Asian Journal of Probability and Statistics

Volume 26, Issue 6, Page 31-40, 2024; Article no.AJPAS.117163

ISSN: 2582-0230

Sentiment Classification of Safaricom PLC Social Media Sentiments on X(Formerly Twitter)

Joshua Kimani [a*], Anthony Karanjah [a] and Pius Kihara [b]

[a] *Department of Mathematics, Multimedia University of Kenya, Kenya.*

[b] *Department of Financial and Actuarial Mathematics, Technical University of Kenya, Kenya.*

Authors' contributions

This work was carried out in collaboration among all authors.All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajpas/2024/v26i6622

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/117163

Received: 15/03/2024

Accepted: 22/05/2024

Published: 31/05/2024

Original Research Article

Abstract

In today's digital era, social media plays a pivotal role in shaping public sentiment, particularly in the financial domain. This study focuses on sentiment analysis of social media discussions, specifically tweets discussing Safaricom PLC on X (formerly Twitter), leveraging Natural Language Processing (NLP) techniques.

**Corresponding author: E-mail: joshuakimani192@gmail.com;*

Cite as: Kimani, J., Karanjah, A., Kihara, P. (2024). Sentiment Classification of Safaricom PLC Social Media Sentiments on X(Formerly Twitter). Asian Journal of Probability and Statistics, 26(6), 31–40. https://doi.org/10.9734/ajpas/2024/v26i6622

By meticulously collecting, cleaning, and analyzing data, valuable insights into the sentiment landscape surrounding Safaricom PLC during a significant period were obtained. The sentiment analysis, conducted using the VADER lexicon, categorized sentiments into positive, negative, and neutral classes. Notably, the analysis revealed a predominant positive sentiment, indicative of an optimistic tone in discussions related to Safaricom PLC. This study highlights the potential of integrating sentiments and sentiment analysis techniques into stock price prediction models to facilitate informed investment decision-making

Keywords: Sentiment analysis; social media; Safaricom PLC; stock price prediction; natural language processing (NLP); VADER lexicon.

1 Introduction

In the rapidly evolving digital era, social media has become an invaluable barometer, dynamically reflecting the pulse of public sentiment and offering a compelling snapshot of collective opinions on unfolding events. The imperative process of sentiment analysis, which delves into the nuanced understanding of textual expressions, has garnered noteworthy attention, especially within the financial community, as illuminated by influential studies [1, 2] . This discernible trend within finance revolves around the strategic leverage of linguistic and text analysis tools, with a primary objective to automate sentiment assessment in news articles. This, in turn, introduces vital contextual information into existing financial models through the seamless integration of automated systems and sophisticated data processing techniques [3].

Expanding into the broader spectrum of natural language processing (NLP), text sentiment analysis has blossomed into a multifaceted discipline with applications reaching far and wide. Its pervasive influence extends across domains such as algorithmic trading, social media analytics, consumer behavior analysis, and strategic initiatives in sales and marketing [4]. At its core, this methodology hinges on the adept utilization of lexicons—predefined collections of terms intricately paired with sentiment ratings. The practical application of this methodology is further highlighted through the meticulous testing of various sentiment scoring algorithms on a comprehensive set of pre-labeled news items [5].

This research distinctly focuses on the intricate task of event sentiment classification — a methodological endeavor involving the identification and meticulous categorization of articles into distinct classes: positive, negative, or neutral, contingent upon the sentiments they convey. Our chosen methodology strategically incorporates the Vader lexicon library, seamlessly integrated into the Python Natural Language Toolkit (NLTK) to ensure a nuanced and accurate sentiment labeling process.

To further heighten the precision of our text classification, we deploy the snowball stemmer during the preprocessing stage. This sophisticated algorithm, developed by Martin Porter, stands as an advanced iteration of the original Porter Stemmer, effectively addressing and rectifying inherent issues [6]. The subsequent text cleaning process unfolds meticulously through a robust series of regular expression procedures. This comprehensive cleaning regimen includes the judicious removal of punctuation, numerals, website links, special characters, emojis, and extraneous white spaces.

In our earnest pursuit of social media sentiment classification, our chosen approach harmonizes seamlessly with the VADER (Valence Aware Dictionary and sEntiment Reasoning) framework [7]. Employing the Vader lexicon transformer, we harness the potency of this tool to methodically assign sentiments to articles within the defined categories of positive, negative, or neutral.

