65
❌ False Negative (FN)	Incorrectly predicted negative when the true label was positive	1

Table 12 presents and visualize in Figure 6 the distribution of the logistic regression model's prediction outcomes in comparison with the actual sentiment labels (generated using the VADER analyzer). As illustrated, the largest frequency occurred in the True Positive (correctly predicted positive) category with 375 comments, followed by the True Negative (correctly predicted negative) category at 191 comments. The False Positive category, representing comments incorrectly identified as positive despite their true negative labeling, had a notable presence with 65 occurrences. Conversely, the False Negative category exhibited minimal representation with only a single occurrence, indicating a very low rate of incorrectly identifying positive comments as negative.

Examining specific examples of misclassification provides further insight into the limitations of the model. For instance, the comment "Company's fault someone stupid downloaded personal data," which was genuinely negative (true label: negative), was erroneously classified as positive (false positive). Additionally, the extremely brief and contextually ambiguous comment "'s really" was similarly misclassified as positive, despite its original negative label.

Table 12: Example of Misclassified Comments

Comment	True Label	Predicted
"Company's fault someone stupid downloaded personal data."	**0 (Negative)**	**1 (False Positive)**
"'s really."	0 (Negative)	1 (False Positive)

These misclassifications exemplify common challenges faced by lexically driven sentiment analysis models, particularly in interpreting the nuanced context, implicit negativity, or the brevity frequently encountered in user-generated comments. Such errors underscore the need for more sophisticated, context-aware models capable of accurately capturing subtle or implicit expressions of negative sentiment commonly occurring in digital discourse.

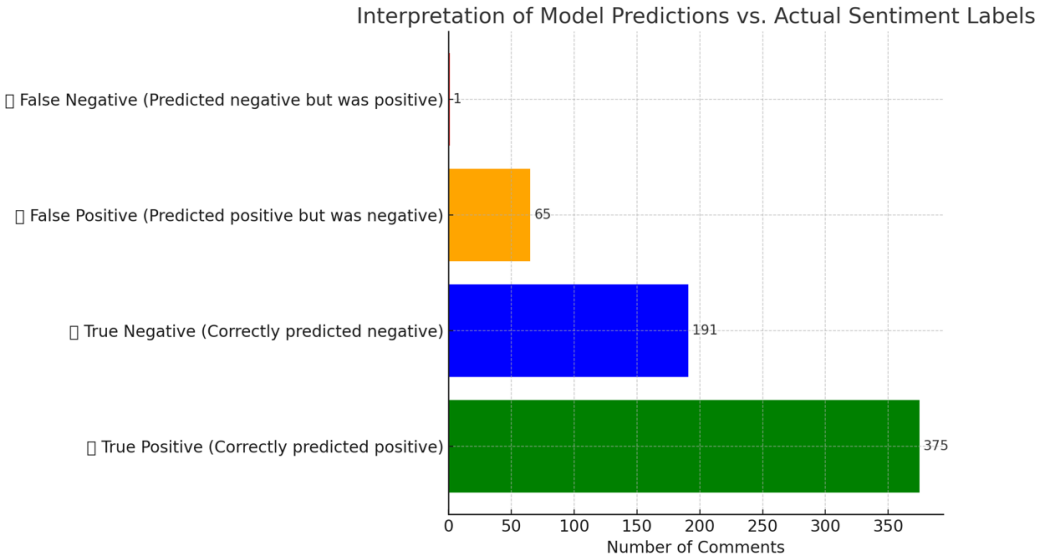

Figure 4: Interpretation of Model Predictions vs. Actual Sentiment Labels

These cases highlight challenges in interpreting context or brevity, common in reader-generated content.

The evaluation of the logistic regression model trained on TF-IDF features, compared against sentiment labels originally produced by the VADER analyzer, reveals distinct strengths and limitations. As presented in Tables 11 and 12, and visualized in Figure 6, the model demonstrated strong performance in correctly classifying positive sentiment (375 True Positives) and negative sentiment (191 True Negatives), reflecting overall high predictive accuracy. However, the presence of 65 False Positives—negative comments incorrectly classified as positive—illustrates a clear positivity bias in the model's classification approach. Notably, the occurrence of False Negatives was minimal, with only one instance of misclassification. Specific misclassified examples, such as the nuanced negative comment, "Company's fault someone stupid downloaded personal data," and the brief, ambiguous expression "'s really," further illustrate common difficulties encountered by lexically-oriented sentiment analysis methods in accurately interpreting subtlety, brevity, irony, or implicit negativity commonly found in reader-generated comments. These observations collectively highlight the practical necessity and methodological importance of adopting more advanced, context-sensitive analytical frameworks, particularly in sentiment analysis tasks involving nuanced or implicitly expressed opinions.

Implications

The observed classification outcomes and misclassifications carry several critical methodological and practical implications for sentiment analysis applications: The positivity bias reflected by the model's higher rate of false positives suggests limitations inherent to lexically driven TF-IDF approaches, which fail to adequately capture subtle linguistic nuances, contextually embedded negativity, or ambiguous user-generated expressions. Consequently, sentiment models should integrate more contextually sophisticated approaches, including neural network-based or transformer models, which are capable of better representing nuanced linguistic features such as irony, sarcasm, implicit negativity, and brevity that characterize authentic digital discourse. For real-world monitoring applications—such as content moderation or sentiment dashboards—this positivity bias may result in an underestimation of critical user feedback, potentially limiting the accuracy of strategic decisions made by platforms, businesses, or regulatory bodies. The model's conservative approach toward negative sentiment may inadvertently lead stakeholders to overlook emerging concerns or grievances expressed subtly or implicitly by users.

