Go-Food Sentiment Analysis Using Twitter Data, Compared the Performance of the Random Forest Algorithm with That of the Linear Support Vector Classifier

Muhammad Abdullah Hadi, Nizirwan Anwar(✉), Budi Tjahjono, Lina, Binastya Anggara Sekti, Yunita Fauzi Achmad, and Yulhendri

Esa Unggul University, Jakarta 11510, Indonesia
nizirwan.anwar@esaunggul.ac.id

Abstract. As a generalization, many modern consumers now favor using one of the many available e-commerce websites to do their shopping. Customers can save time and energy by shopping online instead of going out to physical stores because they can do so whenever they like, from wherever they like. Eighty percent of the dataset is used for training, while twenty percent is used for validation. With these default settings for the training data, the random forest algorithm is applied to the classification with 40 n estimators and linear SVC. Accuracy, precision, recall, and the F-measure are just a few of the quantitative metrics we employ to assess the quality of the model. Random forest has a 98.6% success rate, while linear SVC only achieves a success rate of 98%. Training data for a random forest can take up to 5 min, but training data for a linear SVC only takes 1 min. Sentiment analysis performed with machine learning's random forest algorithm and linear SVC on Go-Food reviews in Indonesian found that positive sentiment was still higher than negative sentiment as of June 2022.

Keywords: Association Rules · A priori Algorithm · Data Mining

1 Introduction

In recent years, the use of social media to share information across the globe or universe has grown in popularity. They disseminate the information via Facebook, Twitter, and other social media platforms. Not only do they share the information, but they also comment on it, whether the comment is positive or negative [3]. With the rapid growth and widespread adoption of e-commerce technology, an increasing number of users prefer to shop on various e-commerce platforms. In contrast to offline shopping in physical stores, users can shop online at any time and from anywhere, saving time and effort [16]. Due to the virtual nature of e-commerce platforms, there are numerous issues with the products sold on these platforms, such as inconsistency between descriptive information and actual goods, poor quality of goods, and inadequate after-sales services, among others [25].

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 3–13, 2022.
https://doi.org/10.2991/978-94-6463-084-8_2

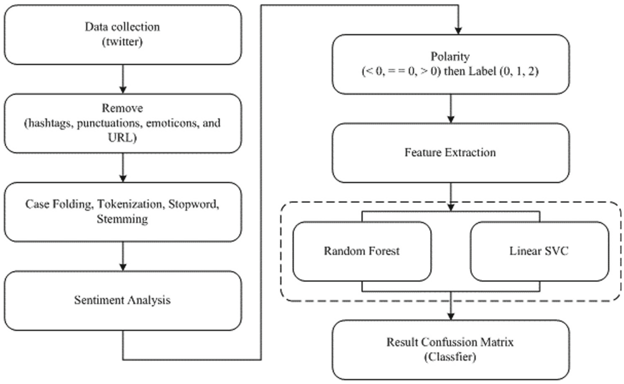


Fig. 1. Sentiment analysis process

The objective of sentiment analysis is to determine the polarity of text sentiment. The analysis of sentiment could be viewed as a classification problem [3]. The paper primarily investigates three class problems that categorize the text as positive, negative, or neutral. Using sentiment classification techniques to automatically collect diverse perspectives from various platforms, this analysis helps researchers and decision-makers gain a deeper understanding of opinions and client satisfaction. There has been a great deal of research on the classification of emotions. Historically, the majority of it has centered on classifying longer pieces of text, such as reviews in [21]. Sentiment analysis for product reviews, also known as text orientation analysis or opinion mining, is the process of automatically analyzing the subjective commentary text with the customer's emotional hue and deriving the customer's emotional disposition [8].

Sentiment analysis benefits both the provider and the purchaser, as it assists the provider in introducing new products and the purchaser in locating the original product by utilizing user reviews on online websites or mobile devices [26]. It requires significantly more language comprehension than text analysis and subject classification. Indeed, if the simplest algorithms only consider the statistics of the frequency of occurrence of words, it is typically insufficient to identify the dominant opinion in a document [34]. In this research, unstructured data of Go-Food Reviews have been taken from Twitter. It has been filtered to remove noisy data like emoticons, typos, punctuations, etc. It has been preprocessed to evaluate the sentiment of the reviews using polarity analysis, feature extraction using TF-IDF, and supervised learning using random forest and linear support vector classifier.