2 Related Works

Introducing SentiStrength, a rule-based sentiment analysis system, researchers emphasized its precision and hybrid approach, showcasing its efficacy in analyzing sentiment in social media through the integration of rule-based methods [8]. Subsequent research provided an in-depth exploration of various strategies for sentiment analysis on Twitter data, elucidating both the challenges faced and the methodologies employed in dissecting sentiment within the concise confines of Twitter (now X) [9].

Building upon these foundational studies, further research proposed innovative techniques for sentiment analysis in financial news articles, particularly those pertaining to stock market analysis. Their method involved the extraction of features from noun phrases, subsequently scoring them based on a negative/positive subjectivity measure. To augment accuracy, researchers implemented a feature selection process grounded in a contextual entropy model [10]. This model unearthed comparable emotion-laden words and their corresponding intensities from online stock market news events, thereby enriching the existing repository of seed words [11].

Continuing along this trajectory of innovation, another pivotal contribution to sentiment classification emerged with the introduction of the Vader lexicon, a transformative element within the natural language processing (NLP) toolkit. This indispensable tool played a pivotal role in advancing the field of sentiment analysis, heralding notable advancements [7].

Further enriching the landscape of sentiment analysis methodologies, subsequent research provided a comprehensive overview of deep learning techniques. Delving into their applications across various domains, including social media, this work served as an indispensable guide to the evolving methodologies in sentiment analysis [12]. In parallel, additional research explored techniques and methodologies pertinent to sentiment analysis of news articles. Through their rigorous examination, they shed light on the challenges, advancements, and practical applications, offering invaluable insights into the extraction of sentiments from textual news data [13].

However, a significant hurdle remains in effectively analyzing sentiments conveyed in tweets discussing Kenyan firms like Safaricom PLC on X (formerly Twitter). Despite the strides made in sentiment analysis, the concise and informal nature of tweets poses challenges in capturing the subtleties of sentiments expressed. This underscores the need for continued exploration and enhancement of sentiment analysis methodologies specifically tailored to the Kenyan context, with Safaricom PLC serving as a pertinent case study.

3 Methodology

The research methodology employed in this study is outlined below. We gathered data from articles about Safaricom PLC obtained from the social media platform X (formerly Twitter). For analyzing sentiment and determining sentiment classes, we utilized the Python programming language and its associated packages.

3.1 Research design

The research design consisted of three primary steps. Initially, we sourced and collected data from X (formerly Twitter). Following this, the collected data underwent a thorough cleaning process to ensure its quality and reliability. Subsequently, sentiments expressed in the collected articles were categorized into positive, negative, or neutral classifications.

The data collection process extended throughout the year 2022, which was particularly notable as an election year in Kenya. Data were gathered on a daily basis, excluding weekends and holidays. On average, we collected between 70 to 2000 articles or tweets per day. Sentiment classification was carried out on these tweets using Natural Language Processing (NLP) techniques. This comprehensive approach allowed for a detailed analysis of sentiments expressed on the platform during this significant period.

3.2 Text cleaning

In this study, the focus was on classifying social media articles (posts) into negative, positive, and neutral sentiments, constituting a three-class classification problem. Specifically, the research concentrated on articles sourced from the social media platform X, formerly known as Twitter. During this phase, data collection involved scraping for posts (articles) related to the organization of interest, Safaricom PLC, utilizing a web scraping code developed in Python.

Prior to text classification, it was imperative to undertake text cleaning procedures. This entailed several steps aimed at enhancing the quality and consistency of the data. Firstly, lemmatiza-tion was applied to reduce words in the dataset to their stems and roots. Subsequently, a regex method was employed to remove various elements such as punctuations, numericals, website links, special characters, emojis, and white spaces.

The algorithmic process for text cleaning involved the following steps:

1. Converting the text to lowercase to ensure uniformity.
2. Removing square bracket enclosed content to eliminate unnecessary elements.
3. Eliminating URLs (both http and www) to enhance readability.
4. Removing HTML tags to strip away any formatting artifacts.
5. Eliminating punctuation marks to focus solely on textual content.
6. Removing newline characters to maintain consistency in formatting.
7. Eliminating alphanumeric words to streamline the dataset.
8. Tokenizing the text using spaces to facilitate further processing.
9. Removing stop words to reduce noise in the dataset.
10. Employing stemming to reduce words to their root form, aiding in simplification.
11. Joining the cleaned words back into a cohesive string to prepare for subsequent analysis.
12. Returning the cleaned text, now deemed suitable for further processing.