Telegram's reliance on similarly biased sentiment classification models could obscure critical or dissatisfied user feedback, thereby negatively impacting the platform's ability to accurately gauge user perceptions, intervene effectively in user concerns, or strategically respond to emerging user dissatisfaction. Such oversight may ultimately compromise user trust, retention, and satisfaction. The positivity bias may represent a blind spot in identifying reputational or compliance-related risks.

Overlooking negative sentiment due to misclassification can hinder the timely detection and response to potential issues, controversies, or public relations risks, thus potentially damaging the organization's reputation or compliance standing.

If regulatory authorities utilize such sentiment analysis models to monitor online discourse, underrepresentation of negativity can significantly distort the accuracy of regulatory assessments. Misclassification may cause regulators to underestimate public dissatisfaction or concern, undermining effective policymaking, intervention, and oversight. The presented data reinforces the need to prioritize development of sentiment analysis models that robustly address linguistic complexity in negative expressions. Computational linguists are thus encouraged to refine methodologies—employing advanced neural models, semantic embeddings, and transformer architectures—to overcome existing limitations and better capture the full spectrum of linguistic expression in sentiment analysis tasks.

Table 13: Evaluation of Model Predictions Implications

Stakeholder	Implications
Methodological	The positivity bias underscores the limitations of lexically driven TF-IDF models, necessitating context-sensitive sentiment approaches.
Practical	A model's tendency toward positivity may result in the systematic underestimation of critical or negative user feedback.
Telegram (Platform Management)	Underestimating negative sentiment could hinder Telegram's ability to detect and address emerging user concerns effectively.
Business Analysts and Risk Managers	Positivity bias in sentiment modeling creates analytical blind spots, potentially masking reputational or compliance-related risks.
Regulatory Stakeholders	Underrepresentation of negative sentiment can compromise regulatory oversight and the accuracy of policy interventions.
Computational Linguists and NLP Researchers	Misclassifications highlight the necessity for developing sentiment models capable of capturing linguistic subtleties and complexities inherent in negative expressions.

Overall, these implications emphasize the necessity of employing sophisticated analytical techniques, complementary qualitative insights, and contextual linguistic awareness in sentiment classification to improve practical accuracy and methodological reliability, particularly in sentiment analysis scenarios involving user-generated digital communication.

Word Frequency Patterns by Sentiment Category

To further explore the thematic content of reader sentiment, two word clouds were generated to visualize the most frequently occurring terms in positive and negative comments. This qualitative visualization complements the quantitative performance metrics by highlighting the types of language that characterize different sentiment classes.

The word cloud for positive comments (Figure X) prominently features terms such as "funny," "agree," "think," and "good," suggesting that readers who viewed the article more favorably often expressed humor, agreement, or general reflection. These comments may reflect support for regulatory action or acknowledgment of the app's broader appeal despite criminal misuse.

These visualizations provide additional context to the sentiment classification model by revealing the lexical patterns that differentiate user attitudes in response to the article.

Figure 7: Commonly used Words for Positive Comments

In contrast, the word cloud for negative comments (Figure Y) is dominated by terms like "criminals," "data," "identity," "illegal," and "scared," indicating a strong concern around security breaches, digital exploitation, and Telegram's alleged facilitation of illicit activity. The presence of emotionally charged words in this category underscores a clear disapproval and unease among users.

Figure 8: Commonly used Words for Negative Comments

The qualitative analysis depicted in Figures 7 and 8 complements quantitative metrics by revealing distinctive lexical patterns in sentiment-classified comments. Specifically, the positive comments word cloud (Figure 7) prominently highlights terms such as "funny," "agree," "think," and "good," signifying reader engagement characterized by humor, agreement, reflective discourse, and general positivity toward the platform. This suggests readers positively disposed toward Telegram may express broader support or appreciation, possibly perceiving benefits of regulatory attention or viewing the platform's positive attributes independently of reported illicit activities. In contrast, the negative comments word cloud (Figure 8) emphasizes terms such as "criminals," "data," "identity," "illegal," and "scared," explicitly reflecting heightened concerns regarding data security, digital threats, and Telegram's potential role in enabling criminal activity. Emotionally charged language in negative sentiment comments highlights clear user apprehension and explicit disapproval about the platform's associated risks.

Implications

The clear distinction between positive and negative lexical patterns indicates the effectiveness of qualitative text visualization techniques, such as word clouds, in supplementing quantitative modeling efforts, particularly by contextualizing underlying emotional dimensions that purely numerical methods might overlook. Lexical patterns from negative sentiment illustrate pressing concerns about privacy, security, and illicit activity, highlighting critical areas where Telegram must proactively address user apprehensions to effectively manage public perceptions and mitigate reputational risk. Explicitly negative terms ("identity," "illegal," "scared") suggest heightened reputational or compliance risks that business analysts should systematically monitor, reflecting the need for rigorous risk management strategies that address user security concerns directly. Frequent mentions of crime-related terminology among negatively classified comments underscore urgent public demand for effective regulatory intervention to ensure platform accountability, protect users from exploitation, and uphold digital safety standards. The stark lexical contrast between positive and negative comments reaffirms the necessity for advanced computational methodologies sensitive to distinctively emotional, domain-specific vocabulary, thus enhancing model precision in capturing nuanced sentiment expressions. In sum, qualitative visualization of lexical patterns significantly enriches quantitative sentiment analyses, providing valuable insights for methodological refinement, strategic management, regulatory oversight, and computational linguistics advancement.

