JURNAL INFOTEL
VOL. 17, NO. 3, AUGUST 2025, PP. 627–648. DOI:10.20895/INFOTEL.V17I3.1289

RESEARCH ARTICLE

Data Mining's Impact on Company Performance Using Consumer Reviews on Social Media: A Case Study of the Telecommunication Industry

Dwi Adi Purnama[1,*]

[1]Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia

*Corresponding email: dwiadipurnama@uii.ac.id

Received: December 16, 2024; Revised: June 04, 2025; Accepted: June 30, 2025.

Abstract: The industry is currently faced with rapid technological developments, including the challenges of Industry 5.0. Advanced technology is necessary to improve automation and digitalization in the industrial sector. One of them involves mining information from social media data, which generates large amounts of data (big data) and offers the potential to improve company performance. This study takes a case study of the telecommunications industry in Indonesia, using Principal Component Analysis (PCA) and Principal Component Regression (PCR) methods. Big data is obtained from social media review data collected over 33 weeks, encompassing unstructured data on telecommunications service products in Indonesia. The text mining stage produces 30 selected words for further analysis with PCA to create the main components. Based on the evaluation results, the main components formed show a good correlation with the company's performance in the stock market; at least, there is one main component equation that shows a strong correlation. Principal component 2 demonstrates the optimal and most suitable R-squared value to predict the stock price index variable. PC2 comprises the textual elements "smartfrencare", "service", "obstacle", "modem", "location"', "voucher", "Wi-Fi", "promotion", and "interference", exhibiting the highest R-value exceeding 0.5 on the PCR stock index-open, index-high, index-low, and index-close, demonstrating strong or moderate correlation strength. This shows the potential to use a data mining approach based on social media reviews as a basis for decision making to improve company performance. Furthermore, the dominant variables formed from the PCA are considered to obtain a simple mathematical model.

Keywords: Company Performance, Data Mining, Principal Component Analysis, Social Media, Telecommunication Industry

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

The current technological era is entering the fifth industrial phase, marked by rapid technological developments that are leading to automation and digitalization in the industrial sector. The development of technology and digitalization enables easy access to data from various social media platforms. There will be 2.95 billion people in the world using social media in 2022, and there will be more than 1.3 billion social users using Twitter [1]. Various Twitter user information is very useful to explore, especially for business purposes. Customer information on social media is essential to explore for business purposes. In product development, even the best ideas come from customers [2]. Furthermore, Twitter data exploitation can be used for various prediction purposes [3–6]. The increasing data capacity provides more information that is easy to access and lower-cost. Alternative uses of Twitter social media data need to be further analyzed to improve company performance.

The amount of data generated by social media users is growing exponentially, and big data is being produced, which creates new opportunities for companies to gain valuable insights into consumer preferences and product perceptions. One increasingly popular method for harnessing this data is data mining, which allows companies to identify patterns and trends in big textual data [7–9]. Text mining and sentiment analysis, as a branch of data mining, have become essential tools for companies to measure consumer sentiment towards their brands, products, or services [10, 11]. By analyzing the text of consumer reviews, companies can identify positive, negative, or neutral opinions, as well as the most frequently discussed topics. So, big social media data can be used to improve company performance by providing cheap and quick access to big data.

In terms of business, one crucial factor that attracts the attention of different business people and academics is related to the company's performance in the stock market. The involvement of businesspeople in the stock market can provide benefits to various parties if they can read stock trends well. Still, it will be detrimental if they cannot accurately predict. As with a product business, where there are far more failed products than successful products in the market, so too in the stock market business. Stock indexes, derived from stock price values and volumes, can reflect company performance indicators. However, estimating stock prices and volumes as indicators of company performance is a significant challenge due to dynamic trend changes. The difficulties found from previous studies conducted using historical data approaches show that the stock market pattern is very volatile. Stock price trend patterns are nonlinear and non-stationary time series data, making it difficult to predict stock price pattern estimates. However, if you can analyze the right analysis pattern related to stock prices, it can generate significant profits and contribute to designing profitable trading strategies.

Previous research has shown a significant relationship between data mining and various company performance metrics, such as sales, market share, and stock prices. Text mining and sentiment analysis were previously used for sales forecasting [12–14], market share [15,16], stock prices [17,18]. Furthermore, the Latent Dirichlet Allocation (LDA) technique, which is similarly designed to model consumer review topics, organizes subjects by synthesizing a collection of keywords [19]. The Social Network Analysis (SNA) method enables the identification of subjects and the investigation of relationships between topics or keywords, facilitating the mapping and recognition of consumer review trends on social media [20]. Sentiment analysis can be utilized to determine the polarity and emotions of clients, classifying them as positive, neutral, or negative [20]. Although previous studies

have explored data mining for company performance, none have utilized big data from social media data based on customer reviews to predict and improve company performance. Then, this study uses the Principal Component Analysis (PCA) and Principal Component Regression (PCR) approaches to predict company performance. Thus, the novelty and contribution of this study lie in its use of big data from customer reviews to predict company performance, as well as the application of PCA and PCR as analysis techniques. To the best of our knowledge, utilizing the approach, concept, and big data mining from social media has yet to be done in previous studies.

This paper examines the role of data mining in utilizing abundant data (big data) from social media, such as Twitter, to analyze company performance. However, the problems associated with using big data from social media stem from the unstructured, informal, and noisy nature of the data. Moreover, the variables obtained from the data exhibit a high correlation among themselves. So, a formal procedure for analyzing social media data is required, and a multivariate statistical approach supports this. It is necessary to understand the potential of social media data for company performance and to develop a framework for large-scale unstructured data mining from social media, serving as a basis for policies that improve company performance and competitiveness.

2 Research Method

Information mining through social media offers the advantages of being cheap, fast, and comprehensive, making it an effective tool for finding important information from various customers about a product. Social media analysis provides an alternative approach with effective and real-time cost analysis based on customer opinions about the product [21–23]. Previous studies on online customer evaluations through comment review analysis have employed data mining and machine learning approaches, including text mining, sentiment analysis, topic modeling, principal component analysis, and social network analysis. Text mining is the most straightforward technique for analyzing the words and elements most frequently discussed by customers on social media [23]. The topic modeling method, shown by the Latent Dirichlet Allocation (LDA) technique, is similarly developed to model consumer review subjects; however, LDA organizes topics by synthesizing a set of keywords [20]. The Social Network Analysis (SNA) method facilitates the identification of subjects and the examination of relationships between topics or keywords, enabling the mapping and recognition of consumer review trends on social media [24]. Sentiment analysis can be employed to ascertain the polarity and feelings of clients, categorizing them as positive, neutral, or negative [25]. This study employs text mining techniques to identify significant terms and aspects, followed by principal component analysis to group and reduce key variables from consumer evaluations on social media. This work contributes by utilizing big data from customer evaluations to forecast corporate performance, employing Principal Component Analysis (PCA) and Principal Component Regression (PCR) as analytical techniques. The research Flowchart is explained in Figure 1.