2 Research Methodology

Figure 1 depicts the design of the proposed method for sentiment analysis of tweets. For preprocessing, text cleaning and preprocessing are utilized. The training phase involves training the tweets. The final step is the evaluation of sentiment using the test dataset.

2.1 Data Collection

This study's sentiment analysis data was collected from Twitter. Using the SNScrape Python library, tweets from users will be collected in real-time based on specific keywords. From 1 June 2022 to 30 June 2022, Twitter provided data for 28763 tweets containing the keyword Go-Food in Indonesian. The data is processed with Python using NLTK, Sastrawi for the Indonesian language, TextBlob, TF-IDF, and Random Forest Classification, among others. Several steps are required to achieve optimal results when analyzing emotions. There are four steps involved: data collection, preprocessing, feature extraction, and sentiment analysis.

E-Commerce. As e-commerce continues to expand, it is natural for barriers that limit dynamics to emerge. According to multiple studies, e-commerce businesses face a variety of obstacles and challenges. Customer service, distribution and logistics, payment processing, infrastructure, the economy, fiscal and legal issues, and marketing are among the fundamental obstacles that impede the widespread use of the Internet in the e-commerce industry. Due to the shorter distance and delivery time required, this phenomenon is less prevalent among food manufacturers who sell primarily to local consumers. A specialized route for delivery vehicles is unnecessary in the case of perishable food [7]. Occasionally, a stimulus, such as a random event [1] or a global problem like the COVID-19 epidemic [32], is required for the expansion of the online food trade.

Social Network. The objective of social networking sites (SNS) is to facilitate social interactions between users. Users generate the content of social networking sites. It contains voluminous information about users, shared ideas and thoughts, and real-time data regarding users' conversations and statuses. In addition to the growth of SNS users, the rate of data in SNS demonstrates the importance of SNSs in numerous fields for real-time analysis and forecasting [2]. Twitter is a microblogging service that was officially launched on 13 July 2006 [17]. The primary purpose of Twitter is to publish brief messages (tweets) via the web or mobile devices. 280 characters is the maximum character count for a tweet. Twitter is an almost inexhaustible source for classifying text. Twitter tweets have numerous attributes [11].

2.2 Text Cleaning and Data Preprocessing

During the method proposal phase, Twitter data was collected. Preprocessing includes data cleansing, case folding, tokenization, stop word elimination, and stemming. Next, conduct sentiment analysis with a random forest algorithm and linear SVC.

- The process of cleaning irrelevant tweet data to make it relevant is known as data cleansing.
- Case folding is the process of converting words to the same case, such as all capital letters or all lowercase letters.
- Tokenization is the process of separating sentences into tokens. Tokens can be composed of words, phrases, or other meaningful elements.
- Stopword removal is the elimination of common and frequently used words without affecting sentence structure significantly. According to the list of stop words, Indonesian Twitter messages contain and, or, etc.

- Stemming is the removal of affixes and suffixes from a word to obtain its root [23]. Continue conducting sentiment analysis with the random forest algorithm.

Text Mining. Text mining is the rapidly evolving practice of discovering and extracting information from vast unstructured textual resources. Text mining is capable of utilizing unstructured data sources [15]. Text mining involves three steps:

- Information extraction
- Partial, superficial, and in-depth language analysis
- Identify pertinent entities and facts about entities
- Data mining
- Combine and link facts Information extraction.

Natural Language Processing (NLP). Theoretically, natural language processing (NLP) is a computational technique for analyzing and describing text naturally in one or more levels of linguistic analysis, with the goal of achieving the ability to comprehend the essence of a sentence as humans do, for various tasks and applications. The theory and implementation of natural language processing are applicable to a variety of applications [31]. Bhuvneshwar Kumar et al. [8] focused on Natural Language Processing (NLP) issues to differentiate between positive and negative customer reviews of products on the internet. This information was compiled using data from Amazon.com, Rediff.com, and Flipkart.com.

