applied sciences

MDPI

Article

Innovative Forward Fusion Feature Selection Algorithm for Sentiment Analysis Using Supervised Classification

Ayman Mohamed Mostafa *, Meeaad Aljasir, Meshrif Alruily, Ahmed Alsayat and Mohamed Ezz

College of Computer and Information Sciences, Jouf University, Sakaka 72388, Aljouf, Saudi Arabia
* Correspondence: amhassane@ju.edu.sa

Abstract: Sentiment analysis is considered one of the significant trends of the recent few years. Due to the high importance and increasing use of social media and electronic services, the need for reviewing and enhancing the provided services has become crucial. Revising the user services is based mainly on sentiment analysis methodologies for analyzing users' polarities to different products and applications. Sentiment analysis for Arabic reviews is a major concern due to high morphological linguistics and complex polarity terms expressed in the reviews. In addition, the users can present their orientation towards a service or a product by using a hybrid or mix of polarity terms related to slang and standard terminologies. This paper provides a comprehensive review of recent sentiment analysis methods based on lexicon or machine learning (ML). The comparison provides a clear vision of the number of classes, the used dialect, the annotated algorithms, and their performance. The proposed methodology is based on cross-validation of Arabic data using a *k*-fold mechanism that splits the dataset into training and testing folds; subsequently, the data preprocessing is executed to clean sentiments from unwanted terms that can affect data analysis. A vectorization of the dataset is then applied using TF–IDF for counting word and polarity terms. Furthermore, a feature selection stage is processed using Pearson, Chi^2, and Random Forest (RF) methods for mapping the compatibility between input and target features. This paper also proposed an algorithm called the forward fusion feature for sentiment analysis (FFF-SA) to provide a feature selection that applied different machine learning (ML) classification models for each chunk of *k* features and accumulative features on the Arabic dataset. The experimental results measured and scored all accuracies between the feature importance method and ML models. The best accuracy is recorded with the Naïve Bayes (NB) model with the RF method.

Keywords: sentiment analysis; machine learning; cross-validation; vectorization; feature importance

Citation: Mostafa, A.M.; Aljasir, M.; Alruily, M.; Alsayat, A.; Ezz, M. Innovative Forward Fusion Feature Selection Algorithm for Sentiment Analysis Using Supervised Classification. *Appl. Sci.* **2023**, *13*, 2074. https://doi.org/10.3390/app13042074

Academic Editors: Yue Wu, Xinglong Zhang and Pengfei Jia

Received: 2 January 2023
Revised: 2 February 2023
Accepted: 3 February 2023
Published: 5 February 2023

1. Introduction

Sentiment analysis is considered a natural language processing (NLP) method for analyzing users' orientations toward services and topics under consideration. The goal of sentiment analysis mechanisms is to differentiate between subjective and objective sentiments. Objective sentiment is used to express general facts, while subjective sentiment is based on polarity terms that express user reviews or opinions. Objective sentences are excluded during the analysis of sentiments, whereas subjective sentiments can be classified into positive, negative, or neutral polarities.

Most peoples' and users' feelings towards different topics are reflected on social media reviews and sites [1]. Social media allow users to share their views, opinions, and emotions to classify the main service and enhance its specifications in the future. The Arabic language is widely applied on most social media platforms, such as Twitter and Facebook. The Arabic language is considered the official language of Middle East countries and North Africa, comprising 27 countries in addition to the other countries that consider the Arabic language one of its popularly used dialects. It has recently attracted more attention due to the increasing use of Arabic in social media platforms [2].

In addition, the Arabic language contains different dialects with high morphological meanings that can be categorized as standard Arabic and colloquial. Most Twitter users, especially those writing and speaking in English, can express their opinions using traditional or colloquial sentences. They can also use mixed terms that make the preprocessing and analysis processes more complex [3]. Since the Arabic language contains a vast number of linguistics and terminology that are challenging to clean and analyze, sentiment analysis has recently attracted a lot of attention. In addition, sentiment analysis and prediction become more difficult with free writing on social media, particularly in the Arabic language [4].

Few research methodologies are conducted to analyze Arabic sentiments on social media due to the high morphological linguistics in each Arabic sentence that is difficult to classify and analyze. Arabic Social Media Analysis for Arabic (ASAD) aims to fill an important gap in analyzing social media in the Arabic language [5]. Marketing analysis of services or products and public responses to events, persons, and pandemics are considered major examples of the dire need to analyze Arabic sentiments efficiently and accurately, especially when the dataset is new and has not been trained earlier [6]. Recent research methodologies of sentiment analysis depend mainly on collecting datasets from social media such as Facebook and Twitter that provide expressive sentiments from several domains. Twitter users create huge volumes of text to convey their views [7] with a wide range of terms, fields, services, and products [8–10]. Twitter datasets are collected from different sources and categories for classifying public events, pandemics, and product marketing [11–13]. The analysis of sentiments on the Twitter dataset has gained more interest as large companies and institutions depend mainly on user reviews to enhance and upgrade their business services. In addition, the simplicity of the Twitter platform makes it one of the most powerful social networks in the world, with a high volume of daily-generated sentiments [14]. Sentiment analysis methodologies concentrate on mining texts and sentences that can explore deep visions and insights into users' attitudes and opinions.

The main analysis strategy is extracting, classifying, and analyzing the sentiments related to several categories, such as emotional, cognition, social, and theoretical, and analyzing complex texts. Furthermore, most retrieved texts from the Twitter dataset contain unstructured texts that need more preprocessing steps to become more concise and clearer. This can increase the complexity of the selected analysis methodology [15]. In addition to the extracted user text, user-generated data is another additional direction for retrieving data. These data can reduce the users' uncertainty towards business or E-commerce products, which helps analyze user opinions and polarity sentiments for applying decision-making strategies [16]. Whenever a machine learning (ML) technique is applied to analyze polarity sentiments, there must be a set baseline and accuracy parameters to follow. The first step in analyzing sentiments is to remove stop words, elongation terms, symbols, and irony terms that can affect the accuracy and performance of the analysis process [17]. The training phase performs a feature extraction applied to the ML technique. In contrast, the testing or prediction phase applies the features to the classifier models to determine the term polarity. The contribution of the paper is presented as follows:

1. Presenting a current perspective of the primary strategies and algorithms for Arabic sentiment analysis with a thorough examination of applicable dialects, binary and multi-classification, and annotation algorithms.
2. Providing a multi-stage methodology for feature generation and selection of Arabic terms using TF–IDF.
3. Proposing a model using the FFF-SA algorithm that adopts a forward filter for the feature selection method for scoring and registering the accuracy of each *k*-chunk feature of Arabic terms and the subsequent accumulative features.
4. Measuring and scoring the accuracy for each conducted result, proving the high performance of the NB model with the RF method.