Upon completion of these operations, the data was considered sufficiently "clean" and ready to undergo sentiment analysis using the Vader lexicon transformer. This meticulous cleaning process was essential to ensure the accuracy and reliability of the subsequent sentiment analysis results.

3.3 Sentiment analysis using VADER lexicon

To gauge the sentiment of the tweets, the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon transformer was employed. This lexicon-based approach assigns sentiment scores to each tweet, including positive, negative, and neutral sentiment probability scores, along with a compound sentiment score that encapsulates the overall sentiment of the tweet.

Valence scores, representing the emotional intensity or polarity of words, were utilized to assess the sentiment of individual words in tweets. The compound sentiment score is derived by aggregating these valence scores, considering both the polarity and intensity of the words.

The compound score (C) is calculated using the following formula:
Let S be the set of valence scores for all the words in a given text:

$S = s_1, s_2, \ldots, s_n$, then,

$$C = \frac{\sum_{i=1}^{n} s_i}{\sqrt{\sum_{i=1}^{n} {s_i}^2}} \tag{3.1}$$

where;

(i) C is the compound score.

(ii) S is the set of valence scores for all the words in the tweet.

(iii) n the number of words in the tweet.

(iv) s_i is the valence score of the i-th word.

This method provides a comprehensive assessment of sentiment, taking into account the nuanced emotional tones within the analyzed text.

4 Results and Discussion

4.1 Text cleaning

Text cleaning serves as a foundational preprocessing step vital for enhancing the accuracy and interpretability of sentiment analysis results. In the study, a meticulous approach to text cleaning was undertaken to ensure the integrity of the data used for sentiment analysis.

The text cleaning process encompassed several key steps aimed at eliminating noise and irrelevant elements from the dataset. Initially, special characters and symbols were systematically removed from the text corpus to streamline the data and facilitate subsequent analysis. Additionally, the removal of white spaces was performed to enhance data consistency and readability.

Furthermore, the handling of website links was prioritized to prevent their interference with sentiment analysis algorithms. By excluding website links from the text corpus, the focus remained solely on the textual content, thereby minimizing potential confounding factors.

Punctuation marks, known for their varied usage and potential impact on sentiment interpretat-ion, were also addressed during the cleaning process. Careful consideration was given to the appropriate handling of punctuation marks to ensure that their presence did not unduly influence the sentiment analysis outcomes.

By systematically removing extraneous elements and noise, the resulting dataset was optimized for accurate sentiment analysis.

Fig. 1. A word cloud of frequently used words

Fig. 1. illustrates the most frequently used words in our text data following the completion of the text cleaning procedure. The visualization provides valuable insight into the contents of the dataset, highlighting various terms and phrases associated with Safaricom. This analysis sheds light on the diverse conversations surrounding the organization, offering a comprehensive understanding of the topics discussed within the dataset.

4.2 Sentiment analysis using VADER lexicon

Table 1 presents the results obtained from the sentiment analysis phase, showcasing the sentiment probability scores of sample tweets. Each tweet is evaluated based on its compound, positive, negative, and neutral sentiment probabilities, allowing for a nuanced understanding of the sentiment expressed.

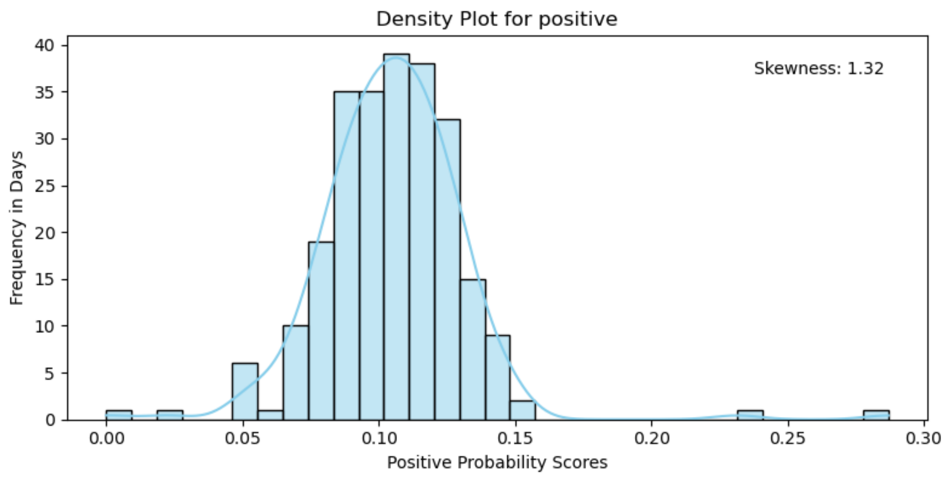
Table 1. Sample tweets sentiment probability scores