2.1 Data

The type of data analyzed is unsupervised data obtained from Twitter comment reviews. Utilizing online data from social media provides several advantageous aspects for data

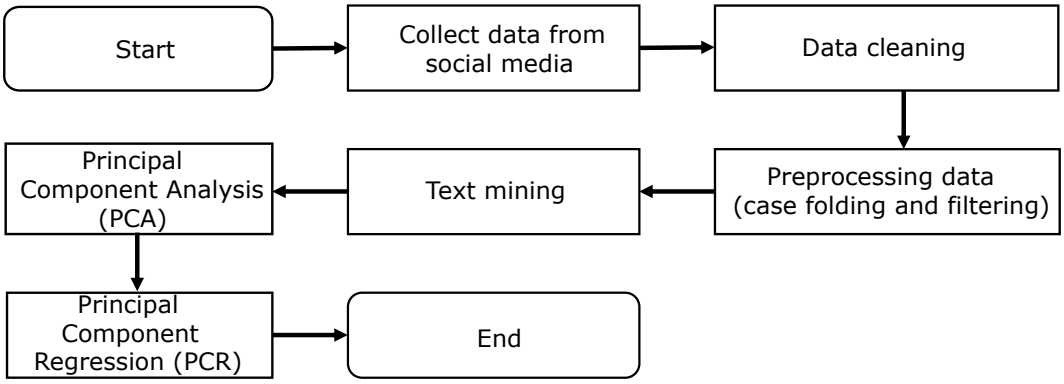


Figure 1: Research workflow.

collection, such as increased sample numbers, improved sample variety, heightened flexibility and ease, and diminished time and financial costs [26]. Twitter data (social media) from various user opinions related to 'smartfren' is obtained using Python software, with parameters set for data search time, keywords, and language. The keywords of this study focus on the object of research, Smartfren, using Twitter tweet data collected over 33 weeks from January to April 2019, as presented in Table 1. This study utilizes past data, specifically one year before COVID-19, to ensure an objective analysis of data mining's role in enhancing company performance, unaffected by numerous variables or uncertain situations. This study uses a case study of a telecommunications company, supported by the fact that data mining techniques can be used as a company improvement strategy [27, 28]. The utilisation of 33 periods is considered adequate for statistical analysis and data mining; additionally, research by D'Uggento et al. [29] has recognized text mining and data mining employing temporal data for 12 periods. Moreover, the utilization of 33 data periods is based on the Central Limit Theorem, which asserts that the sampling distribution of the sample mean will converge to a normal distribution, irrespective of the original population distribution's form, provided the sample size is sufficiently large [30]. The number 30 is the minimal threshold at which the Central Limit Theorem impact becomes relevant [30].

2.2 Data Pre-processing

Data pre-processing involves cleaning the data to prepare it for further technical processes. Data cleaning involves extracting only the most relevant information from the textual content of comments made by Twitter users discussing the target keyword "smartfren". This means that researchers remove information on usernames, account names, dates, hashtags, and other information that is not related to user comments. In pre-processing, researchers prepare a dictionary consisting of vocabulary words, as well as synonyms, to avoid the appearance of similar words that are repeated. The data to be processed is in the form of a comment column only; for example, it can be shown in the Twitter data excerpt in Table 2.

Data for each comment review is preprocessed using a text mining approach; the results of this preprocessing are then used as input for the process, which employs principal component analysis. Data preprocessing is carried out through several processes, namely tokenization, case folding, and filtering using stop word removal.

Table 1: The number of twitter data

Period	Number of reviews	Period	Number of reviews	The form of the data
1	1833	18	2560	textual
2	1995	19	2656	textual
3	2577	20	3325	textual
4	1803	21	3638	textual
5	1799	22	4617	textual
6	1812	23	3814	textual
7	1888	24	2802	textual
8	2282	25	3097	textual
9	2788	26	4335	textual
10	2830	27	6918	textual
11	2357	28	8922	textual
12	1634	29	8198	textual
13	2128	30	5242	textual
14	2451	31	4540	textual
15	2770	32	4465	textual
16	2357	33	5826	textual
17	2348			textual

Table 2: Twitter data excerpt

Yes, smartfren is good. For downloading it is also super fast
Wow, smartfren is still slow even though using mifi
Do those who use Smartfren have good signals?
You don't want to use smartfren
It's terrible, with disruptions everywhere,
but I'm now relieved after switching to GoUnlimited smartfren.

1. Tokenization is the initial step taken by separating the text available in each comment into pieces of words, often called tokens. The transformation steps in the tokenization process are shown in Table 3.
2. Case folding is standardizing the form of letters or words to lowercase. This is implemented because two-word characters with varying capital and lowercase letters can possess distinct meanings recognized by the computer system. Furthermore, it can also be executed to eliminate redundant space characters.
3. Stop word removal. The filtering step is done by removing stop words (the, is, ...), punctuation (@, #, ...), abbreviations (omg, ...), typos, and symbols.

2.3 Text Mining

Data mining starts with text mining, which involves finding keywords and calculating their frequency of occurrence. This allows for the identification of the types of words that frequently appear and are discussed on social media, particularly about opinions that are often reviewed. The results of the text mining are then used to select the types of relevant words related to the target object 'smartfen' based on the frequency of occurrence, so that

Table 3: The Tokenization Process

Yes, Smartfren is really good for downloading, it's also super fast	yes smartfren is really good for download is also super fast
Wow, Smartfren is still slow eventhough using MiFi	wow smartfren is still slow even though using MiFi

unnecessary noisy data can be removed. Text mining involves several processes, namely tokenization, case folding, and filtering. Tokenization is the initial step, which consists of separating the text available in each comment into pieces of words called tokens.

The theory of text mining is based on the idea that hidden value is frequently concealed in large amounts of textual data (such as emails, documents, social media posts, and customer reviews). It focuses on the process of converting unstructured text into a machine-readable format so that algorithms can analyze it [23]. Pre-processing (tokenization, normalization, stop-word removal, stemming/lemmatization), text representation (e.g., Bag-of-Words, TF-IDF, or word embeddings like Word2Vec), and the use of sophisticated analytical techniques (e.g., classification, clustering, topic modelling, sentiment analysis, and entity extraction) are some of the important steps involved in this process. To facilitate data-driven decision-making, improve contextual understanding, and uncover patterns that are challenging to identify manually in large and complex text data, the main objective is to extract both explicit and implicit knowledge. The key formulas commonly used in text mining, using Term Frequency (TF), are explained in Eq.(1). Term Frequency measures how often a word appears in a specific document.

$$\mathrm{TF(t,d)} = \text{Count of word } t \text{ in document } d \tag{1}$$

2.4 Principal Component Analysis (PCA)

Variable analysis using Principal Component Analysis (PCA) is performed as a multivariate analysis, based on the value of the covariance matrix to obtain eigenvalues. Consideration for using PCA is that the variables under investigation are often closely connected, efficiently conveying the same information. To analyze the relationships among a collection of p-correlated variables, it may be advantageous to convert the original variables

into a new set of uncorrelated variables known as principal components [31]. Principal component analysis can also be regarded as eliminating multicolinearity in the data. Furthermore, the eigenvector value explains the correlation between the original variable and the new variable (principal component) formed by PCA. The eigenvector value calculates the maximum number of components obtained, namely, thirty components.