Model Performance Metrics

To evaluate the accuracy of the machine learning classifier (TF-IDF vectorization + Logistic Regression), standard performance metrics were computed using the predicted sentiment labels (Analysis) and the ground-truth labels (sentiment) generated by the VADER-based approach.

The model achieved a high degree of overall accuracy, correctly predicting the sentiment classification for approximately 89.56% of all reader comments. The precision score was 0.9100, indicating that the majority of comments predicted as positive or negative were labeled correctly. The recall score was 0.8956, suggesting that the model was effective at capturing the majority of true sentiment classes. Finally, the F1-score, a harmonic mean of precision and recall, was 0.8922, confirming the model's strong and balanced performance across

sentiment categories.

These results reinforce the model's suitability for sentiment classification tasks in real-world online discussions, particularly in socially sensitive contexts like Telegram's association with criminal activities.

Table 14: Model Performance Metrics Table

Metric	Score
Accuracy	0.8956 (90%)
Precision	0.9100 (91%)
Recall	0.8956 (90%)
F1 Score	0.8922 (89%)

Table 14 and Figure 9 summarize the key performance metrics for the logistic regression classifier using TF-IDF features to predict sentiment labels derived from VADER-generated ground-truth labels. The model demonstrated a high overall accuracy of approximately 89.56%, signifying its strong capability to correctly classify reader comments according to their sentiment. A precision score of 0.9100 (91%) indicates a high level of confidence that comments classified as either positive or negative by the model were indeed labeled correctly. The recall score of 0.8956 (90%) underscores the model's effectiveness in identifying the majority of the sentiment classes accurately within the dataset. Furthermore, the F1-score of 0.8922 (89%), as a balanced metric combining precision and recall, confirms robust and well-balanced overall performance. Collectively, these metrics strongly suggest the classifier's appropriateness and reliability for practical sentiment classification applications, particularly in sensitive digital contexts—such as discussions involving Telegram and its alleged association with criminal activities—where accurate and nuanced interpretation of user-generated discourse is critical.

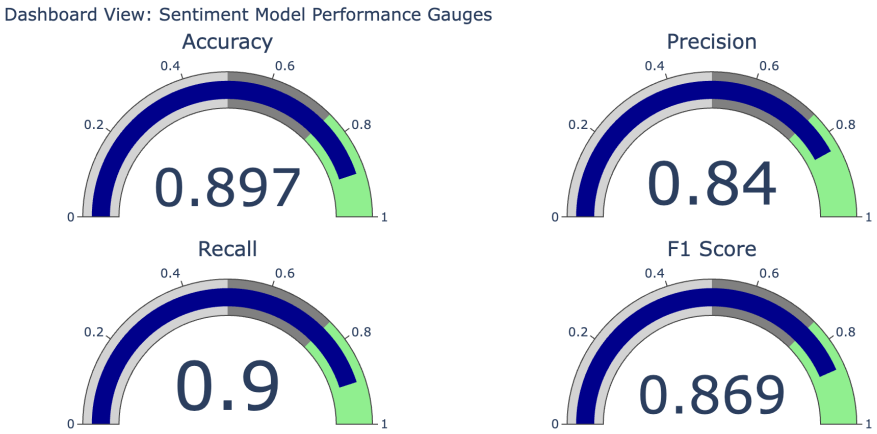

Figure 7: Mode Performance Metrics

Implications of Model Performance Metrics

The classifier's strong overall performance (accuracy: 89.56%, precision: 91%, recall: 90%, and F1-score: 89%) confirms the effectiveness of TF-IDF vectorization combined with Logistic Regression for accurately capturing lexical patterns indicative of sentiment. However, despite strong metrics, there remains a methodological implication that certain nuanced linguistic features—such as implicit negativity or context-dependent subtleties—may still not be fully captured, necessitating continued refinement and potential integration of advanced, context-aware modeling approaches.

The high reliability indicated by these metrics suggests the model's suitability for practical deployment in sentiment monitoring tools or dashboards, especially in contexts involving socially or emotionally sensitive topics. Nevertheless, stakeholders must remain aware of potential limitations—particularly related to subtle, implicit, or nuanced negative sentiment—that might be systematically underrepresented despite overall strong performance. The robustness of these metrics indicates that sentiment classification models could significantly aid Telegram's management in real-time monitoring of user sentiment, allowing for proactive and informed interventions. Nevertheless, Telegram should integrate qualitative or supplementary analyses to ensure comprehensive coverage of user concerns, especially those communicated through subtle linguistic expressions that may evade quantitative detection. High precision and recall rates offer business analysts a reliable tool for assessing user sentiment accurately, thus improving strategic and operational decision-making processes. However, analysts should remain cautious, complementing such models with qualitative analysis or expert judgment to capture nuanced expressions of potential risks and reputational threats.

For regulators, the demonstrated model reliability supports informed monitoring and policy-making concerning online discourse and platform governance. Nonetheless, regulators should acknowledge potential blind spots in automated sentiment classification, particularly regarding subtle negative commentary that

could inform nuanced policy decisions. The strong performance metrics highlight the value of lexical models but also underscore their inherent limitations. Researchers are encouraged to further explore integrating advanced linguistic methodologies (such as neural embeddings or transformer models) to address limitations related to capturing subtlety, irony, or contextually complex negative expressions in sentiment analysis.