Date	Compound	Positive	Negative	Neutral	Class Type
30/12/2022	0.1027	0.186	0.109	0.705	POSITIVE
30/12/2022	0.3818	0.088	0.000	0.912	POSITIVE
30/12/2022	0.6705	0.524	0.000	0.476	POSITIVE
30/12/2022	-0.5719	0.000	0.190	0.810	NEGATIVE
30/12/2022	0.0000	0.000	0.000	1.000	NEUTRAL

The compound score, in particular, is valuable as it aids in categorizing tweets as overall positive, negative, or neutral based on a predefined threshold. The compound score values fall on a continuous scale of -0.959500 to 0.965700 which is really close to the continuous scale of -1 to 1. Hence, we adopt a common and intuitive approach, where, a score of approximately 0.0 serves as the threshold for neutrality, tweets with sentiment scores comfortably above the neutral threshold (i.e., > 0.1) are categorized as expressing positive sentiment. The higher the sentiment score, the stronger the positive sentiment conveyed by the tweet. This class captures tweets with content that is predominantly positive in tone and finally, tweets with sentiment scores clearly below the neutral threshold (i.e., < -0.1) are categorized as expressing negative sentiment. This is well illustrated in Table 1 which shows a sample of the transformed tweets' scores achieving the first objective of our study by classifying the social media sentiments into positive, and neutral.

4.3 Analysis of sentiment results

Fig. 2. displays the use of density plots to visualize the distribution of sentiments scores. This involved the positive, neutral and negative probability scores.