2.3 Feature Extraction

Maximum accuracy can be achieved in machine learning by selecting the most advantageous features. Feature selection is therefore a crucial step in any classification problem. The set of features in text classification is a subset of words that can be used to differentiate between different classes [27]. The chosen words should provide classification-related information that is useful. Consequently, it is essential to consider various techniques for converting text into a form that can be processed to extract the necessary information. In this study, the user characteristics consist of term-based, sentiment-based, and Go-Food-related keyword characteristics. Features Based on Emotion. Using contextual polarity, sentiment analysis is the process of extracting the sentiment of a text. It is frequently used to categorize reviews of various products on the Internet, such as the tone of movie reviews. In this study, we assigned a sentiment to each tweet using the TextBlob library [22]. TextBlob is a Python library used for textual data analysis. A positive, neutral, or negative sentiment value is assigned to a tweet based on its polarity score. Only the IDF and TF-IDF techniques, out of the four types of word weighting techniques, consider the importance of a word/term in the entire corpus, as opposed to the importance of the word/term in a single document. It has been demonstrated in [31] that TF-IDF is more precise than IDF. Where n is the total number of documents and NJ is the total number of documents that contain the term j.

Term Frequency–Inverse Document Frequency (TF-IDF). As a weighting method, TF-IDF is widely known and utilized, and its performance is comparable to that of innovative methods. Considered to be term weighting factors are documents. The primary preprocessing step required to index documents is considered when selecting the feature for the feature selection procedure [19]. STW is used to weight the features in information retrieval applications involving supervised learning, such as text categorization and filtering [10, 11]. Although the term frequency-inverse document frequency (TF-IDF) technique was proposed by [12], it is regarded as the standard technique for weighting keywords in feature extraction. Combining the concepts of term frequency and inverse document frequency results in a composite weight for each term in every document.

$$tf - idf = tf \times idf \tag{1}$$

2.4 Data Preprocessing

In the first step of data processing, the researcher extracts features using TF-IDF by reducing the maximum number of features to 10000. Then, distinguish training data from test data. The percentage of training data is 80%, while testing data is 20%. The classification is then performed on the training data using the random forest algorithm. The classification results obtained from the training data are then applied to the test data. After classifying the data, the performance evaluation is conducted.

Sentiment Analysis. Sentiment analysis is an application of Natural Language Processing, computational linguistics, and text analysis that deals with people's opinions [30]. In recent years, sentiment analysis has been a valuable tool in social media for determining the general public's opinion on a specific topic. Extensive research has been conducted in the field of sentiment, ranging from hotel reviews [20] to movie reviews [24], with the goal of eliciting opinions regarding topics, trends, etc.

Random Forest. In 2001, Breiman presented the RF classification algorithm [32]. The Random Forest is a collection of unpruned regression and classification trees [5]. A RF is constructed from the extracted Bootstrap sample. The Bootstrap sample then implements recursive partitioning. The q predictors at each node are selected at random from the p predictors. The recursive partitioning is completed, and a tree is constructed. The preceding procedures are repeated until a forest is created. The forest-based classification is determined by the majority vote of all trees [11].

Linear Support Vector Classifier. Linear SVM classifiers are sufficient for a variety of applications [27]. Not all problems, however, are even roughly linearly separable [6]. In the vast majority of cases, the data for specific classes lies on multiple disjoint lower-dimensional manifolds, rendering linear classifiers inapplicable [18]. Regarding the conventional linear SVM, the following formulation is used:

$$H(x) = w^T x + b = \sum\nolimits_{i=1}^{n} w_i x_i + b \tag{2}$$

2.5 Data Processing

After training the data and loading the random forest-based classification model, we examine the metrics to determine the performance of the constructed classification model. These metrics include of precision, Remember, and the F-measure. If the dataset is unbalanced, the precision metric is faulty. Precision and remember are the most advantageous measures for unbalanced data sets. The selected metrics can be computed using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) measures, where TP refers to the rate of correctly classified instances as positive, TN refers to the rate of correctly classified instances as negative, and FP refers to the rate of incorrectly classified instances as positive. The fraction of erroneously identified negative instances is denoted by FN. Accuracy. It is a statistic used to evaluate the performance of a prediction model. It is the proportion of labels that are appropriately categorized. It is determined by Eq. 3:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision. It measures the actual positive forecasts. The precision of a model is determined using Eq. 4:

$$Precision = \frac{TP + TN}{TP + FP} \tag{4}$$

Remember. This is a sensitivity measurement. It is used to assess a model's ability to predict positive labels. It is computed using Eq. 5:

$$Remember = \frac{TP + TN}{TP + FN} \tag{5}$$

F-Measure. This measurement takes both Remember and precision into account. It can be viewed as a weighted average of precision and Remember values ranging from 0 (worst) to 1 (best) (best). F-measure is computed with Eq. 6:

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

3 Methodology

This section describes the outcome of our experiment and evaluates the performance of the random forest classification method. This study employs a training dataset of 23010 tweets and a test dataset of 5753 tweets. While random forest is classified as machine learning, confusion matrix is divided 80:20 for training and test data. According to polarity analysis, data is labeled as positive if >0, negative if <0, and neutral if $==0$. In Table 1, several examples of such tweets are labeled with various emotions.