The paper sections are organized as follows. Section 2 explains the main sentiment analysis mechanisms that are conducted to predict user opinions. Section 3 highlights

the comparative analysis of recent annotation algorithms with their applied dialects and performance. Section 4 explores the proposed methodology for cross-validation, feature generation, and feature selection of data. Section 5 provides the proposed FFF-SA algorithm for sorting, selecting, and filtering polarity term features. The experimental results are explained in Section 6, and the conclusion and future works conclude the paper in Section 7.

2. Sentiment Analysis Mechanisms

The sentiment analysis mechanism is based on text analysis and natural language processing (NLP), which aims to identify, extract, and analyze the polarity of sentiments from different sources and languages. The first process for identifying the sentiments is to discriminate between subjective and objective sentences. Subjective sentiments contain polarity terms that reflect the users' attitudes in social media reviews. In contrast, objective sentiments are based on general facts or information that do not reflect the user's orientation.

The sentiment polarity can be verified based on several weights and scales. Most research methods depend mainly on analyzing the sentiments, whether they are positive, negative, or neutral. Positive polarities express the positive orientation of the users towards a service or a topic under consideration. In contrast, negative polarities denote the opposite meaning to express the user's negative orientation towards the service. Neutral orientations mean that the detected positive and negative polarities are equal; therefore, the user orientation towards a topic or a service is fair. In addition, the sentiment classified as neutral may contain neither positive nor negative terms to be detected. In Figure 1, the overall sentiment analysis mechanism is explained. The main process for analyzing user reviews and tweets centrally depends on lexicon-based and machine-learning (ML) approaches. Each approach has its advantage, methodology for implementation, and methods for dealing with input data and user reviews. In addition, each conducted approach must be measured based on data processing, accuracy, and performance. The methodology's performance can also be changed according to the level of analysis of the polarity terms. There are three main analysis levels: aspect, sentence, and document. In addition, analyzing Arabic sentiments with highly complex terms with different linguistics is considered a challenging process.

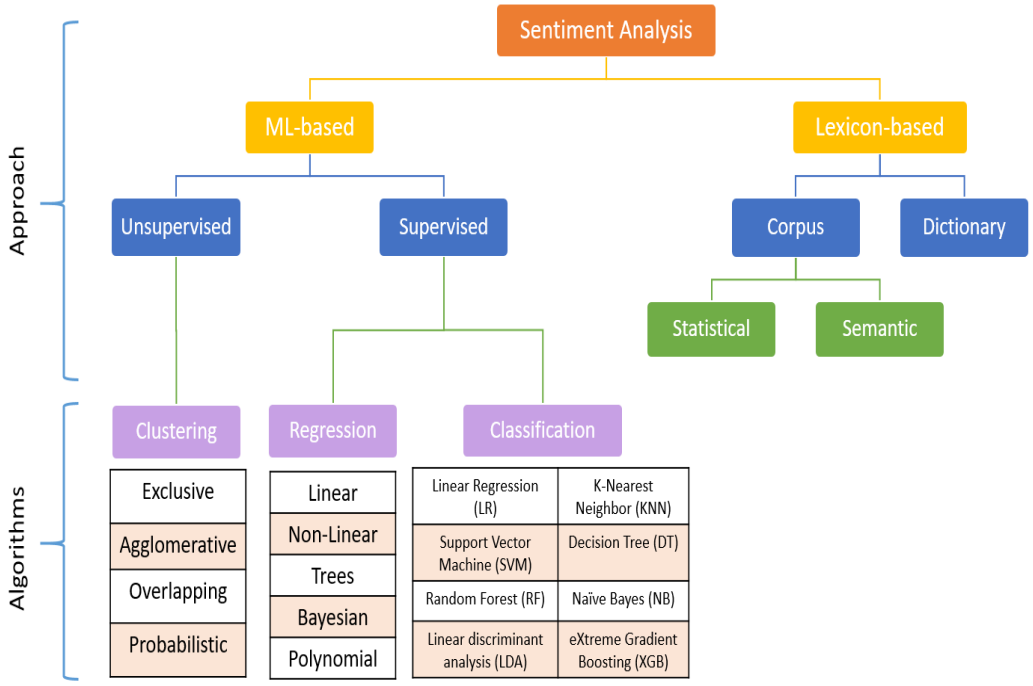

Figure 1. Sentiment analysis approaches and algorithms.

2.1. Lexicon-Based Approach

The lexicon-based approach aims to score every extracted polarity term from the sentiment sentences, and then the overall polarity of the sentence is calculated. To explain the mechanism of the lexicon-based approach, the polarity weight is first defined to distribute and score each detected polarity term according to its meaning and orientation. Second, the number of detected positive and negative terms is calculated according to their predefined weight score in the sentence. Therefore, the main sentence polarity is defined, and the whole document polarity is analyzed accordingly.

The lexicon-based approach has two main categories: dictionary-based and corpus-based. The dictionary-based stores initial word terms from different sources, and then the dictionary is extended by incorporating additional synonym terms using automated and manual annotations. Therefore, the performance of the dictionary method is constantly changing according to the size of the stored word and polarity terms. The corpus-based method is based on building a corpus that can store different Arabic dialects along with their meaning and orientations. Building a corpus is considered very time-consuming, especially for the Arabic language, which requires adding each term with its corresponding meanings from different dialects.