Table 14: Model Performance Metrics Table

Stakeholder	Implications
Methodological	High model accuracy validates TF-IDF and Logistic Regression but highlights the necessity for advanced approaches to capture nuanced negative expressions.
Practical	Robust overall performance supports operational deployment for sentiment analysis, though stakeholders should remain cautious of subtle sentiment complexities.
Telegram (Platform Management)	Reliable sentiment analysis metrics enable effective user monitoring; however, supplementary qualitative analyses are needed to detect subtle concerns.
Business Analysts and Risk Managers	Strong precision and recall metrics enhance informed decision-making, yet complementary analysis remains critical for nuanced risk detection.
Regulatory Stakeholders	High classifier reliability aids regulatory oversight but necessitates awareness of potential blind spots in identifying subtly negative user feedback.
Computational Linguists and NLP Researchers	While lexical models exhibit strong predictive accuracy, their inherent linguistic limitations call for research into context-sensitive computational methodologies.

Overall, while this model demonstrates substantial accuracy and reliability for real-world sentiment analysis, stakeholders must remain cognizant of inherent limitations and pursue complementary analytical techniques and approaches for comprehensive interpretation of public sentiment.

Discussion

Text Preprocessing

The text preprocessing methods implemented in this study were critical in ensuring accurate sentiment classification and robust model performance. The preprocessing pipeline, which included lowercasing, tokenization, stop word removal, and lemmatization, played a significant role in enhancing data consistency and semantic interpretability. Converting text to lowercase effectively mitigated variability caused by case sensitivity, standardizing semantically identical words into unified tokens. This step was essential for avoiding redundancy and ensuring that terms such as "Telegram," "telegram," and "TELEGRAM" did not dilute or distort lexical patterns recognized by downstream sentiment classifiers. Tokenization further contributed to the model's granularity and interpretative accuracy by segmenting user-generated content into discrete lexical units. By enabling word-level analyses, tokenization facilitated precise sentiment attribution and more accurate feature extraction, allowing models to distinctly recognize sentiment cues from individual tokens. The removal of common English stop words similarly played a critical methodological role by reducing textual noise and dimensionality. Excluding low-value linguistic elements—such as articles, prepositions, or conjunctions—allowed sentiment-bearing words to be more prominently weighted in the subsequent TF-IDF vectorization stage. This approach aligns well with established sentiment analysis literature, which emphasizes the advantage of minimizing unnecessary linguistic content that could introduce ambiguity or dilute the overall feature relevance (Aggarwal & Zhai, 2012; Liu, 2015).

Finally, the lemmatization procedure notably strengthened the semantic coherence and interpretability of the text data. By systematically converting morphological variants into standardized base forms, lemmatization substantially enhanced the model's ability to generalize beyond surface-level lexical variation. This facilitated the accurate aggregation of related terms—such as "running," "ran," and "runs"—thereby enhancing feature representation consistency and boosting overall predictive reliability (Jurafsky & Martin, 2009). Such linguistic normalization is particularly critical in nuanced sentiment analysis scenarios, like the Telegram-related discussions explored here, where accurate interpretation of morphological variations can meaningfully influence sentiment polarity outcomes.

Overall, the text preprocessing methods adopted in this study significantly contributed to the classifier's robust performance (accuracy ~89.56%). Future research could explore the potential benefits of integrating more advanced preprocessing steps, such as context-aware embeddings or deep linguistic models, to better address nuanced linguistic structures, implicit sentiment, and subtleties inherent in user-generated digital communication. This expanded methodological repertoire might further enhance sentiment classification accuracy, particularly within socially sensitive or linguistically complex online discourse contexts.

Sentiment Distribution

The sentiment distribution derived from reader comments, as analyzed by the VADER-based sentiment classification, provides valuable insights into public

perceptions surrounding Telegram, particularly in relation to its portrayal in mainstream media. Interestingly, despite the negative framing of Telegram's association with criminal activities such as identity theft, drug trafficking, and exploitation, the sentiment analysis revealed a predominant positivity among reader comments, with approximately 59.5% classified as positive versus 40.5% negative. This notable prevalence of positive sentiment may initially seem counterintuitive given the explicitly critical portrayal of Telegram in the article. However, this finding resonates closely with existing scholarship emphasizing polarized public responses toward digital platforms associated with privacy, encryption, and user autonomy. Such platforms frequently engender divided public opinion, often shaped significantly by users' ideological leanings, platform loyalty, or inherent skepticism towards traditional media narratives (Marwick & Lewis, 2017).

The evident divergence between media framing and audience sentiment underscores a substantial sociotechnical dynamic that stakeholders—such as digital platforms, policymakers, business strategists, and media professionals—must carefully navigate. For Telegram, strong positive sentiment indicates a solid foundation of user support, potentially insulating the platform against reputational damage and regulatory scrutiny, thereby reinforcing its public position as a privacy-respecting communication medium. Policymakers, however, face pronounced challenges, as public resistance to negative framing and subsequent policy interventions illustrates the complex relationship between regulatory enforcement, public perception, and platform autonomy. Consequently, policymakers should approach regulation with greater sensitivity to user sentiment, perhaps engaging in proactive public dialogue to communicate clearly the rationale and objectives underlying platform governance decisions.