A multivariate, unsupervised statistical method essential to data analysis and machine learning, principal component analysis (PCA) aims to minimize the variance in a data set while reducing its dimensionality [32]. The idea of orthogonal linear transformations, which transform a collection of possibly correlated variables into a new set of uncorrelated variable values known as principal components, is the foundation of PCA theory. The first principal component explains the largest variance, followed by the second, which is orthogonal to the first, and so on. Each principal component is a linear combination of the original variables and is arranged according to the amount of variance it explains. PCA enables researchers to simplify data complexity, reduce noise, address multicolinearity issues, and visualize hidden structures in high-dimensional datasets by identifying the most significant components (typically those with the highest eigenvalues, as seen in a Scree Plot). This improves predictive model performance and facilitates interpretation.

2.5 Principal Component Regression

Regression and correlation analysis of the PCA analysis components results were conducted to identify the relationship between various variables obtained from Twitter social media data and company performance in the stock market. This was achieved by connecting several word characters from text mining based on the frequency of word occurrence and stock price and volume data in each period. The Principal Component Regression enables the precise identification of significant elements, the exclusion of irrelevant ones, and the assessment of their interrelationships [31]. Regression analysis is frequently employed to model or examine data. It is used to comprehend the correlation between the variables, which may subsequently be leveraged to forecast the exact result [31].

Key Formulas in Principal Component Regression (PCR)
PCR combines Principal Component Analysis (PCA) with Multiple Linear Regression. This is the first and primary step in PCR. The independent variables (predictors) X undergo a PCA transformation.

1. Standardization of Predictor Data As in pure PCA, the predictor matrix X must be standardized.
2. Calculation of the Predictor Covariance/Correlation Matrix.
3. Eigen-decomposition of the Predictor Covariance Matrix.
   Find the eigenvalues (λ) and eigenvectors (v) of CX:

$$CXv = \lambda v \tag{2}$$

   Order the eigenvalues from largest to smallest: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The corresponding eigenvectors, $v_1, v_2, \ldots, v_p$, are the Principal Components (PC).
4. Selection of Principal Components Select k top principal components (based on the largest eigenvalues or other criteria such as a Scree Plot, cumulative explained vari-

ance, or, as you noted in your paper, the highest correlation with the dependent variable). Let V_k be the matrix whose columns are the chosen k eigenvectors (a $p \times k$ matrix).

5. Calculation of Principal Component Scores Project the standardized predictor data onto the k-dimensional space formed by the selected principal components. This yields the principal component scores matrix T (an $n \times k$ matrix). T=XstdVk Each column of T represents a Principal Component score ($PC_1, PC_2, \ldots, PC_k$).

3 Results

3.1 Text Mining Results

Text mining results provide information on the variables that are frequently reviewed in Twitter comments concerning Smartfren's performance. The number of words that appear based on frequency can be associated with the character of Smartfren's performance at that time in various periods, allowing for an analysis of the product's character. An example of text mining results is shown in Figure 2.

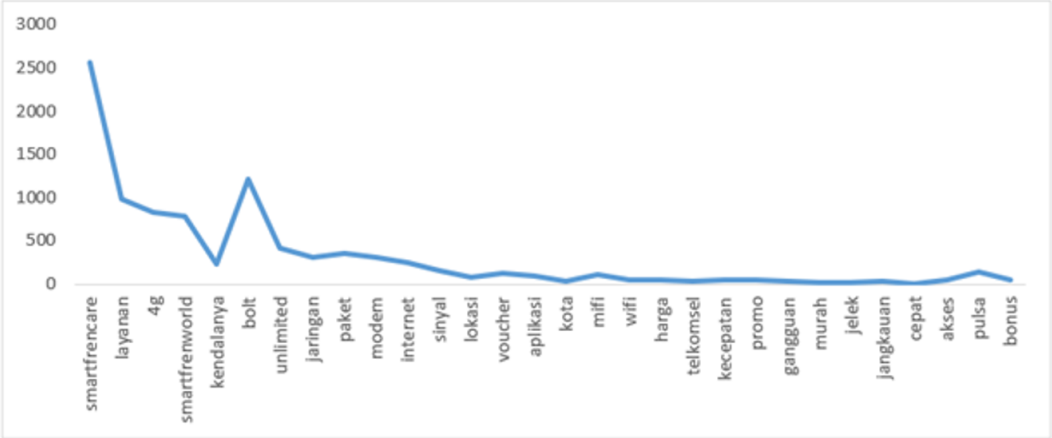


Figure 2: The example of text mining in the first period.

Figure 2 displays a line graph showing the frequency of various keywords extracted from Twitter comments related to telecommunication. It appears to be the result of a text mining process, likely indicating the prevalence of certain terms in user reviews or discussions. Many of the high-frequency terms are related to specific telecommunication providers, general service aspects, or common problems. The highest frequency of customer reviews is "smartfrencare", "layanan", and "4g".

1. "smartfrencare" is by far the most frequent term. "Smartfren Care" likely refers to Smartfren's customer care or service channel. Its high frequency suggests that users are actively engaging with or mentioning Smartfren's customer support. This could indicate a significant volume of inquiries, complaints, or feedback directed at their service.

2. "layanan" is an Indonesian word that means "service". Its high frequency indicates that the general quality or type of service provided by telecommunication companies is a major topic of discussion on Twitter. This is a broad term, so further drilling down into its context would be beneficial (e.g., "layanan cepat", "layanan buruk", "layanan pelanggan").
3. "4g" refers to the mobile network technology. Its high frequency suggests that network speed, availability, or performance (specifically for 4G) is a significant concern or point of discussion among users.

3.2 Principal Component Analysis (PCA)

Based on the results of text mining, 31 variables are often reviewed in Twitter social media comments from the keyword smartfren, and some of them have a high correlation with the smartfren stock price index. The number of variables involved in predicting the stock index can provide the complexity of the analysis and require more time. Thus, it is necessary to simplify the variables involved based on their importance to the stock price index to predict the dynamics of the rise and fall of stock values in each period based on tweet data on Twitter social media. The PCA results provide the weight of the variance of each variable involved. Then, 7 Principal Components are selected based on Figure 3, explaining that using seven components can explain the total variance of more than 80%.