Arabic is a rich language with additional linguistic terms and several meanings in different dialects. For example, the term "حلو" which means "Good" has several synonyms in different Arabic dialects. The Egyptian and Sudanese dialects use "كويس", the Saudi dialect uses "زين", while some North African dialects use "باهي" or "طيب". All terms reflect the same meaning of "Good" but with different Arabic dialects.

The creation of the corpus depends mainly on statistical or semantic methods. The statistical method measures the behavior of the detected polarity terms in each sentence. If the orientation of the terms is positive, then the sentence polarity will be positive and vice versa. On the other hand, the semantic method assigns a score value to each word term. Words with similar or closer intensified meaning to the word term will have the same score value.

One of the recent methods for managing sentiment sentences was presented in [18], where a mechanism was proposed to embed words to reduce the length of the sentiment. By performing word embedding, each word was converted to its embedded word to reduce its dimension. The dimension reduction can help increase the prediction of sentiment orientation. Therefore, the mapping between predicted and actual polarity scores showed high results. Another enhanced sentiment analysis framework for normalizing the morphological terms of the Arabic language was proposed in [19]. The authors considered two main methods based on the aspect level of sentiments. The two methods were based on the orientation of both category and term polarities, where the normalization of text was executed after the classification. In addition, the authors built a word encoder and decoder to match the word term and polarity term for the given sentiment with their corresponding target meaning in another Arabic dialect.

The process of handling the Arabic language based on its dialects and idioms is considered another major concern. Many social media users and followers use different expressions and idioms based on aphorisms, wisdom, and popular proverbs that can highly affect the analysis performance. The authors of [20] proposed an algorithm for handling this issue. The algorithm's objective was to store the root of the polarity word with the emotions in the sentence that can guide the possible orientation of the sentence.

The authors of [17] built a corpus for measuring the sample percentage with its accuracy and error rate to explore the major concerns that affect the analysis of the Arabic dataset. The authors performed manual, mixed, double-check, and non-check experiments and compared the efficiency of the analysis process. As proposed in [21], a lexicon-based mechanism was presented for analyzing Arabic polarity terms from a Twitter dataset. The proposed method applied a mechanism for distributing different multi-weight polarities based on the number of detected polarity terms in the same sentiment. Therefore, if the number of detected terms increases, the weight polarity will increase due to the diversity

and orientation of polarity terms. As presented in [22,23], a set of positive and negative sentences were assigned based on Arabic tweets, where a hybrid strategy was proposed to combine different machine learning approaches. The Lexical-based classifier applied this method to label the training data.

As presented in [24], an automatic sentiment analysis based on supervised classification on the Arabic dataset was proposed. Most sentiment analysis methods verify the sentiments based on their polarity. When the negation term in the sentiment is detected immediately before the polarity term, the sentiment polarity is converted to its reverse polarity. The authors in this paper proposed a methodology for detecting negation terms even if they do not precede the polarity term. Another enhanced sentiment analysis mechanism for analyzing Arabic reviews was proposed in [25], where a lexicon-based analyzer was constructed to analyze polarity terms with different weights to increase efficiency.

Aspect-based sentiment analysis was proposed in [26], where a model was provided for estimating the polarity of user reviews. A lexicon was constructed for acquiring tweets from Twitter, and the dataset was preprocessed to remove any stop words and non-English words. The subjective sentences were processed using Senti-Word-Net to apply a score for each sentence. Based on the provided score, the aspect of each sentence was verified. Another recent research for applying aspect and content analysis of sentiments is presented in [27]. The authors provided a framework for collecting marketing and customer service information from reviews of different universities. The published or posted content of each university was classified into links, photos, videos, and statuses and then the frequency of each content was defined and scored.

The authors of [28] provided a framework for analyzing sentiment opinions during the COVID-19 outbreak. The dataset was collected, classified based on seven clusters, and categorized based on five annotators. The positive, negative, and neutral polarities were determined in each cluster, and the polarity score for each cluster was defined. Another methodology for analyzing sentiments during COVID-19 was proposed in [29]. The authors aimed to examine and measure the influencing factors that affect the orientation of people during the epidemic. The dataset was collected from different social media forums. Based on the proposed influencing factors, the dataset was classified, and the sentiments were analyzed to explore their polarities. The authors of [30] proposed another method of NLP for analyzing user sentiments during the COVID-19 epidemic. A framework was proposed to measure the performance of a set of relevant word terms from different aspects such as economy, social, and health to view the major orientation of the reviews that explained a neutral polarity orientation.

2.2. Machine Learning-Based Approach

Machine learning-based (ML) approaches provide powerful methods for analyzing polarity sentiments from different domains. The ML methods are adapted to learn from the input dataset and then provide the prediction from the hidden patterns. Learning and prediction are considered the two major steps in all ML algorithms. As proposed in Figure 1, ML has two major approaches for handling input data for training. These approaches are supervised and unsupervised. In the supervised approach, the input data must be trained first before applying the testing dataset, as the ML algorithms can perform the prediction based on previous experience. Therefore, different classification and regression algorithms are applied to measure the relationship between input and target features [31], and then the accuracy of the prediction is explored. The unsupervised approach depends on an unlabeled dataset where the patterns are discovered by performing complex tasks using clustering algorithms that can group the dataset and learn from unknown patterns. In addition to supervised and unsupervised approaches, the semi-supervised methods can perform the training and prediction with a less labeled dataset that is linked with the unlabeled dataset to produce the result.

Recent research methodologies for applying machine learning (ML) algorithms on sentiment analysis from social media datasets have been proposed. As presented in [32],

a sentiment analysis mechanism based on machine learning was applied to Russian and Kazakh languages. The applied dataset was categorized into positive, negative, and neutral. To increase their efficiency, different resampling techniques were used to resample the unbalanced datasets. Another enhanced mechanism for analyzing sentiments from the Twitter dataset was proposed in [33], where different classifiers of machine learning were adopted using the TF–IDF algorithm for extracting features. The sentiments were collected from multi-classes emotion data, and the accuracy was tested. Another sentiment analysis approach was presented in [34] for classifying tweets based on learning models. The research aimed to analyze a large number of social media tweets from Twitter using the Apache Spark model. The experiments were conducted to measure the time consumed using Apache Spark compared with other classifier models. As presented in [35], a machine learning-based approach was applied to analyzing Arabic sentiments. Different ML classifiers were used on the cleaned dataset to remove stop words and word elongation.