For business analysts and platform strategists, the sustained positive user sentiment signals a strategic opportunity to deepen user engagement, brand loyalty, and resilience to reputational crises. Incorporating comprehensive sentiment analysis into reputation management frameworks thus provides actionable insights, enabling proactive strategic responses aligned with genuine user attitudes. Meanwhile, the evident disconnect between journalistic framing and reader sentiment highlights a critical area for professional reflection among journalists and media organizations. Specifically, sentiment analysis offers a powerful feedback mechanism to gauge the impact of journalistic narratives on public opinion. Integrating sentiment analytics into journalistic practice can improve audience engagement and enhance credibility by encouraging more nuanced, responsive reporting practices, particularly on controversial technological issues.

Overall, this sentiment distribution analysis illustrates how sentiment modelling techniques—such as the lexicon-based VADER analyzer—can significantly contribute to understanding and interpreting complex public discourse around digital platforms. Such analyses not only inform methodological refinement in computational sentiment studies but also carry direct implications for the practical management of public perception, policy development, strategic communication, and journalistic integrity in the digital age.

Model Evaluation: TF-IDF + Logistic Regression

The evaluation of the Logistic Regression classifier using TF-IDF-transformed features revealed notable insights into the strengths and limitations of lexical sentiment analysis approaches in classifying public sentiment. The confusion matrix (Table 6 and Figure 2) demonstrates that the model achieved substantial predictive accuracy, correctly classifying 369 comments as positive (True Positives) and 189 as negative (True Negatives). However, it also exhibited an asymmetrical performance with a noteworthy positivity bias—misclassifying 67 negative comments as positive (False Positives), while only misclassifying 7 positive comments as negative (False Negatives). This asymmetry, while affirming the model's overall reliability, underscores a pronounced methodological limitation wherein nuanced expressions of negative sentiment are frequently missed or incorrectly interpreted.

The classification report (Table 9) further reinforces these findings through precision, recall, and F1-score metrics. The model exhibited high precision (96%) in negative sentiment classification, suggesting that when it did predict negativity, it was highly reliable. Nevertheless, the significantly lower recall for negative sentiment (74%) compared to positive sentiment recall (98%) signals a clear methodological challenge in capturing subtler forms of negativity. Such challenges align closely with established literature in sentiment analysis, which frequently identifies implicit negativity, irony, sarcasm, and contextually embedded critiques as complex linguistic constructs often inadequately captured by surface-level lexical analysis (Cambria *et al.*, 2017).

From a methodological standpoint, these results highlight limitations inherent in bag-of-words approaches such as TF-IDF, which rely primarily on explicit lexical cues rather than sophisticated semantic interpretation. Thus, the model's robust identification of positive sentiment likely reflects the clearer lexical and syntactic patterns commonly associated with explicit agreement or approval. In contrast, negative sentiment frequently involves linguistic subtlety and complexity, explaining the observed positivity bias. Future research directions could address this limitation by integrating context-sensitive approaches, such as transformer-based language models or neural embeddings, which can better interpret subtle linguistic signals and implicit semantic content.

Practically, the observed positivity skew has substantial implications for stakeholders. For Telegram, deploying similar sentiment models operationally—such as in content moderation or user-sentiment dashboards—may

inadvertently underestimate user dissatisfaction, obscuring critical feedback and inhibiting effective responsiveness. For business analysts and risk managers, positivity bias represents a potential analytical blind spot, potentially compromising the timely detection of reputational and compliance risks. Regulatory stakeholders should similarly remain cautious, recognizing that a sentiment model biased toward positivity could diminish the visibility of user concerns or critical perspectives essential for informed policymaking and oversight. Finally, computational linguists and NLP researchers are encouraged to address these biases and improve sentiment models by developing nuanced methods capable of recognizing linguistically complex negative expressions, ultimately ensuring more balanced and accurate sentiment assessments in public discourse.

In conclusion, this model evaluation indicates strong overall reliability for lexical sentiment analysis using TF-IDF vectorization and Logistic Regression. However, the asymmetrical performance pattern emphasizes the need for methodological enhancements to better capture nuanced negativity. Future sentiment analysis models should therefore adopt advanced computational methods, ensuring comprehensive and contextually accurate interpretation of public sentiment, particularly within emotionally charged or contentious digital communication contexts.

Comparison of Actual vs Predicted Sentiment Counts

The comparison between actual and predicted sentiment counts from the logistic regression classifier trained on TF-IDF-transformed features provides critical insights into both the methodological strengths and limitations of the approach. The classifier exhibited a notable bias toward positive sentiment, systematically predicting a higher number of positive comments (440) compared to the actual positive count (376), while correspondingly underestimating negative sentiment (192 predicted versus 256 actual). This positivity skew aligns closely with earlier classification metrics, particularly the recall scores that highlighted significantly higher recall for positive sentiment (0.98) compared to negative sentiment (0.74). These observations suggest that positive sentiment expressions typically feature more explicit, straightforward linguistic patterns, which are easier for lexical-based TF-IDF models to detect consistently. Conversely, the underrepresentation of negative sentiment indicates a substantial methodological shortcoming in capturing nuanced linguistic cues, subtle criticism, irony, and implicit forms of negativity.

From a methodological perspective, the identified bias underscores significant limitations inherent in traditional bag-of-words approaches like TF-IDF, which rely predominantly on surface-level lexical features and frequency counts. While effective in capturing explicit sentiment expressions, these approaches often fail to adequately interpret the context-dependent subtleties and complexity inherent in negative user-generated comments (Cambria *et al.*, 2017; Liu, 2015). Therefore, enhancing model accuracy—particularly for negative sentiment—would require integration of advanced contextualization methods such as neural embeddings, transformer-based architectures, or deep learning techniques that offer deeper semantic understanding of linguistic nuances.