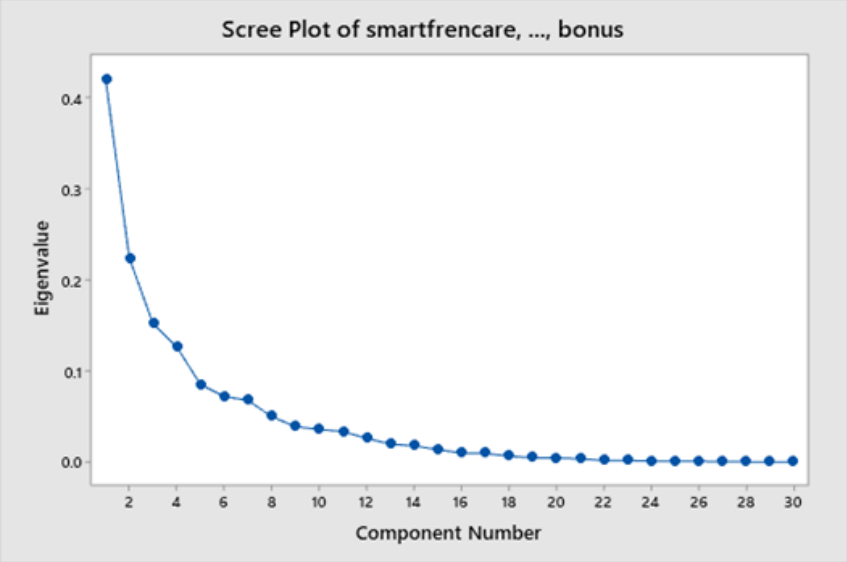


Figure 3: Eigen value.

Figure 3 displays a Scree Plot, which is a visualization tool used in Principal Component Analysis (PCA). It helps determine the optimal number of principal components to retain in an analysis. X-axis (Component Number) represents the number of principal components. Each point corresponds to a specific component (e.g., Component 1, Component 2, etc.). The plot goes up to 30 components, suggesting that there were 30 initial variables (likely the keywords from your previous text mining output, "smartfrencare" to "bonus").

Then, the Y-axis (Eigenvalue) represents the eigenvalue associated with each principal component. Eigenvalues quantify the amount of variance explained by each component. A larger eigenvalue means the component explains more variance in the original data. Based on the involvement of the seven main components described in Figure 3, the results converge to represent the entire variant formed.

3.2.1 Eigen Vector

The eigenvector value explains the correlation between the original variable and the new variable (principal component) formed by PCA. The eigenvector value calculates the maximum number of elements that are the same as the number of variables, namely 31 components. Based on the previous scree plot analysis, seven main components can explain the principal component value.

Table 4: Eigen vector principal component analysis

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
smartfren	0.179	0.303	0.095	-0.132	0.066	-0.18	-0.05
Service	-0.147	-0.474	-0.181	0.059	0.027	-0.011	-0.057
4g	0.311	-0.119	-0.208	-0.192	0.205	0.119	-0.02
smartfrenworld	0.292	-0.087	0.129	0.065	-0.183	0.033	-0.039
obstacle	0.145	0.264	0.132	-0.024	0.002	-0.074	-0.07
bolt	-0.112	-0.152	-0.124	0.128	0.011	-0.179	-0.034
unlimited	-0.276	-0.083	-0.334	0.171	0.07	0.028	0.021
network	0.216	0.095	0.104	-0.092	-0.03	0.038	-0.034
package	0.171	0.054	0.228	-0.087	0.053	-0.021	-0.033
modem	-0.072	0.329	-0.225	0.105	-0.095	-0.098	0.073
internet	0.247	-0.057	-0.009	-0.153	0.317	-0.05	-0.19
signal	0.213	0.007	-0.087	0.128	-0.312	-0.095	0.143
location	0.23	0.053	0.127	-0.109	-0.189	-0.009	0.049
voucher	0.157	-0.325	0.168	0.106	0.273	0.079	0.343
application	-0.07	-0.137	0.185	0.225	-0.122	0.01	-0.523
city	-0.031	-0.006	0.058	-0.064	-0.041	0.309	-0.12
wifi	0.1	-0.034	-0.116	0.032	-0.163	-0.144	0.049
wifi	-0.011	-0.018	-0.156	-0.034	-0.078	-0.087	-0.068
price	0.087	-0.078	0.112	0.445	0.234	0.067	0.141
telkomsel	0.104	-0.15	-0.077	-0.044	-0.168	0.064	0.086
speed	-0.042	0.035	-0.157	-0.064	-0.121	-0.068	-0.014
promotion	0.125	-0.232	-0.295	-0.087	0.099	-0.213	0.109
credit	0.223	-0.206	0.041	-0.129	-0.086	-0.056	0.11
cheap	0.147	-0.088	-0.133	-0.147	-0.387	-0.047	0.151
bad	-0.236	-0.115	-0.165	0.167	-0.054	-0.121	-0.097
reach	0.112	-0.142	0.092	-0.167	-0.365	0.057	-0.271
fast	0.221	-0.172	0.109	-0.213	-0.265	-0.221	-0.021
access	0.257	0.105	0.097	0.002	-0.131	-0.031	0.237
pulsa	0.086	-0.095	-0.137	-0.097	-0.068	-0.138	-0.183
bonus	0.053	-0.082	0.134	0.141	0.245	-0.273	-0.233

Dominant Variables from Principal Component Analysis

Principal Component 1

The components in PC 1 are formed from 33 text-mining variables that have been obtained previously. These variables can be simplified based on the results of vector loading in Table 3, which considers a few variables, but the dominant variables can compose the components. The dominant variables that form PC1 include 'smartfrenworld', '4g', 'unlimited', 'network', 'internet', 'signal', 'telkomsel', 'cheap', 'bad', 'interference', 'fast', 'access', and 'credit'.

Principal Component 2

The dominant components that form principal component 2 include 'smartfrencare', 'service', 'obstacle', 'modem', 'location', 'application', 'wifi', 'promotion', 'interference', and 'fast'. The variables 'smartfrencare', 'service', 'obstacle', and 'modem' are more dominant in forming PC2 compared to other PCs. There are several variables, such as 'location', 'voucher', 'application', and 'promotion', that not only form PC2 but also dominantly form other PCs.

Principal Component 3

The dominant components that form principal component 3 include '4g', 'smartfrenworld', 'unlimited', 'location', 'application', 'telkomsel', 'access', and 'pulsa'.

Principal Component 4

The dominant components that form principal component 4 include 'smartfrenworld', 'bolt', 'package', 'internet', and 'application'.

Principal Component 5

The dominant components that form principal component 5 include '4g', 'package', 'internet', 'voucher', 'price', 'promotion', 'cheap', and 'coverage'.

Principal Component 6

The dominant components that form principal component 6 include 'city', 'mifi', 'ugly', 'access', and 'bonus'.

Principal Component 7

The dominant components that form principal component 7 include 'signal', 'voucher', 'speed', 'range', and 'bonus'.