As shown in [36], ML classification algorithms were applied to multi-language datasets based on predefined Key Performance Indicators (KPIs). The idea was to group a set of opinions from the leaders of a company and then analyze the comments about the company and distribute these comments over the predefined KPIs. Analyzing and enhancing the accuracy of sentiment analysis can be proposed based on the semantic knowledge and content analysis of the sentiment polarities. This method is called aspect-based sentiment analysis. As presented in [37], aspect-based sentiment analysis was provided to identify sentiment polarities based on different attributes or aspects. The authors applied a framework for extracting aspects from Twitter datasets, and sentiment analysis was applied based on machine learning methods. One of the recent research methodologies for analyzing Arabic user opinions was presented in [38]. The research focused on analyzing Saudi citizens' and residents' opinions about the downloaded programs from Google Play and App Store.

Different machine learning classifiers were provided to measure the dataset's accuracy, classified into negative, positive, neutral, unique, and stem words. Another interactive methodology for analyzing Arabic Twitter sentiments was proposed in [39]. The authors of this research focused on a new topic related to detecting depression terms in Arabic sentiments from Twitter. The authors created three lexicons for storing depression terms and counted the number of tweets for each symptom. Different machine-learning algorithms were deployed to measure the accuracy of the results. Another interactive method for classifying sentiments is shown in [40], where teaching-learning-based optimizers were applied to a Twitter dataset. Different preprocessing steps were executed to remove stop words and symbols before applying four text-processing models. A feature selection algorithm was proposed to classify polarity features into positive and negative polarities, and finally, the results of the models were listed.

As proposed in [41], machine-learning-based algorithms have been applied to different polarity languages. The dataset was collected from Twitter, where each tweet was converted and counted to an integer. The integer was converted to its TF–IDF score value, and finally, the score was applied to ML classifiers for prediction. The authors of [42] proposed another aspect-based sentiment analysis methodology by providing an automatic annotation of datasets from YouTube songs. These songs were extracted and applied based on the number of views and reviews and then an aspect filtration was provided based on five aspect categories. The overall reviews were optimized again based on the predefined aspects to determine the most important aspect category.

Sentiment analysis of users' orientations is also conducted in health sectors to improve the users' feedback and provide deep insight into the provided services. As presented in [43], a sentiment analysis tool was provided for measuring people's attitudes in smart cities during the COVID-19 epidemic. The dataset was tested five times to measure the average accuracy of the classified instances. As shown in [44], sentiment analysis of user polarities based on natural language processing (NLP) was proposed to explore the attitudes of Gulf countries during COVID-19. The goal of the paper was to check whether there were mixed emotions among people or not. The dataset was extracted from Twitter API, and the

polarities were classified in different countries. The authors stated that most Gulf people's attitude toward the epidemic was neutral. In addition to previous research, the authors of [45] provided an approach based on semi-supervised machine learning that measured the analyses of datasets collected from different social networks about several epidemics. The first analysis clustered the data from Word-2-Vec and Fast-Text and then explored the sentiment orientation based on the applied classification algorithms. The authors of [46] measured the sentiment analysis of public health but from the financial aspect. The dataset was collected from financial news and then the reviews were grouped based on four polarity attributes to view the overall orientation. Furthermore, different classification algorithms were applied to learn from the polarity attributes and explore sentiment accuracy.

In addition to lexicon-based and ML-based methodologies, different deep learning and transfer learning methods can be applied to Arabic sentiment analysis. As presented in [47], a deep learning model for multitasking was applied for classifying Arabic sentiments. The research aimed to enhance the performance as the Arabic language has low resources and contains high morphological and linguistic terms. The authors proposed a long short-term memory (LSTM) deep learning model that explored the relationship between three and five sentiment polarities into a private layer. This layer contained an encoder for words to add flexibility to the features. The authors of [48] proposed a deep transfer-learning model for manipulating Arabic text. The Convolutional Neural Network (CNN) training was conducted to classify the sentiments based on a pre-annotated dataset. The authors collected different classes of the dataset and performed augmentation of data to enhance the accuracy. Another deep learning model for manipulating Arabic sentiments was presented in [49], where the data was collected based on three classes and then classified using the LSTM model to explore the results. One of the recent methods for extracting and detecting Arabic polarity text was proposed in [50], where transfer learning (TL) techniques were applied to aspect-based Arabic text. The authors proposed an architecture using the BERT model on the HAAD dataset to measure the approach's effectiveness, which showed high results.

Arabic is considered among the richest languages worldwide with many linguistic terms. Arabic sentiment analysis has not been studied as highly as other languages. The Arabic language suffers from a lack of high-quality terms and large-scale training data with the difficulty of manipulating ironic and slang expressions [47]. The Arabic language is considered an unstructured language with many inconsistencies in the spelling of terms and difficulties in identifying key features. Furthermore, using Arabic tweets in social networks cause many word elongations and repetition of terms to convey the user's feelings. This paper proposes a novel algorithm based on a forward fusion feature selection for sentiment analysis using different models such as Pearson, Chi^2, and RF, and then different ML classification models for each chunk of k features and accumulative features on the Arabic dataset to explore the winning model with high accuracy.

3. Comparative Analysis with Annotation Algorithms of Sentiment Analysis

The analysis of sentiments depends mainly on the extraction of subjective sentences from different social media streams. The customers and users reflect and express their attitudes and opinions towards different services and products using several languages and linguistics with a different number of classes. As presented in Tables 1 and 2, a comprehensive review of recent sentiment analysis researches is conducted. As explained, only positive and negative polarities will be predicted if two classes are applied for classification. The use of three classes for classification adds neutral sentiments where the number of positive and negative sentiments are equal or the overall orientation of the sentiment is fair. Arabic is considered one of the most languages in the world in the context of linguistics and terms that are difficult to identify and analyze. Two main annotation methods are applied for analyzing and predicting sentiments: machine learning (ML) and lexicon-based. The ML-based method applies different algorithms for measuring the accuracy of predicting the sentiments' orientations. In contrast, the lexicon method is based on building a corpus for storing Arabic dialects with their polarities or weight scores.