Practically, the identified positivity bias bears substantial implications for operational sentiment monitoring in real-world digital platforms. For Telegram, the systematic underestimation of negative sentiment could obscure critical user feedback, grievances, or dissatisfaction, potentially compromising timely intervention, user satisfaction, and trust. Similarly, business analysts and risk managers relying on such sentiment analysis tools may underestimate reputational and compliance risks, misinforming strategic decisions and hindering proactive risk mitigation. Regulatory stakeholders may also be adversely affected, as this positivity bias could diminish the accuracy and effectiveness of regulatory assessments, particularly around sensitive or contentious platform practices. Consequently, these stakeholders must approach sentiment analysis results cautiously, supplementing lexical models with qualitative or additional analytical methods to achieve comprehensive sentiment understanding.

Furthermore, computational linguists and natural language processing researchers should interpret these findings as a clear mandate for continued innovation. This bias highlights the necessity for more nuanced, contextually sensitive modeling approaches capable of accurately interpreting the implicit, subtle, and often culturally specific cues of negative sentiment expressions. Advancing analytical techniques beyond traditional lexical frameworks—towards sophisticated, contextually aware computational linguistics—will enable more accurate, balanced, and insightful sentiment analysis outcomes, particularly within emotionally charged digital discourse contexts.

In conclusion, while the logistic regression classifier demonstrates overall robust predictive reliability, its positive sentiment skew identifies a significant methodological and practical limitation. Addressing this limitation through the adoption of advanced semantic modeling approaches promises more balanced, accurate, and practically valuable sentiment analyses for digital communication platforms and stakeholders.

Evaluation of Model Predictions

The detailed evaluation of the Logistic Regression model utilizing TF-IDF features, compared against original sentiment labels generated by the VADER analyzer, reveals key methodological strengths and critical limitations. The analysis categorizes model predictions into four distinct outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Notably, the model demonstrated strong overall predictive accuracy, correctly identifying positive sentiment in 375 instances and negative sentiment in 191 instances. This robustness in correctly classifying sentiment underscores

the general effectiveness of TF-IDF–based logistic regression approaches for explicit sentiment recognition. However, the presence of 65 False Positive predictions—negative comments incorrectly classified as positive—highlights a significant positivity bias within the model's classification approach. Conversely, the occurrence of False Negatives was minimal, with only one such instance observed, reflecting minimal risk in overlooking positive user-generated feedback.

An examination of specific misclassified examples further illuminates inherent methodological challenges. Comments like "Company's fault someone stupid downloaded personal data," a negatively intended critique incorrectly predicted as positive, exemplify the difficulty lexically oriented sentiment models encounter in accurately interpreting implicit negativity, subtle criticism, or nuanced linguistic context. Similarly, brief and ambiguous comments like "'s really" further emphasize these limitations, where brevity and lack of explicit lexical cues impede accurate classification. These instances underscore a critical shortcoming in lexically driven models such as TF-IDF, demonstrating their inadequate handling of subtle, contextually embedded negativity, ambiguity, and implicit expressions prevalent in authentic digital communication.

The positivity bias observed has substantial methodological implications, calling attention to the inherent constraints of frequency-based lexical approaches when interpreting user-generated content. Such limitations advocate strongly for the integration of more advanced computational methods—particularly deep learning models, context-aware neural embeddings, or transformer-based architectures—that possess greater semantic understanding and can better account for linguistic subtleties, irony, brevity, and implicit expressions of negative sentiment.

Practically, the implications of the model's positivity bias are significant for real-world sentiment monitoring. Telegram, as a platform provider, risks systematically underestimating genuine user dissatisfaction or critical feedback if similarly biased models are operationalized. This could compromise accurate detection and timely management of user concerns, potentially affecting user satisfaction and retention adversely. Business analysts and risk managers could likewise misinterpret sentiment data due to this positivity skew, possibly overlooking subtle but critical reputational or compliance-related risks. Regulatory stakeholders relying on such models might similarly underestimate negative sentiment, reducing their effectiveness in policy oversight and intervention. For computational linguists and natural language processing researchers, these findings reinforce the necessity of developing models capable of nuanced semantic interpretation, motivating research into sophisticated, contextually sensitive linguistic methodologies.

Overall, these insights emphasize a clear need to complement lexically driven sentiment analyses with advanced semantic modeling techniques and qualitative interpretation strategies. Ensuring comprehensive and balanced detection of nuanced negative sentiment expressions will significantly enhance methodological robustness, practical utility, and stakeholder trust in computational sentiment analysis tools—particularly within sensitive, nuanced digital communication contexts.

Word Frequency Patterns by Sentiment Category

The qualitative exploration of lexical patterns through word clouds (Figures 5 and 6) offers meaningful contextual insight, complementing the quantitative sentiment classification results. The visualization distinctly highlights differences in language usage between positive and negative comments, providing deeper understanding into user sentiment. Specifically, the positive sentiment word cloud (Figure 5) prominently features words such as "funny," "agree," "think," and "good," indicating that readers positively disposed toward Telegram frequently employed language associated with humor, agreement, reflective discourse, and explicit positivity. These lexical choices likely signify broader support, either for Telegram's foundational principles (privacy and autonomy) or for regulatory approaches towards the platform, notwithstanding its controversial associations. Such qualitative insights reinforce previous quantitative findings of prevalent positivity, underscoring user loyalty or skepticism toward media portrayals of Telegram's negative associations.