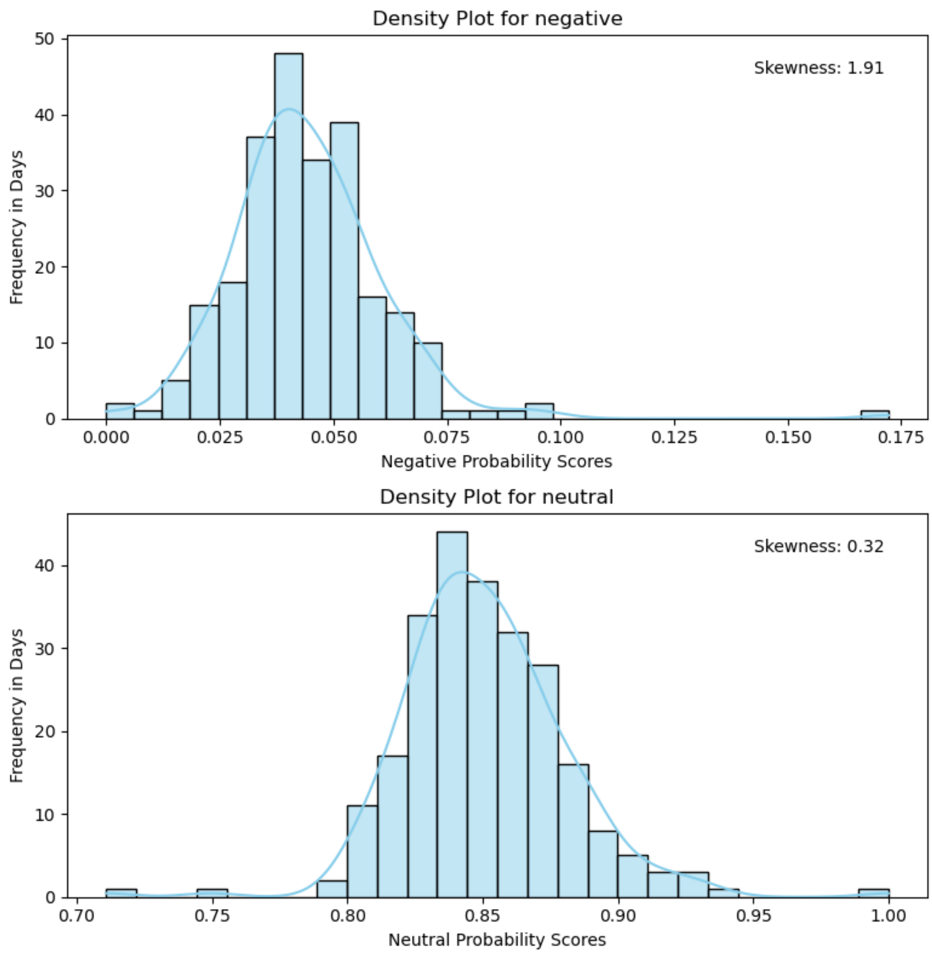


Fig. 2. Sentiment scores probability histograms

From Fig. 2, the positive sentiment probability score distribution exhibits a rightward skew with a skewness of 1.32. This indicates that the distribution is more heavily concentrated on lower positive sentiment probability scores as observed in the plot, with a smaller proportion of tweets having very high positive sentiment probability scores. The range of positive sentiment probability scores is 0.28, spanning from 0 to 0.28, further illustrates the spread of positive sentiment probability values within the dataset.

The distribution of negative sentiment probability scores demonstrates a more pronounced rightward skew, as evidenced by a skewness value of 1.91. This implies that the majority of tweets with negative sentiments exhibit lower negative probability scores, while instances of extremely negative sentiment probability scores are less frequent. The range of negative sentiment probability scores is 0.17, which spans from 0 to 0.17, further reinforces the concentration of negative sentiment values towards the lower end of the spectrum.

In contrast to the other sentiments, the distribution of neutral sentiment probability scores exhibits a slight rightward skew, as indicated by a skewness value of 0.32 almost tending to normal with a bell shaped as illustrated by the plot. This suggests that a majority of tweets have a higher likelihood of being classified as neutral. The range of neutral sentiment probability scores is 0.29, spanning from 0.71 to 1, further demonstrates the higher spread in neutral sentiment scores compared to positive and negative sentiments.

To gain valuable insights into the dataset, we performe a comprehensive analysis by calculating essential summary statistics. These statistics provide us with a clear understanding of the dataset's central tendencies and distributions. By examining measures such as mean, median, standard deviation, and quartiles, we uncovered key information about the dataset's overall characteristics.

Table 2. Summary statistics

	Compound	**Positive**	**Negative**	**Neutral**
count	245	245	245	245
mean	0.090667	0.104904	0.044273	0.850700
std	0.049817	0.026276	0.017116	0.030183
min	-0.305100	0.000000	0.000000	0.710940
25%	0.063408	0.090210	0.033820	0.832200
50%	0.091175	0.105260	0.042337	0.847500
75%	0.120720	0.119690	0.053140	0.867700
max	0.224739	0.286990	0.172167	1.000000

From Table 2, it is seen that the dataset consists of 245 records, indicating the number of observations analyzed in the sentiment analysis.The average sentiment scores for compound, positive, negative, and neutral sentiments are approximately 0.0907, 0.1049, 0.0443, and 0.8507, respectively. These values represent the central tendency of sentiment scores across the dataset. The standard deviation measures the dispersion of sentiment scores around the mean. For compound, positive, negative, and neutral sentiments, the standard deviation values are approximately 0.0498, 0.0263, 0.0171, and 0.0302, respectively.

The minimum sentiment scores represent the lowest observed values in the dataset. The lowest compound sentiment score is approximately -0.3051, while the lowest positive and negative sentiment scores are both 0.0000. The lowest neutral sentiment score is approximately 0.7109. The 25th percentile values indicate the sentiment scores below which 25% of the observations fall. These values are approximately 0.0634, 0.0902, 0.0338, and 0.8322 for compound, positive, negative, and neutral sentiments, respectively. The median sentiment scores represent the middle value of the dataset. The median values for compound, positive, negative, and neutral sentiments are approximately 0.0912, 0.1053, 0.0423, and 0.8475, respectively. The 75th percentile values indicate the sentiment scores below which 75% of the observations fall. These values are approximately 0.1207, 0.1197, 0.0531, and 0.8677 for compound, positive, negative, and neutral sentiments, respecti-vely. The maximum sentiment scores represent the highest observed values in the dataset. The highest compound sentiment score is approximately 0.2247, while the highest positive and negative sentiment scores are approximately 0.2870 and 0.1722, respectively. The highest neutral sentiment score is 1.0000.