Table 1. Comparison of lexicon-based methods with their performance.

Ref	Dataset Size	No. of Classes	Dialect	Language	No. of Annotators	Performance
[17]	30K	Pos.–Neg.	Single	Arabic	2	77%
	8K					86%
[19]	16.9K	Pos.–Neg.	Single	Arabic	2	76.89%
	18.9K					76.48%
[21]	5k	Pos.–Neg.	Single	Arabic	2	Senti. Score
		Pos.–Neut.–Neg.			3	Senti. Score
[22]	100K	Pos.–Neut.–Neg.	Single	Arabic	3	66%
[23]	1.1K	Pos.–Neg.	Single	Arabic	2	84%
[24]	3.4K	Pos.–Neut.–Neg.	Single	Arabic	3	Senti. Score
[25]	6.3K	High Pos.–Pos. Neut. Neg.–High Neg.	Single	Arabic	5	Senti. Score
[26]	6K	Pos.–Neut.–Neg.	Single	English	3	Senti. Score
[27]	3.9K	Pos.–Neut.–Neg.	Single	English	3	Senti. Score
[28]	12.5K	High Pos.–Pos. Neut. Neg.–High Neg.	Single	English	5	46.4%
[29]	22.5K	Pos.–Neut.–Neg.	Single	English	3	86.2%

Table 2. Comparison of ML-based models with their performance.

Ref	Dataset Size	No. of Classes	Dialect	Language	No. of Annotators	Performance
[1]	20K	Pos.–Neut.–Neg.	Single	Arabic	3	79%
[5]	21K	Pos.–Neut.–Neg.	Single	Arabic	3	75.1%
[31]	24K	Pos.–Neut.–Neg.	Single	Turkish	3	91.57%
[32]	80.8K	Pos.–Neut.–Neg.	Multi	Russian	3	77%
	15.9K			Kazakh		
[33]	39.8K	Pos.–Neut.–Neg.	Single	English	3	89.92%
[34]	10K	Pos.–Neut.–Neg.	Single	Arabic	3	83%
[36]	685	Pos.–Neg.	Multi	Eng. Spa. Ita. Ger. Fre.	2	88.17%
[37]	6.7K	Pos.–Neut.–Neg.	Single	English	3	78.76%
[38]	8K	Pos.–Neut.–Neg.	Single	Arabic	3	78.46%
[39]	4.5K	Pos.–Neut.–Neg.	Single	Arabic	3	82.39%
[40]	14.6K	Pos.–Neg.	Single	English	2	76.9%
[41]	11K	Pos.–Neg.	Multi	Urdu	2	84%
				Roman Urdu		85%
				English		77%
[43]	0.5K	Pos.–Neut.–Neg.	Single	English	3	89.4%

4. Proposed Sentiment Analysis Methodology

As discussed in the previous sections, recent sentiment analysis mechanisms depend mainly on different aspects. Firstly, the applied language or dialect contains users' reviews and comments. Secondly, the data preprocessing mechanism for cleaning and adapting

datasets and corpus documents for sentiment analysis. Thirdly, the implemented lexicon-based or ML-based methods for manipulating datasets. Finally, the proposed methodology and algorithm for exploring and enhancing the analysis of user polarity is based on word terms and polarity terms.

The methodology of this paper is based on applying ML-based sentiment analysis algorithms on Arabic datasets from [1,51]. The Arabic language is a difficult language in the ramifications of its terminology as it contains many dialects and huge linguistic and morphological terms, which makes analyzing user sentiments a great challenge. The Arabic language has low resources and contains high morphological and linguistic terms [47]. Therefore, the analysis and classification of polarity terms is considered a challenging process. In addition, the Arabic sentiment analysis has not been studied at a level as high as other languages, such as English, Chinese, and French.

Furthermore, the user reviews that use the Arabic language to express their views and orientations can use multi-dialect in the same comment, which causes an additional overhead during the data preprocessing and analysis. As presented in Figure 2, the proposed mechanism depends on interactive multi-level processes for splitting the dataset into training and testing and then performing a 10-fold cross-validation for interchanging the *k* folds. The next step is data preprocessing, where several stages are applied to clean the polarity sentiments from unwanted terms and particles to increase the analysis efficiency. The next feature generation stage transforms the tweets into a set of feature vectors with an encoder that will be applied to the training and testing dataset. The vectorization of features is executed using the Term Frequency—Inverse Document Frequency (TF–IDF) mechanism that generates the features from the overall corpus of documents and sentiment analysis documents with their word and polarity terms. The generated features are selected using a filter method that divides the input features into different feature vectors related to the target feature. The correlation is executed using three methods: Person, Chi^2, and Random Forest (RF). The correlated features are sorted according to their relevance to the target features where the best *k* features are ranked and selected. The next *k* features are added until all the features are processed. The final stage is based on modeling the selected features using differing ML algorithms, where the best accuracy for each correlation method with the ML modeling algorithm is recorded.