3.3 Mathematical Model of Principal Component Analysis (PCA)

The mathematical equation of variable X is formed based on the dominant variables of the principal component values as follows:

$$f_{PC1} = 0.311(\text{4g}) + 0.292(\text{smartfrenworld}) + 0.276(\text{unlimited}) + 0.216(\text{network}) + 0.247(\text{internet}) + 0.213(\text{signal}) + 0.195(\text{telkomsel}) + 0.223(\text{interference}) + 0.176(\text{bad}) + 0.221(\text{fast}) + 0.257(\text{access}) + 0.258(\text{credit}) \quad (3)$$

$$f_{PC2} = 0.303(\text{smartfrencare}) + 0.474(\text{service}) + 0.264(\text{obstacle}) + 0.329(\text{modem}) + 0.205(\text{location}) - 0.325(\text{voucher}) + 0.212(\text{wifi}) - 0.232(\text{promotion}) + 0.206(\text{interference}) \quad (4)$$

$$f_{PC3} = -0.281(\text{service}) - 0.334(\text{unlimited}) + 0.359(\text{location}) + 0.185(\text{application}) - 0.277(\text{telkomsel}) + 0.295(\text{access}) - 0.174(\text{credit}) \quad (5)$$

$$f_{PC4} = -0.239(\text{smartfrenworld}) + 0.128(\text{bolt}) + 0.536(\text{package}) + 0.257(\text{internet}) \quad (6)$$
$$+ 0.225(\text{application}) \quad (7)$$

$$f_{PC5} = 0.205(\text{4g}) + 0.251(\text{package}) - 0.285(\text{internet}) + 0.273(\text{voucher}) + 0.234(\text{price}) + 0.290(\text{promotion}) - 0.387(\text{cheap}) - 0.365(\text{coverage}) - 0.077(\text{fast}) - 0.153(\text{access}) + 0.022(\text{credit}) + 0.245(\text{bonus}) \quad (8)$$

$$f_{PC6} = 0.322(\text{city}) - 0.374(\text{mifi}) - 0.018(\text{bad}) - 0.153(\text{access}) + 0.245(\text{bonus}) \quad (9)$$

$$f_{PC7} = 0.143(\text{signal}) + 0.343(\text{voucher}) - 0.508(\text{speed}) - 0.271(\text{reach}) - 0.233(\text{bonus}) \quad (10)$$

Furthermore, for each main component, a linear equation will be built against the Z variable as an indicator of company performance (stock index) [33].

$$Z_i = a_{i,0} + a_{i,1} \times \text{PC}_{i,j} + E_i \quad (11)$$

where Z_i is the success indicator, $a_{i,0}$ and $a_{i,1}$ = constants E_i = error

3.4 Linear Equation of Principal Component Analysis (PCA)

The prediction model formulation process is built based on a linear approach involving variables X and Y. Variable X is the result of a mathematical equation of the principal components formed from Twitter data variables (shown from the equation function fPC1 to fPC7), and variable Y is obtained from company performance indicators. This study analyzes five predictable Y variables, including the price index (open, high, low, close) and stock volume.

The equation produced by the Principal Component Analysis model is used as the X variable for thirty-three periods of historical stock data by entering the variable values into the PCA equation formed (equation function fPC1 to fPC7). The results of the calculation of each principal component value are shown in Table 5. Furthermore, a linear equation of each principal component in Table 5 will be formed against the normalized value of company performance in Table 6.

Table 5: The value of Principal Component Analysis (PCA)

Periode	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	0.189512	0.803133	-0.66383	0.823459	0.380801	-0.32926	-0.31298
2	0.530555	0.914159	-0.43584	0.605631	0.354658	0.225591	-0.67925
3	1.357038	1.662369	-0.32805	1.295552	0.564369	0.08506	-0.01556
4	0.411450	0.569557	-0.07487	0.736012	0.367354	0.137648	-0.31233
5	0.512643	0.762765	-0.25913	0.329899	0.214272	0.074554	-0.09425
6	0.340984	0.661639	-0.23576	0.113073	0.227197	0.310181	-0.32827
7	0.444076	0.575995	-0.24888	0.353515	0.389328	0.091058	-0.33818
8	0.983750	0.443924	0.22417	0.762296	0.340834	0.564623	-0.62370
9	1.341572	0.484137	0.14319	0.756165	0.480119	0.469120	-0.25941
10	1.466402	0.321845	0.30471	0.837747	0.244194	0.557617	-0.23673
11	1.017967	0.221401	0.37154	0.549878	0.233854	0.346870	-0.16189
12	0.662974	0.143907	0.23053	0.372313	0.192434	0.366371	0.09356
13	1.249060	0.013203	0.50519	1.204998	0.757455	0.069854	-0.44677
14	1.381877	0.163154	0.30208	0.882215	0.527423	0.245883	0.05904
15	1.286849	-0.08290	0.33352	1.097329	0.827729	0.124074	0.03962
16	1.025152	-0.12603	0.24565	0.713615	0.595577	0.099245	-0.16393
17	1.312837	0.206150	0.33260	0.588031	0.473025	0.185236	-0.29784
18	0.943857	0.141244	0.29938	0.140488	0.391668	0.060923	-0.06967
19	0.959154	0.203763	0.32324	0.124916	0.025152	0.330666	-0.46535
20	1.048536	0.110029	0.25999	0.187136	0.282245	0.280539	-0.45587
21	1.551848	0.293436	0.54724	0.880953	0.164751	-0.051880	-0.89563
22	1.949559	-0.33729	-0.60068	1.446578	-0.75776	0.152239	-0.26773
23	1.003050	-0.10967	-0.12159	0.532562	-0.02311	0.032660	-0.16446
24	0.717143	-0.11906	0.14819	0.382231	0.103428	0.100414	-0.23471
25	0.780826	-0.06190	0.13027	0.278736	0.209717	0.027888	-0.24208
26	1.192346	-0.38962	-0.28710	0.457084	0.702312	-0.271720	-0.15364
27	2.067697	-0.20335	-0.80149	0.300251	0.854688	0.472388	-0.95393
28	3.502444	1.426407	0.12118	0.051709	0.213402	0.114485	-0.10093
29	2.269640	-0.36228	-0.91137	0.253378	0.575773	0.269453	0.06662
30	1.626458	0.075155	0.10686	0.627191	0.363737	-0.549640	-0.59019
31	1.295270	0.093658	0.01204	0.358895	0.318761	-0.403160	-0.06958
32	1.524681	0.570909	0.31785	0.341733	0.173417	0.127902	-0.11193
33	1.732009	0.651060	0.55704	0.321270	0.237979	-0.394980	-0.54592

Table 6: The normalization of company performance (stock index)

Open	High	Low	Close	Adj Close	Volume
0.00	0.00	0.00	0.00	0.00	0.49
0.06	0.01	0.07	0.02	0.02	0.27
0.08	0.08	0.09	0.09	0.09	0.48