5 Summary and Recommendations

The study embarked on sentiment classification of social media articles, particularly focusing on tweets discussing Safaricom PLC on X (formerly Twitter). Leveraging Natural Language Processing (NLP) techniques, specifically the VADER lexicon sentiment analyzer, we categorized sentiments into three classes: positive, negative, and neutral. Through meticulous data collection, cleaning, and sentiment analysis, we gained valuable insights into the sentiment landscape surrounding Safaricom PLC during a significant period, notably the 2022 election year in Kenya.

Our analysis revealed a predominant positive sentiment class, followed by neutral and negative sentiments. This finding underscores the optimistic tone prevalent in social media discussions concerning Safaricom PLC. Moreover, the comprehensive sentiment analysis provided by VADER facilitated a nuanced understanding of sentiment distributions, enabling us to discern trends and patterns within the dataset.

However, it's essential to acknowledge that capturing the nuances of sentiments expressed in tweets discussing Kenyan firms like Safaricom PLC on social media platforms presents unique challenges. The concise and informal nature of tweets, coupled with cultural and contextual factors, requires careful adaptation and enhancement of sentiment analysis methodologies.

Moreover, integrating sentiments into stock price prediction models introduces additional challe-nges. These challenges include ensuring the quality and reliability of social media data, addressing the complexity of model integration, and accurately interpreting sentiment analysis results in the context of stock market dynamics. Despite these challenges, the potential benefits of incorporating sentiment analysis outweigh the obstacles, providing investors with valuable tools for informed decision-making.

To address this need, future research should explore innovative approaches to sentiment analysis tailored to the Kenyan context. This may involve the development of specialized lexicons or sentiment analysis models trained on Kenyan-specific linguistic data. Additionally, incorporating domain-specific knowledge and cultural insights into sentiment analysis algorithms can further enhance their effectiveness in capturing the intricacies of sentiments expressed in Kenyan social media discussions.

By advancing sentiment analysis methodologies to better suit the nuances of Kenyan social media discourse, researchers and practitioners can unlock new opportunities for understanding public sentiment, informing strategic decision-making, and driving actionable insights in the context of Kenyan firms and industries.

This study recommends the integration of sentiments and sentiment analysis techniques into stock price prediction models for Safaricom PLC, a telecommunications company. By incorporating sentiments extracted from social media discussions, investors can gain valuable insights into market sentiment dynamics and make more informed investment decisions. Integrating sentiments into stock price prediction models offers several benefits, including improved prediction accuracy, enhanc-ed risk management, and opportunity identification.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Tetlock PC. Giving content to investor sentiment: The role of media in the stock market. J. Financ. 2007;62:1139–1168.

[2] Xing FZ, Cambria E, Welsch RE. Natural language based financial forecasting: A survey. Artif. Intell. Rev. 2017;50:49–73.

[3] Li X, Xie H, Chen L, Wang J, Deng X. News impact on stock price return via sentiment analysis. Knowl. Based Syst. 2014;69:14–23.

[4] Aryaman. Natural language processing and sentiment analysis. International Research Journal of Computer Science. 2021 – October; X:237–242. J Diff. Equ. Appl. 2005;11(7):605-618.

[5] Shapiro AH, Sudhof M. Wilson DJ. Measuring news sentiment. J.Econom; 2020.

[6] Porter, M. “An algorithm for suffix stripping. Program. 1980;14(3):130-137.

[7] Hutto CJ, Gilbert EE. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI; June.2014;34(11):1585-1598.

[8] Thelwall M, Buckley K, Paltoglou G, Cai D. sentistrength: A high-precision, rule-based, and hybrid system for sentiment analysis. IEEE Transactions on Affective Computing; 2010.

[9] Pak F, Paroubek P. Mining twitter data for sentiment analysis: An Overview of Current Approaches. LREC; 2010

[10] Schumaker RP, Zhang Y, Huang C-N, Chen H. Evaluating sentiment in financial news articles. Decis. Support Syst. 2012;53:458–464

[11] Yu L-C, Wu J-L, Chang P-C, Chu H-S. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl. Based Syst. 2013;41:89–97.

[12] Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery; 2018.

[13] Gupta A, Kumar S, Varma R. Sentiment analysis in news articles: A comprehensive review. Journal of Information Science; 2018.

© Copyright (2024): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/117163