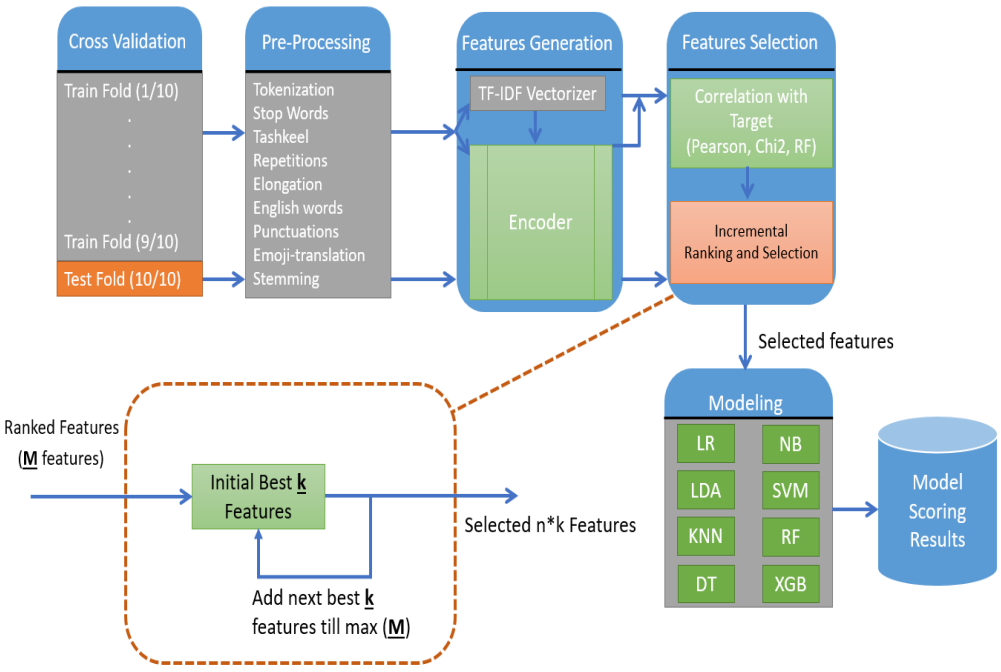

Figure 2. Proposed methodology of multi-stage feature generation and selection mechanism.

4.1. Cross Validation

The preprocessing stage should be in the correct position during the cross-validation process. Preprocessing steps are intended to be created using training data folds (preprocessing adaptors), and then the procedure is repeated using test data folds (using preprocessing adaptors for transformation). This ensures that the model is only exposed to preprocessed training data during the training phase. This approach aids in avoiding "data leakage," which occurs when the model is exposed to information from the test set during training, resulting in overfitting.

The cross-validation process is applied based on the *k*-fold mechanism. The main objective of *k*-fold cross-validation is to divide or split the dataset into a set of K groups [52]. Each group is treated as a validation unit to evaluate the overall model. In this paper, a 10-fold cross-validation on the dataset is executed randomly where $k = 10$. A number of $k - 1$ folds are used for dataset training while the remaining *i* fold is used for testing. On the next splitting process, another *i* fold is used for testing while the remaining $k - 1$ folds are used for dataset training. The 10-fold cross-validation continues until the *i* testing fold is applied on all splitting stages, where the final model is validated on each *i* testing fold. Equation (1) summarizes the overall 10-fold process as follows:

$$\forall\, i \ \subseteq k \ \ni i = 1 \,\&\, k - 1 < 10 \tag{1}$$

4.2. Data Preprocessing

After performing cross-validation on the sentiment analysis dataset, data preprocessing is executed to eliminate and clean the sentiments and their polarity terms from incorrect or unwanted terms that may affect the accuracy of the analysis process. As presented in Figure 3, data preprocessing depends on a set of sequential steps for removing inconsistent terms from the sentences. The stemming process is applied to the overall dataset to reduce the polarity term length so that the polarity term returns to its root. Stop words and tokenization are eliminated from the sentiment analysis dataset, where the stop words are terms that do not affect the sentence's overall meaning. For example, the stop word terms such as "أو–في–على–لا–و–قبل–بعد" that mean "or–in–on–not–and–before–after", respectively, are removed from the preprocessed data. Tokenization is separating polarity terms using a space or a unique character to be analyzed more efficiently during the machine learning mechanism. For example, the sentence "لم يقدم التطبيق أي خدمة إضافية!" that means "The application does not provide any additional service!" is separated into individual terms to increase the machine learning ability to understand the whole sentiment. Due to the use of the Arabic dataset in our paper, some words that contain "Tashkeel" are also eliminated. The term "Tashkeel" means a set of special characters added to the formation of the words to change the word pronunciation. For example, the sentence "التَطبيقُ جَيِد لَكِنَ الخِدْمَاتِ بَطيئَةٌ جِدًاْ" that means "The application is good but the services are very slow" contains many special characters to set the word terms. After removing the "Tashkeel" characters, the new sentence becomes "التطبيق جيد لكن الخدمات بطيئة جدا". The English words, punctuation, and repetition of the terms are also removed during data preprocessing to reduce the sentence length during the sentiment analysis process. Emojis are also removed from the sentences as they can contain different ironical or emotional terms that can affect the orientation of the overall sentences from positive to negative and vice versa. Finally, the word elongation is also removed from the sentiments to eliminate any repetition of letters. For example, the polarity term "جمييييل" is processed to be "جميل" that means "nice" or "beautiful".

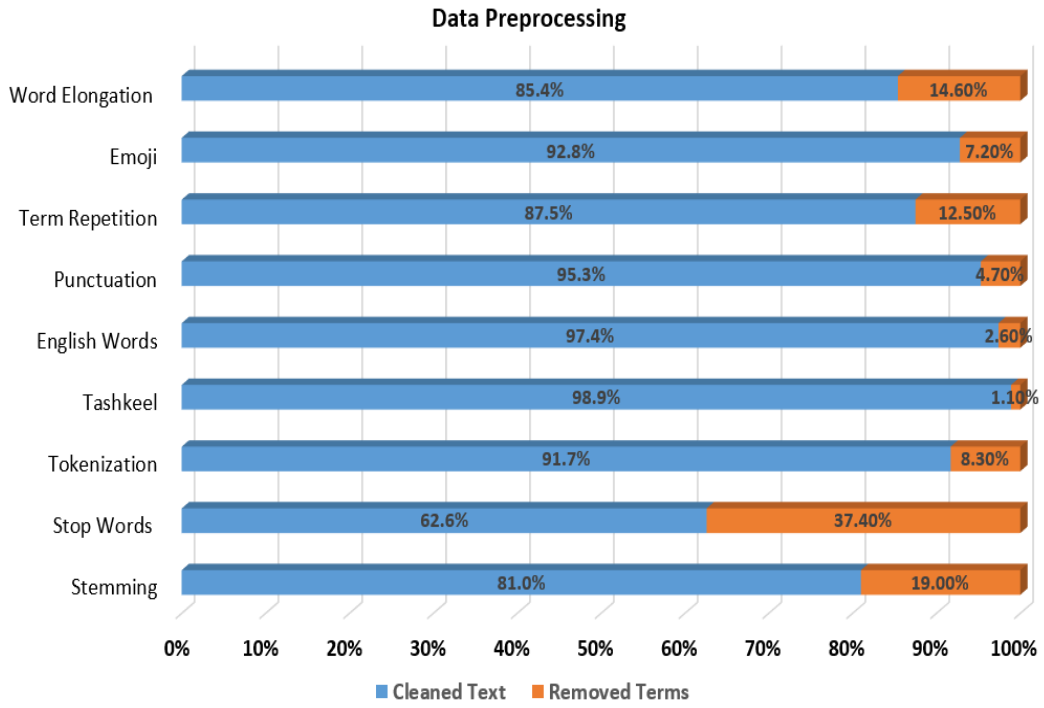

Figure 3. Data preprocessing levels on sentiment analysis documents.