Continued on next page

Table 6 – continued from previous page

Open	High	Low	Close	Adj Close	Volume
0.14	0.18	0.15	0.17	0.17	0.48
0.21	0.49	0.24	0.48	0.48	1.00
0.53	0.65	0.47	0.70	0.70	0.79
0.67	0.89	0.73	0.77	0.77	0.96
0.78	0.93	0.83	0.87	0.87	0.66
0.88	0.86	0.85	0.79	0.79	0.23
0.81	1.00	0.88	0.99	0.99	0.37
0.98	1.00	0.94	0.94	0.94	0.28
0.94	0.99	0.98	0.89	0.89	0.26
0.91	0.87	0.66	0.66	0.66	0.27
0.66	0.67	0.63	0.65	0.65	0.19
0.66	0.75	0.74	0.80	0.80	0.14
0.81	0.90	0.88	0.88	0.88	0.30
0.88	0.91	0.97	0.92	0.92	0.19
0.91	0.94	0.98	1.00	1.00	0.22
1.00	0.98	0.68	0.65	0.65	0.24
0.67	0.73	0.66	0.78	0.78	0.15
0.81	0.78	0.85	0.78	0.78	0.09
0.78	0.73	0.87	0.78	0.78	0.00
0.79	0.93	0.88	0.90	0.90	0.19
0.90	0.93	0.97	0.93	0.93	0.15
0.91	0.94	1.00	0.93	0.93	0.13
0.94	0.92	0.96	0.91	0.91	0.08
0.91	0.89	0.97	0.89	0.89	0.06
0.90	0.90	0.97	0.88	0.88	0.09
0.88	0.85	0.48	0.43	0.43	0.95
0.46	0.40	0.35	0.33	0.33	0.85
0.37	0.29	0.28	0.30	0.30	0.26
0.33	0.31	0.34	0.31	0.31	0.33
0.35	0.29	0.35	0.28	0.28	0.13

3.5 Principal Component Regression (PCR)

Table 6–Table 10 shows the linear relationship between the principal component results and the stock index variables that produce the R-value. The principal component analysis approach from Twitter data yields a fairly good R-value for predicting company performance based on the stock index indicator. The correlation value is shown by the R-value for each principal component (PC1 to PC7) in Table 7–Table 11. The results of the principal component analysis approach show that there is at least one principal component that can be used for prediction, indicated by the R-value of more than 0.5, based on Table 7 and Figure 4 in PC2. This is supported by PC2, which provides the best R-value for each variable that predicts stock prices with an R-value of more than 0.5. PCA is powerful and highly valuable for specific purposes, primarily dimensionality reduction and data visualization. Thus, the most appropriate principal component for predicting company performance variables

(stock price index) is equation 2, and for predicting company performance variables (stock volume), it is equation 3. This is in line with research [34] The best regression results are not always based on the highest eigenvalue. The linear Principal Component Regression equation and R-value are shown in Table 7–Table 11.

Table 7: Principal component regression (stock index-open)

Open		
Principal Component	**PCR**	**R**
PC1	$y = 0.1499x + 0.4834$	0.321
PC2	$y = -0.3807x + 0.7758$	0.595
PC3	$y = 0.2228x + 0.6579$	0.286
PC4	$y = -0.1542x + 0.751$	0.181
PC5	$y = -0.0182x + 0.6697$	0.017
PC6	$y = 0.4458x + 0.6106$	0.395
PC7	$y = 0.0381x + 0.6744$	0.033

Table 8: Principal component regression (stock index-high)

High		
Principal Component	**PCR**	**R**
PC1	$y = 0.0962x + 0.5811$	0.198
PC2	$y = -0.3578x + 0.8022$	0.537
PC3	$y = 0.2292x + 0.691$	0.283
PC4	$y = -0.1811x + 0.7995$	0.204
PC5	$y = -0.0208x + 0.7038$	0.020
PC6	$y = 0.5594x + 0.6304$	0.476
PC7	$y = 0.0981x + 0.7246$	0.081

Table 9: Principal component regression (stock index-low)

Low		
Principal Component	**PCR**	**R**
PC1	$y = 0.1312x + 0.5$	0.267
PC2	$y = -0.3458x + 0.7596$	0.514
PC3	$y = 0.2563x + 0.6511$	0.313
PC4	$y = -0.1161x + 0.7235$	0.130
PC5	$y = -0.0823x + 0.6852$	0.075
PC6	$y = 0.4567x + 0.6035$	0.385
PC7	$y = 0.029x + 0.6659$	0.024

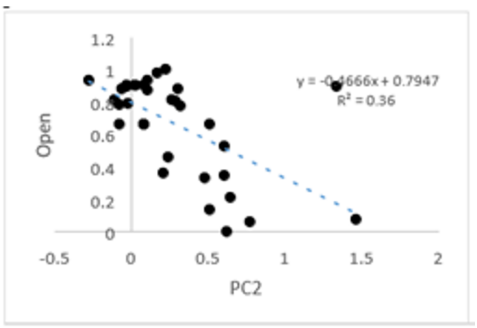
4 Discussion

Principal Component Relationship Analysis with Company Performance Indicators (Stock Index)
The principal component analysis with the best and most appropriate R-squared value for predicting the stock price index variable is shown by principal component 2 (shown in Table 7–Table 11). The results of the component analysis involve thirty variables from the Twitter social media that are considered related to various customer reviews of the company's performance. The involvement of all these variables provides better prediction results. However, the involvement of many variables will provide more complexity to the equation, such as causing data search and collection to be complicated or take a long time. The results of the dominant variables in the previous analysis can be used as variables considered in building a simpler mathematical model for each main component formed. The correlation analysis of Twitter data against the stock index is shown in Figure 4. The results of the linear equation analysis shown by the Twitter data variable against company performance (stock index) show that the analysis of Twitter social media data can be used to predict company performance through five stock index indicators quite well, as indicated by the R-squared results of more than 0.25. This shows that the correlation between the two variables is more than 50%.

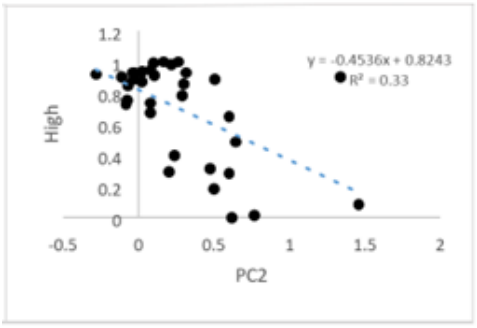
Furthermore, from the five stock index indicators, four indicators were predicted well between Twitter data and stock prices (open, high, low, and close) with strong or moderate correlation strength using the principal component equation 2 and moderate correlation is shown from Twitter data against stock sales volume indicated by the principal component 3. The principal component analysis approach from mining Twitter data has the potential to evaluate company performance in the stock market and can even be used for prediction purposes. In the research of Song et al. [35] also identified investor sentiment indicators to predict stock volatility in the Chinese stock market, which obtained the results of characterizing investor sentiment with good performance. Previous research also showed the importance of financial data such as stock prices are rich time series data that contain valuable information for investors and financial professionals, and a study was also conducted with the aim of investigating the integration of PCA and deep learning models into the Turkish stock market using indicator values and to assess the effect of this integration on market prediction performance [36].