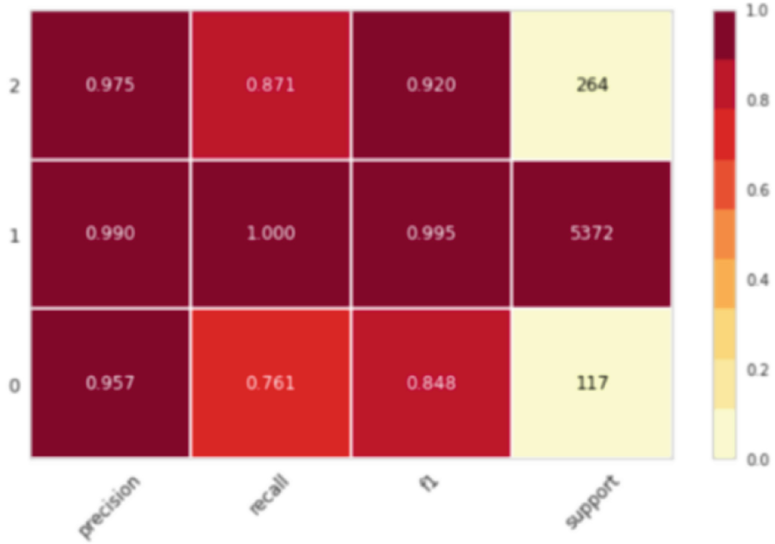
Table 1. Examples of Classified Tweets

Tweets	Sentiment	Polarity
wkwkwkwk inget banget pas puasa nih orang saran enak Go-Food	Neutral	0
yukk sisa voucher goride gocar Go-Food zonauang zonajajan Go-Food	Neutral	0
Go-Food jogja taiwan ongkirnya sih	Neutral	0
fix sih abis dpt kuliah dmn kudu bljr masak serius smp at least dah bikin takeran bumbu pas yakali ngekost Go-Food trs	Negative	−0.3
seliwer mulu mad coco ig gue ikut Go-Food kepo ramenya buset untung abang gojek nungguin	Negative	−0.625
huhu sorry uda sold Go-Food yaa uda bookingan open soree doang	Negative	−0.25
gilee Go-Food tokped brani pake bts kerenn pokok iklan apa pake bts udh tw lahh aplikasi sultan babang gojek lucky bgt gess jimin pas cemberut manis bgt Go-Food tokopedia bts iklan bts	Positive	0.33
menu rekomendasikGo-Food nemenin nonton bts dimsum nemenin nonton make it right	Positive	0.28
menu rekomendasikGo-Food nemenin nonton bts suka army udah mixue favorite banget kalo nonton siang siang wajib banget sen udah sedia Go-Food tinggal satsetsatset dateng minum nonton seger banget	Positive	0.5

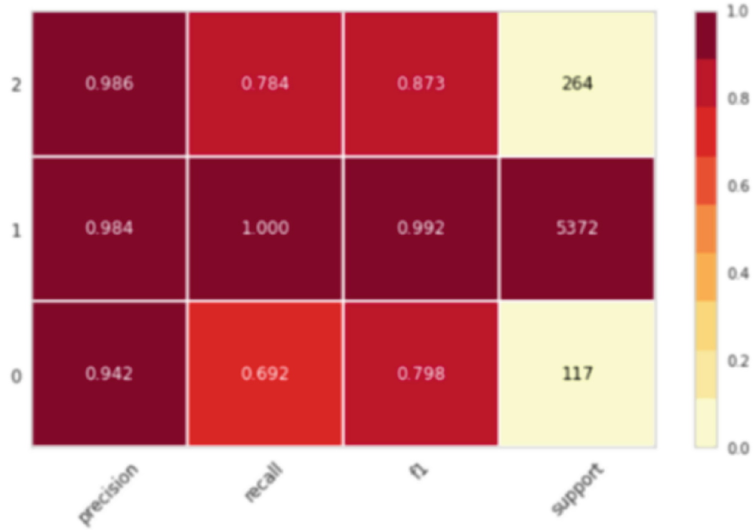
3.1 Metric Evaluations and Classification Report

As shown in Fig. 2, a total of 5753 comments/tweets generated 220 true positive comments, 92 true negative comments, and 5364 true neutral comments based on the data in the Table 2. The random forest algorithm requires approximately five minutes to computer.