The foundation of the classification models is based on data segmentation. The imbalanced dataset uses one chunk of data that contains a large portion, which is called the majority class, while the remaining chunk of data represents the minority class. As a result, an imbalanced dataset is one in which one class has a higher number of occurrences or sentences than the other class. Equal or almost equal numbers of occurrences or sentences from each class are presented in the balanced dataset.

As presented in Table 3, the imbalanced dataset contains 16.7K sentiments that has approximately positive sentiments as 17%, negative sentiments as 16.5%, and neutral sentiments as 66.5%. The dataset was balanced into which two classes of dataset, positive and negative, are selected with 5451 sentiments.

Table 3. Balanced dataset with binary classification.

Dataset \ Polarity	Positive	Negative	Neutral	Total
Imbalanced	2843	2751	11,129	16,723
%	17%	16.5%	66.5%	100%
Balanced with 2 Classes	2725	2726	-	5451
%	50%	50%	-	100%

4.3. Feature Generation

In this research, the tweets are manually annotated using a unigram model that offers a reasonable coverage degree for the dataset. In order to extract the most important features from the training dataset, data preprocessing and feature generation are executed to convert the tweets into a feature vector. An encoder is generated from the vectorization process of the training dataset to be applied to the testing dataset. The generated encoder converts the tweets into a set of feature vectors that are applied on the training and testing datasets. Text and tweets are applied as vectors using the TF–IDF technique that converts the given text into finite feature vectors. The term frequency (TF) computes the number of times the selected term is repeated in a given document. In contrast, the inverse document frequency (IDF) computes the number of times the selected term is repeated in the overall dataset or corpus.

The TF–IDF method has a linear computational complexity regarding the number of text lines and words per line. In contrast, the RF feature selection algorithm has a computational complexity of $O(n\ estimators \times m \times \log(n))$, where n is the number of samples, m is the number of features, and log is the base-2 logarithm. Finally, the NB classifier is a simple and computationally efficient algorithm with $O(n \times d)$ computational complexity, where n is the number of samples and d is the number of features.

The corpus data contains different sentiment analysis documents with different word and polarity terms. The documents are collected from Twitter to analyze users' orientation and attitudes based on their positive or negative polarity. As presented in Equations (2)–(4), the overall vectorization of the dataset is explained:

$$TF\ (p_t\ ,\ s_d) = \frac{\sum_{i=1}^{n} p_{ti}\ \in\ w_t}{\sum_{j=1}^{m} w_{tj}\ \in\ s_d} \tag{2}$$

$$DF\ (p_t\ ,\ c_d) = log\left(\frac{N}{Count\ (s_d \in\ c_d\ :\ p_t \in\ s_d)}\right) \tag{3}$$

$$TF - IDF\ (p_t\ ,\ s_d\ ,\ \ c_d) = TF\ (p_t\ ,\ s_d) \times IDF\ (p_t\ ,\ c_d) \tag{4}$$

where:
w_t : word terms.
p_t : polarity terms.
s_d : sentiment analysis documents.
c_d : corpus data for all documents.

The corpus data c_d contain different documents of sentiment analysis s_d from different domains and sources that reflect the users' opinions about different services. Each sentiment analysis document s_d contains a large number of sentences with word terms w_t that contain polarity terms p_t that explores users' orientations. As shown in Equation (4), $TF - IDF$ is calculated by measuring the resulting score of the multiplication between TF and IDF. The higher the resulting score, the more relevant the polarity term in the sentiment analysis documents s_d.

4.4. Feature Selection

The feature selection stage aims to reduce the input parameters or features to predict the target values efficiently. In addition, some predictive models contain many variables that can affect the efficiency of the memory or can reduce the system performance due to the incompatibility between the input and the target features. Supervised and unsupervised methods are the main key features for predicting target features. The selection process in unsupervised methods removes redundant features and eliminates the target variable, while supervised methods focus on the target features by removing any insignificant input features.

Other feature selection methods, such as wrapper and filter methods, can be applied by evaluating the model performance on the corpus data c_d and sentiment analysis documents. Regarding the wrapper method for feature selection, the input features are divided into different subsets. The method applies several models on the subsets to select the best model that achieves the highest performance. The filter method for feature selection applies several statistical techniques to estimate the relationship between input and target features. In addition, the filter method scores each resulting value between the input and the output features and then filters the best models based on the recorded scores.

This paper applied the filter method for feature selection by dividing the input features into a different subset of feature vectors and then selecting the subsets that are highly associated or related to the target features. The correlation and selection of the subsets are applied using three feature importance methods: Pearson correlation, Chi^2, and Random Forest (RF).

The Pearson method explores the correlation score between the input and target features, where the score ranges from -1 to +1. If the correlation score is close to +1, then

the relationship to the target is high and vice versa. Firstly, the covariance between each word term feature w_t and the target feature of the expected polarity term p_t is calculated, and then the Person correlation is measured by dividing the covariance value in Equation (5) by the multiplication of the standard deviation of both word term feature w_t and polarity term features p_t as explained in Equation (6).

$$Cov\ (w_t,\ p_t) = \frac{1}{n} \times \sum_{i=1}^{n} ((w_{ti} - \overline{w}_t) \times (p_{ti} - \overline{p}_t)\ ) \tag{5}$$

where:
w_{ti}: each input word term feature in the vector.
$\overline{w}_t$: the mean of the overall word term features.
p_{ti}: each target polarity term feature.
$\overline{p}_t$: the mean of the overall polarity term features.
n: the length of the word terms and polarity terms.

$$PC\ (w_t,\ p_t) = \frac{Cov\ (w_t,\ p_t)}{sd_{wt} \times sd_{pt}} \tag{6}$$

where:
sd_{wt}: the standard deviation of word terms.
sd_{pt}: the standard deviation of polarity terms.