Prior research has similarly employed the PCA approach for stock market forecasting, shown by the study conducted by Srijiranon et al. [37]. A hybrid prediction model, termed PCA-EMD-LSTM, was developed in that study, integrating principal component analysis (PCA), empirical mode decomposition (EMD), and long short-term memory (LSTM) to forecast the closing price of the stock market in Thailand one step ahead. Additionally, the research conducted by Song et al. [35] developed investor sentiment indicators to forecast stock volatility in the Chinese stock market, demonstrating effective results in characterizing investor sentiment with commendable performance. Prior research has shown that financial data, particularly stock prices, constitute rich time series data that provide significant insights for investors and financial professionals. Additionally, investigations were conducted to explore the amalgamation of PCA and deep learning models within the Turkish stock market, utilizing indicator values, and to evaluate the impact of this integration on market prediction efficacy [36]. This research employs social media data as an alternative to stock market data, which has been utilised in numerous nations to discover and pre-

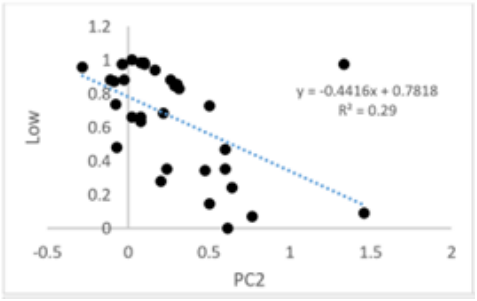
dict corporate performance. This study shifts the focus from prior research, which aimed at forecasting the stock market using historical data beneficial to investors, to identifying company performance derived from multiple online consumer opinion surveys on social media, utilising performance metrics, namely the stock market. This study's contribution is made clear through the use of online social media data to assess company performance based on stock market success metrics and to extract public data from Twitter.



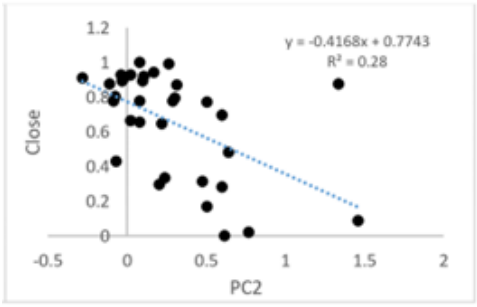
(a) Open



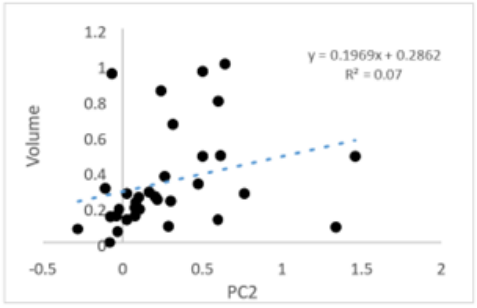
(b) High



(c) Low



(d) Close



(e) Volume

Figure 4: Relationship between principal components and company performance indicators (stock index)

Table 10: Principal component regression (stock index-close)

Close		
Principal Component	**PCR**	**R**
PC1	$y = 0.0735x + 0.5688$	0.157
PC2	$y = -0.3168x + 0.7505$	0.492
PC3	$y = 0.2574x + 0.6506$	0.329
PC4	$y = -0.1363x + 0.7345$	0.159
PC5	$y = -0.0617x + 0.6778$	0.059
PC6	$y = 0.498x + 0.598$	0.439
PC7	$y = 0.0747x + 0.6783$	0.064

Table 11: Principal component regression (stock index-volume)

Volume		
Principal Component	**PCR**	**R**
PC1	$y = -0.127x + 0.4943$	0.293
PC2	$y = 0.1472x + 0.2982$	0.247
PC3	$y = -0.2421x + 0.3477$	0.335
PC4	$y = -0.1058x + 0.4015$	0.133
PC5	$y = 0.1147x + 0.3033$	0.119
PC6	$y = -0.0189x + 0.3438$	0.017
PC7	$y = 0.0959x + 0.3686$	0.089

5 Conclusion

This study presents a method for extracting information from various customer opinions on Twitter social media regarding company performance in the stock market. This study aims to analyze and find the role of data mining in the use of abundant data (big data) from social media such as Twitter on company performance. This study uses Text Mining, Principal Component Analysis (PCA), and Principal Component Regression (PCR) approaches to predict company performance. Thus, the novelty of this study is the use of big data from customer reviews to predict company performance and using PCA and PCR as analysis techniques. One of the telecommunications industries is used as a case study in this study, and mining Twitter social media data that correlates with company performance parameters, namely the stock market index (price-low, high, open, close, and sales volume).

The results of the study found a grouping of words into principal components, which then correlated with company performance, namely the stock price index. The results of the main components formed show that there is at least one equation that is quite strongly correlated to company performance in the stock market, which is indicated by the R and R-squared values. Thus, data from Twitter's social media has the potential for predictive analysis of company performance, especially in the stock market, using a text mining and principal component analysis approach.

Furthermore, as many as thirty variables obtained from the Twitter social media analysis are simplified based on the dominant variables that make up each principal component, showing a simple equation compared to involving all variables, even showing a better cor-

relation value to meet the objectives of forming the model. The linear equation of principal component 2 is most appropriate for predicting stock prices, and principal component 3 is most appropriate for predicting stock sales volume. Thus, the principal component analysis relationship model is not based on the highest eigenvalue but based on the highest R and R-squared results.

Building upon this study's successful application of Text Mining, PCA, and PCR to predict telecommunication stock performance from Twitter data, future research should aim to broaden the scope and refine the predictive power of the model. This involves extending the methodology to diverse industries to assess its generalizability, integrating a wider array of data sources beyond Twitter (e.g., other social media platforms, news articles, traditional financial data) to enrich insights, and exploring more advanced natural language processing (NLP) and machine learning techniques (such as emotion detection, aspect-based sentiment analysis, and deep learning models like LSTMs or Transformers) to capture more complex temporal and non-linear relationships.

References

[1] D. N. Morah and O. A. Nwafor, "Beyond tribal politics for e-participation and development: social media influence on nigeria's 2023 presidential general election," *Journal of Innovation and Digital Transformation*, 2024.

[2] R. G. Cooper, "The drivers of success in new-product development," *Industrial Marketing Management*, vol. 76, pp. 36–47, 2019.