Based on the results of Table 2, we can evaluate the proportion of categorization metrics belonging to each class, with 0 indicating a negative class, 1 neutral, and 2 positive. The percentage of estimated Remember, precision, and F1-score for each class is visualized in Fig. 2. Data with a total of 5753 comments/tweets yielded 209 comments declared true positive, 72 comments declared true negative, and 5367 comments declared true neutral according to the linear support vector classifier classification results in the table below (See Table 3). This algorithm's computation time is approximately 1.57 seconds.



(a)



(b)

Fig. 2. **a** Metrics Evaluations Percentage with Random Forest; **b** Metrics Evaluations Percentage with Linear SVC

The classification of the 28763-item dataset, which was divided into 80 percent for the training set and 20 percent for the test set using the random forest algorithm, yields a reasonably high level of accuracy of 99 percent. as for additional measures such as precision, Remember, and f-measure (See Table 4).

Table 2. Metrics Evaluations of Random Forest Classification

Classification Categgories	True Positive	True Neutral	True Negative
Prediction Positive	220	33	2
Prediction Neutral	2	5364	1
Prediction Negative	3	36	92

Table 3. Metrics Evaluations of Linear SVC

Classification Categgories	True Positive	True Neutral	True Negative
Prediction Positive	209	59	3
Prediction Neutral	3	5367	0
Prediction Negative	4	36	72

The results of sentiment analysis using machine learning by comparing random forest and linear svc on a dataset obtained from Twitter with the keyword Go-Food in Indonesian indicate that the performance of the Go-Food platform is relatively good, as the proportion of positive sentiment analysis results remains greater than the proportion of negative sentiment analysis results.

4 Conclusion

Sentiment analysis is essential for anyone who makes decisions, especially management and product managers. This is useful for calculating, recognizing, and expressing opinions on the platform of the being developed product or application. We compare the classification performance of the linear support vector classifier and the random forest algorithm for sentiment analysis data classification. The model is then evaluated using a metric evaluation methodology consisting of accuracy, precision, and remember, which are useful for diagnosing the model's performance. The result of accuracy for both algorithms 98,6% for random Forest and 98% for Linear SVC. However, during the training phase, the random forest algorithm requires approximately 5 min, while the linear support vector classifier requires only 1.57 s.

References

1. A. Nurwidyantoro and E. Winarko.: Event detection in social media: A survey. Proc. - Int. Conf. ICT Smart Soc.: Think Ecosyst. Act Converg. ICISS 2013, pp. 307–311 (2013)
2. A. M. Ramadhani and H. S. Goo.: Twitter sentiment analysis using deep learning methods, In: 2017 7th International Annual Engineering Seminar (InAES), pp. 1–4 (2017)
3. A. K. Keshri, B. K. Mishra, and B. P. Rukhaiyar.: When rumors create chaos in e-commerce, Chaos, Solitons and Fractals, vol. 131, no. xxxx, p. 109497 (2020)