Conversely, the word cloud representing negative comments (Figure 6) emphasizes emotionally charged and security-focused terms, such as "criminals," "data," "identity," "illegal," and "scared." The prominence of these words underscores explicit user concerns about Telegram's potential role in facilitating illicit activities, including data breaches, identity theft, and digital exploitation. These lexically explicit negative expressions indicate clear apprehension and disapproval, highlighting a substantial segment of public sentiment focused on security, privacy violations, and digital threats associated with the platform.

Methodologically, these visualizations highlight the efficacy of qualitative text analysis techniques such as word clouds, which significantly enrich purely quantitative analytical approaches. By revealing the nuanced, underlying emotional and contextual dimensions of sentiment, these methods provide essential supplementary context to quantitative models, thus enhancing the overall interpretative accuracy of sentiment analyses. Practically, the stark distinction in lexical patterns emphasizes critical areas for platform management—such as Telegram—to proactively address user concerns, especially in security, privacy, and trust. Failing to address these highlighted negative concerns could compromise user satisfaction, reputation, and long-term platform sustainability.

From a risk management perspective, explicitly negative terminology ("illegal," "identity," "scared") suggests significant reputational or compliance risks, which analysts and strategists must rigorously monitor. Such

qualitative insights reinforce the necessity of integrated, contextually nuanced risk assessment frameworks, enabling more precise management responses to public concerns or controversies. For regulatory stakeholders, the frequency of crime-associated terms in negative user comments underscores urgent demands for accountability and effective oversight. Thus, policymakers are advised to closely align regulatory strategies with these publicly articulated user apprehensions, addressing platform accountability and protecting user rights effectively.

For computational linguists and natural language processing (NLP) researchers, the pronounced lexical distinction between positive and negative comments further emphasizes the methodological importance of advanced sentiment analysis techniques. Models sensitive to emotional vocabulary, domain-specific language, and subtle linguistic nuance will enhance precision and interpretative accuracy. Researchers should thus prioritize developing sophisticated computational methodologies, particularly contextually aware semantic models, to accurately capture the complexities of nuanced sentiment expressions.

In conclusion, qualitative lexical analyses, as illustrated by word clouds, significantly enhance quantitative sentiment analyses by illuminating underlying emotional and thematic user sentiments. Integrating qualitative techniques with advanced quantitative methods provides critical insights for methodological refinement, strategic reputation management, informed regulatory oversight, and ongoing computational linguistic research—particularly within contentious or socially sensitive digital discourse contexts.

Model Performance Metrics

The performance metrics summarized in Table 14 and Figure 7 indicate robust effectiveness of the logistic regression classifier using TF-IDF vectorization for sentiment analysis tasks. The model achieved an overall accuracy of approximately 89.56%, precision of 91%, recall of 90%, and an F1-score of 89%, collectively demonstrating its reliable predictive capability for accurately categorizing sentiment in reader-generated comments. High precision indicates substantial confidence in the model's sentiment predictions, reflecting accuracy in distinguishing between positive and negative commentary. Similarly, strong recall underscores its effectiveness in correctly identifying most sentiment classifications present within the dataset. The balanced F1-score further confirms the model's capability to effectively integrate both precision and recall into consistently reliable classification performance.

Despite these strong metrics, methodological limitations inherent in TF-IDF vectorization and logistic regression warrant critical examination. The comparatively lower recall for negative sentiment (74%, as detailed previously) implies that nuanced linguistic features—such as subtle criticism, implicit negativity, sarcasm, or contextually complex expressions—remain inadequately captured by purely lexical approaches. Consequently, the high accuracy achieved may still systematically underrepresent certain nuanced sentiment types, specifically negative or implicit expressions. Thus, future methodological refinement should focus on incorporating advanced modeling approaches, such as context-sensitive neural embeddings, deep learning models, or transformer-based architectures, capable of comprehensively interpreting linguistic subtleties and implicit sentiment cues.

From a practical standpoint, the demonstrated high reliability of this classifier supports its applicability for operational deployment in sentiment monitoring contexts, particularly those involving sensitive issues such as digital privacy or criminal associations. For Telegram's platform management, such reliable metrics indicate valuable potential for real-time sentiment monitoring, informing proactive user-experience interventions and platform responsiveness. However, recognizing inherent limitations in capturing subtle negativity, Telegram should employ supplementary qualitative analyses or human oversight mechanisms to ensure a comprehensive understanding of user sentiment nuances.

For business analysts and risk management professionals, these robust precision and recall metrics significantly enhance decision-making accuracy and strategic confidence, enabling more precise monitoring of user sentiment trends. Nevertheless, analysts must remain vigilant to inherent methodological limitations, complementing automated sentiment analyses with qualitative evaluations or expert judgment to avoid overlooking subtle reputational or compliance-related risks. Regulatory stakeholders similarly benefit from the model's demonstrated reliability, enabling informed oversight and nuanced policy formulation in response to public sentiment. However, regulators should be cautious of potential blind spots related to subtle negative sentiment expression, adopting supplementary analysis methods to ensure comprehensive understanding of public discourse.

For computational linguists and NLP researchers, these performance results highlight the efficacy of lexical models such as TF-IDF logistic regression, but simultaneously underscore their limitations. Addressing these methodological shortcomings necessitates advancing context-sensitive computational methodologies that better interpret subtle linguistic complexities, irony, implicit negativity, and nuanced discourse features prevalent in authentic digital communication.