[3] D. H. Fudholi, R. A. N. Nayoan, and S. Rani, "Stock prediction based on twitter sentiment extraction using bilstm-attention," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 10, no. 1, pp. 187–198, 2022.

[4] A. Gupta and V. K. Tayal, "Analysis of twitter sentiment to predict financial trends," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 1027–1031, IEEE, 2023.

[5] A. Prada and C. A. Iglesias, "Predicting reputation in the sharing economy with twitter social data," *Applied Sciences*, vol. 10, no. 8, p. 2881, 2020.

[6] S. Urolagin and S. Patel, "User-specific loyalty measure and prediction using deep neural network from twitter data," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 1046–1061, 2023.

[7] K. K. Ramachandran, "The use of data mining in education: An overview of state of the art, limitations, and emerging research areas," *Journal of Data Analysis and Research Development (IJDARD)*, vol. 1, no. 1, pp. 1–8, 2023.

[8] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan Kaufmann, 2022. Accessed: Dec. 15, 2024.

[9] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, 2020.

[10] J. Guerreiro and P. Rita, "How to predict explicit recommendations in online reviews using text mining and sentiment analysis," *Journal of Hospitality and Tourism Management*, vol. 43, pp. 269–272, 2020.

[11] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022. Accessed: Dec. 16, 2024.

[12] J. Park, D. Yang, and H. Y. Kim, "Text mining-based four-step framework for smart speaker product improvement and sales planning," *Journal of Retailing and Consumer Services*, vol. 71, p. 103186, 2023.

[13] F. Lyu and J. Choi, "The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews," *Sustainability*, vol. 12, no. 11, p. 4383, 2020.

[14] C. Zhang, Y.-X. Tian, Z.-P. Fan, Y. Liu, and L.-W. Fan, "Product sales forecasting using macroeconomic indicators and online reviews: a method combining prospect theory and sentiment analysis," *Soft Computing*, vol. 24, no. 9, pp. 6213–6226, 2020.

[15] Y. Kim and S. R. Jeong, "Competitive intelligence in korean ramen market using text mining and sentiment analysis," *Journal of Internet Computing and Services*, vol. 19, no. 1, pp. 155–166, 2018.

[16] P. Eachempati, P. R. Srivastava, A. Kumar, J. M. de Prat, and D. Delen, "Can customer sentiment impact firm value? an integrated text mining approach," *Technological Forecasting and Social Change*, vol. 174, p. 121265, 2022.

[17] M. Jaggi, P. Mandal, S. Narang, U. Naseem, and M. Khushi, "Text mining of stocktwits data for predicting stock prices," *Applied System Innovation*, vol. 4, no. 1, p. 13, 2021.

[18] S. Urolagin, "Text mining of tweet for sentiment classification and association with stock prices," in *2017 International Conference on Computer and Applications (ICCA)*, pp. 384–388, IEEE, 2017.

[19] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," *International Journal of Information Management*, vol. 48, pp. 280–290, 2019.

[20] Subagyo, D. A. Purnama, N. A. Masruroh, and R. R. Pratama, "Modeling dynamic consumer preferences in product attributes for social media-based product improvement planning," *Malaysian Journal of Consumer and Family Economics*, vol. 32, no. 1, pp. 104–140, 2024.

[21] U. Ruhi, "Social media analytics as a business intelligence practice: Current landscape & future prospects," *Journal of Internet Social Networking and Virtual Communities*, vol. 2014, 2014.

[22] D. A. Purnama, Subagyo, and N. A. Masruroh, "Online data-driven concurrent product-process-supply chain design in the early stage of new product development," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 9, no. 3, p. 100093, 2023.

[23] R. Arifin and D. A. Purnama, "Identifying customer preferences on two competitive startup products: An analysis of sentiment expressions and text mining from twitter data," *J. Infotel*, vol. 15, no. 1, pp. 66–74, 2023.

[24] S. H. Batool, W. Ahmed, K. Mahmood, and A. Sharif, "Social network analysis of twitter data from pakistan during covid-19," *Information Discovery and Delivery*, vol. 50, no. 4, pp. 353–364, 2021.

[25] S. A. Nugroho and S. Widianto, "Exploring electric vehicle adoption in indonesia using zero-shot aspect-based sentiment analysis," *Sustainable Operations and Computers*, vol. 5, pp. 191–205, 2024.

[26] O. P. Okereka, A. E. Orhero, and U. C. Okolie, "Digital media and data collection in social and management sciences research in nigeria," *Ianna Journal of Interdisciplinary Studies*, vol. 6, no. 1, pp. 76–89, 2024.

[27] P. Kartheek, "Big data analytics on data with the growing telecommunication market in a distributed computing environment," 2023. Accessed: Jun. 17, 2025.

[28] C. Amin, A. W. Hasyim, M. Sun'an, R. M. Hilman, and H. Fahmiasari, "Impact of increasing local economic capacity on reducing maritime logistics costs in island province of eastern indonesia: A dynamic system approach," *Transportation Research Interdisciplinary Perspectives*, vol. 27, p. 101195, 2024.

[29] A. M. D'Uggento, A. Biafora, F. Manca, C. Marin, and M. Bilancia, "A text data mining approach to the study of emotions triggered by new advertising formats during the covid-19 pandemic," *Quality and Quantity*, vol. 57, no. 3, pp. 2303–2325, 2023.

[30] X. Zhang, O. L. O. Astivia, E. Kroc, and B. D. Zumbo, "How to think clearly about the central limit theorem," *Psychological Methods*, vol. 28, no. 6, p. 1427, 2023.

[31] S. K. Sarkar, "Principal component analysis," in *Statistical Procedures for Analysis of Agricultural Data Using R*, p. 139, 2023.

[32] D. A. Purnama, P. C. Marifa, and R. C. Shinta, "Unlocking indonesia's maritime potential: Optimizing hub port development using a principal component analysis and k-means clustering," *Jurnal Sistem dan Teknik Industri*, vol. 27, no. 1, pp. 35–46, 2025.

[33] J. E. Jackson, *A user's guide to principal components*. John Wiley & Sons, 2005. Accessed: Jun. 17, 2025.

[34] J. E. Jackson and A. Edward, *User's guide to principal components*, vol. 40. John Wiley & Sons, 1991.

[35] Z. Song, X. Gong, C. Zhang, and C. Yu, "Investor sentiment based on scaled pca method: A powerful predictor of realized volatility in the chinese stock market," *International Review of Economics & Finance*, vol. 83, pp. 528–545, 2023.

[36] T. Uçkan, "Integrating pca with deep learning models for stock market forecasting: An analysis of turkish stocks markets," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 8, p. 102162, 2024.

[37] K. Srijiranon, Y. Lertratanakham, and T. Tanantong, "A hybrid framework using pca, emd and lstm methods for stock market price prediction with sentiment analysis," *Applied Sciences*, vol. 12, no. 21, p. 10823, 2022.