4. L. Yang, Y. Li, J. Wang, and R. S. Sherratt.: Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning, IEEE Access, vol. 8, pp. 23522–23530 (2020)
5. P. Ji, H. Y. Zhang, and J. Q. Wang.: A Fuzzy Decision Support Model with Sentiment Analysis for Items Comparison in e-Commerce: The Case Study of http://PConline.com, IEEE Trans. Syst. Man, Cybern. Syst., vol. 49, no. 10, pp. 1993–2004 (2019)
6. A. Smith.: Image segmentation scale parameter optimization and land cover classification using the random forest algorithm, J. Spat. Sci., vol. 55, no. 1, pp. 69–79 (2010)
7. S. J. Lewis.: Thumbs up, American Journal of Orthodontics and Oral Surgery, vol. 31, no. 9. Elsevier BV, pp. 481–482 (1945)
8. C. Chauhan and S. Sehgal.: Sentiment analysis on product reviews, In: 2017 International Conference on Computing, pp. 26–31. Communication and Automation (ICCCA) (2017)
9. P. Karthika, R. Murugeswari, and R. Manoranjithem.: Sentiment Analysis of Social Media Network Using Random Forest Algorithm, IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2019, pp. 1–5 (2019)
10. Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang.: RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12343 LNCS, pp. 503–515 (2020)
11. Y. Al Amrani, M. Lazaar, and K. E. El Kadirp.: Random Forest and support vector machine-based hybrid approach to sentiment analysis, Procedia Comput. Sci., vol. 127, pp. 511–520 (2018)
12. R. A. Priyantina and R. Sarno.: Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity and LSTM, Int. J. Intell. Eng. Syst., vol. 12, no. 4, pp. 142–155 (2019)
13. Y. Jeevan Nagendra Kumar, B. Mani Sai, V. Shailaja, S. Renuka, and Panduri.: Python NLTK sentiment inspection using naïve bayes classifier, International Journal of Recent Technology and Engineering, vol. 8, no. 2 Special Issue 11. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 2684–2687 (2019)
14. T. Joachims.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorixation, Proc. ICML97 (1997)
15. V. A. Fitri, R. Andreswari, and M. A. Hasibuan.: Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm, Procedia Comput. Sci., vol. 161, pp. 765–772 (2019)
16. A. M. Florio, D. Feillet, and R. F. Hartl.: The delivery problem: optimizing hit rates in e-commerce deliveries, Transp. Res. Part B Methodol., vol. 117, pp. 455–472 (2018)
17. C. I. Hsu, S. F. Hung, and H. C. Li.: Vehicle routing problem with time-windows for perishable food delivery, J. Food Eng., vol. 80, no. 2, pp. 465–475 (2007)
18. M. M. Mostafa.: More than words: Social networks' text mining for consumer brand sentiments, Expert Syst. Appl., vol. 40, no. 10, pp. 4241–4251 (2013)
19. 2009 Go, Huang, & Bhayani.: Twitter Sentiment Analysis (final project results), J. Inform. (2009)
20. Property Valuation Using Linear Regression and Random Forest Algorithm, International Journal of System Dynamics Applications, vol. 10, no. 4. IGI Global, p. 0 (2021)
21. W. Gao and Z. Ding: Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm. Security and Communication Networks, vol. 2022. Hindawi Limited, pp. 1–9 (2022)
22. Ö. Akar and O. Güngör.: Classification of multispectral images using Random Forest algorithm, Journal of Geodesy and Geoinformation, vol. 1, no. 2. UCTEA Chamber of Surveying and Cadastre Engineers, pp. 105–112 (2012)

23. A. F. Istifani and Sholiq.: Inspirasi Profesional Sistem Informasi, J. Sisfo, vol. 07, no. 02, pp. 165–180 Sistem (2018)
24. N. Yang, X. Cui, C. Hu, W. Zhu, and C. Yang: Chinese Social Media Analysis for Disease Surveillance. In: 2014 International Conference on Identification, Information and Knowledge in the Internet of Things, pp. 17–21 (2014)
25. Rai, R., Shukla, S., Singh, B.: Reactive Power Based MRAS for Speed Estimation of Solar Fed Induction Motor with Improved Feedback Linearization for Water Pumping. IEEE Trans. Ind. Informatics **16**(7), 4714–4725 (2020)
26. A. Alessa, M. Faezipour, and Z. Alhassan: Text Classification of Flu- Related Tweets Using FastText with Sentiment and Keyword Features. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 366–367 (2018)
27. L. Istrati and A. Ciobotaru: Automatic Monitoring and Analysis of Brands Using Data Extracted from Twitter in Romanian. In: Lecture Notes in Networks and Systems, vol. 296, pp. 55–75 (2022)
28. A. Kulai, M. Sankhe, S. Anglekar, and A. Halbe: Emotion Analysis of Covid Tweets using FastText Supervised Classifier model. In: 2021 International Conference on Communication information and Computing Technology (ICCICT), pp. 1–6, 2021
29. Ramya, M., Pinakas, J.A.: Different Type of Feature Selection for Text Classification. Int. J. Comput. Trends Technol. **10**(2), 102–107 (2014)
30. F. Debole and F. Sebastiani: Supervised term weighting for automated text categorization. In: Proceedings of the ACM Symposium on Applied Computing, pp. 784–788 (2003)
31. J. Chen, C. Chen, and Y. Liang: Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word. pp. 114–117 (2016)
32. B. Fulkerson, A. Vedaldi, and S. Soatto: Class segmentation and object localization with superpixel neighborhoods. In: Proceedings of the International Conference on Computer Vision ({ICCV}) (2009)