Overall, while the logistic regression classifier demonstrates substantial practical reliability and predictive accuracy, stakeholders must remain mindful of inherent methodological limitations. Integrating complementary analytical techniques and developing advanced semantic modeling approaches remain critical avenues for improving comprehensive and nuanced sentiment analysis. Such advancements promise significant benefits for methodological rigor, operational decision-making, regulatory effectiveness, and computational linguistics

innovation, particularly in sensitive or controversial digital communication contexts.

CONCLUSIONS

This study systematically evaluated public sentiment regarding Telegram's reported association with criminal activities by applying sentiment analysis methods—specifically VADER and a TF-IDF vectorized Logistic Regression classifier—to reader comments sourced from a Wall Street Journal article. Results revealed notable patterns, both quantitative and qualitative, which carry critical methodological, theoretical, and practical implications. Quantitatively, VADER sentiment analysis classified the majority (59.5%) of reader comments as positive, indicating significant public skepticism towards the negative media framing of Telegram or a broader ideological support for privacy-centric platforms. The Logistic Regression classifier demonstrated robust performance (89.56% accuracy, precision of 91%, recall of 90%, and F1-score of 89%), affirming the utility of lexically-based sentiment classification models for effectively analyzing public discourse. However, it exhibited a clear bias toward positivity, reflected in lower recall for negative sentiment (74%) and a significant presence of false-positive classifications. These results highlight the critical methodological limitation of current lexical-based approaches in accurately capturing nuanced negative sentiment expressions, especially implicit negativity, subtle criticisms, sarcasm, and brevity. Consequently, advanced context-aware methodologies, such as transformer-based neural embeddings, are recommended for future research to enhance accuracy, particularly in detecting nuanced or implicitly negative sentiments.

Qualitative analysis, specifically word cloud visualizations, distinctly captured lexical patterns characterizing both positive and negative sentiment classes. Positive comments frequently included terms reflecting humor, agreement, or reflective discourse, suggesting supportive or skeptical attitudes toward negative reporting on Telegram. Conversely, negative comments explicitly conveyed heightened user concerns about Telegram's role in security breaches and digital criminal activities, using emotionally charged and crime-specific language (e.g., "criminals," "illegal," "scared"). These qualitative insights not only contextualize quantitative sentiment classifications but also underscore significant sociotechnical dynamics that digital platforms, policymakers, and regulatory stakeholders must acknowledge.

Practically, this research carries substantial implications for Telegram's platform management, business analysts, risk managers, regulatory authorities, computational linguists, and media professionals. The identified positivity bias in predictive models emphasizes caution in operationally deploying lexical-based sentiment analysis tools, which might systematically overlook subtle critical feedback, thereby limiting effective user engagement strategies, risk detection, and regulatory responses. Integrating qualitative analyses and advanced computational linguistic techniques alongside traditional lexical methodologies will thus enhance the comprehensive interpretation and responsiveness to public sentiment.

Ultimately, the nuanced examination provided by this study highlights sentiment analysis as an indispensable analytical framework, significantly informing the strategic management of corporate reputation, stakeholder communication, regulatory policy development, and the ethical accountability of digital platforms. Moving forward, continued methodological refinement and a balanced incorporation of qualitative insights will further enable sentiment analysis to robustly support nuanced decision-making in sociotechnical domains, notably digital communication and platform governance.

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 840–847.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A Practical Guide to Sentiment Analysis*. Springer.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*(6), 74–80.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15–21.

Castells, M. (2013). *Communication Power*. Oxford University Press.

Europol. (2022). *Internet Organised Crime Threat Assessment 2022*. Europol Public Information.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter, 12*(1), 49–57.

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science, 369*(6508), 1197–1201. https://doi.org/10.1126/science.abb2428

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society, 22*(6), 854-871.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social

media text. *Proceedings of the International AAAI Conference on Weblogs and Social Media, 8*(1), 216-225.

Kohlmann, E. (2024). Quoted in "How Telegram Became Criminals' Favorite Marketplace," *The Wall Street Journal.*

Kohlmann, E. (2024). Telegram and the evolving digital underground. *Journal of Cybersecurity Research, 8*(1), 12–27.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society Research Institute. https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pfeffer, J., Zorbach, T., & Carley, K. M. (2014). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications, 20*(1-2), 117-128.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Sexton, D. (2024). Quoted in "How Telegram Became Criminals' Favorite Marketplace," *The Wall Street Journal.*

Sexton, J. (2024). Encrypted messaging and criminality: Challenges and responses. *Digital Crime Studies Journal, 10*(2), 45-60.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems, 29*(4), 217-248.

Suzor, N. P. (2019). *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge University Press.

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2021). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science, 8*(9), 210631. https://doi.org/10.1098/rsos.210631

Team, B. (2025, March 20). *How many people use Telegram? 55 Telegram stats*. Backlinko. https://backlinko.com/telegram-users

The Wall Street Journal [WSJ]. (2024). *Telegram CEO Pavel Durov Arrested in France*. Retrieved from https://www.wsj.com/articles/telegram-ceo-arrested-france

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The Platform Society: Public Values in a Connective World*. Oxford University Press.

Wall Street Journal (2024). *How Telegram Became Criminals' Favorite Marketplace*. Retrieved from www.wsj.com.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research, 19,* 315–354.