The Chi^2 method for feature selection is used to measure the independence degree between the observed values of the word term features w_t and the expected values of the polarity term features p_t. Therefore, the Chi^2 method selects the features of both word terms and polarity terms that are highly correlated as shown in Equation (7).

$$Chi^2 = \sum \frac{(w_{ti} - p_{ti}\ )^2}{p_{ti}} \tag{7}$$

The RF is considered a predictive method with high performance and low overfitting value. In addition, the RF combines both filter and wrapper methods and contains several decision trees. Each tree is based on a random extraction of the word term features w_t and a random extraction of the polarity term features p_t. Each tree cannot trace all the word term features and polarity term features to reduce the overfitting.

5. FFF-SA Algorithm

The filter methods for feature selection measure one input feature at a time with the target feature. Therefore, there is no interaction between the input features. To overcome this issue, this paper proposed an algorithm called the innovative forward fusion-based for feature selection (I-FFF). This algorithm performs a forward chain feature selection by calculating the input feature importance with the target value using Pearson correlation, Chi^2, and RF. The Pearson method is applied to numerical variables, while the Chi^2 method is applied to categorical variables. The RF method is applied to both numerical and categorical variables. The FFF-SA algorithm is based on three main stages for selecting the best correlation between the input and the target features.

The first stage is based on sorting the feature vectors based on their importance and association to the target. The second stage starts by selecting the best k batch of features and then measures its score during the tuning and validation of data using different modeling approaches of machine learning (ML). The third stage adds a new k batch of features one at a time until all batches of features are added.

After performing the feature selection for all input features of word terms w_t and target features of polarity terms p_t, different machine learning models are implemented to predict the labels of the input word terms. In addition, the trained dataset is tuned to maximize the models' performance without overfitting data. The tuning process is based on hyper-parameters that can control the trained dataset. In this modeling stage, eight machine-

learning algorithms are applied, where some algorithms have unique hyper-parameters while others have similar hyper-parameters. These algorithms are Logistic Regression (LR) [53], Linear Discriminant Analysis (LDA) [54], K-Nearest Neighbor (KNN) [55], Decision Tree (DT) [56], Naïve Bayes (NB) [57], Support Vector Machine (SVM) [58], Random Forest (RF) [59], and Gradient Boost (XGB) [60]. The overall FFF-SA algorithm is explained as follows:

The proposed FFF-SA algorithm starts by defining the full feature set FF, the maximum size of features M, the k features for each experiment, and the n number of execution times during the experiment. Each experiment is conducted using the three correlation features, Pearson, Chi^2, and RF features, and by dividing the input features to a different subset of feature vectors and then selecting the subsets that are highly associated or related to the target features. The first correlation feature of Pearson is initialized, where the first k features is applied on the experiment, and then the eight ML algorithms are used as models to train the selected k features and then score the accuracy using the testing fold to measure the similarity and correlation between the training and target features. Each time, the first ML algorithm selects the k features and then registers the score. The next $Accum_F + k$ features are added to the previous features as accumulative features to model and score the accuracy on the accumulative features. The process continues until the M features are reached, which represents the maximum size of the features. The second ML algorithm repeats the experiments for the k features and then registers the score at each accumulative feature until all ML algorithms record their accuracy results for the Pearson correlation method. The next experiments are conducted with the Chi^2 and RF correlation methods to score and record the accuracy.

FFF-SA Algorithm	
1	$Input : FF\ ,\ M\ ,\ k\ ,\ n$
2	$Output : Sentiment\ Analysis\ Accuracy$
3	$Initialize\ n\ \leftarrow 1$
4	$FOR\ EACH\ CorrF\ in\ [Pearson\ ,\ Chi^2,\ RF]\ DO$
5	$FF_{Cor} = CorrF_i\ (FF)$
6	$FOR\ n = 1\ to\ t\ DO$
7	$Chunk_F = k$
8	$Accum_F = n \times k$
9	$WHILE\ (Accum_F < M)\ DO$
10	$F_{SELECT} = Extract\ Ranked\ Set\ (FF_{Cor}, Accum_F)$
11	$FOR\ EACH\ Alg_i\ in\ [\ LR,\ LDA,\ KNN,\ DT,\ NB,\ SVM,\ RF,\ XGB\ ]$
12	$Model = Train\ (Alg_i\ ,\ F_{SELECT}\ [\ Train\ k\ fold\ ])$
13	$Score = Evaluate\ (\ Model\ ,\ F_{SELECT}\ [\ Test\ k\ fold\ ]\ )$
14	$Register_{Score} = (Alg_i\ ,\ Accum_F\ ,\ Score)$
15	$Accum_F = Accum_F + k$

6. Experimental Results

The experimental results based on the proposed FFF-SA algorithm are tested, where the three correlation coefficient methods, Pearson, Chi^2, and RF, are used as primary coefficients to measure the accuracy and efficiency of sentiment analysis classification. The hyper-parameters are parameters the user sets rather than learns from the data. Some common hyper-parameters used during the feature selection include the number of selected features, the criteria for selecting features (RF, Chi^2, and Person), the feature selection threshold, and finally, the number of iterations for forward feature selection.

As explained before, the collected dataset is based on the Arabic language that contains high and complex morphological sentiments and terms that can affect the performance of the experiments. The Arabic language is rich in many linguistic terms that indicate more than one meaning and orientation. In order to obtain a higher degree of accuracy, the sentiment analysis process depends on many stages to effectively purify and analyze the sentiment analysis documents and terms. In addition, splitting and training the data depends on several steps to find the relationship between the input feature vectors and the target features. The following sections explore the applied feature selection correlation mechanisms with different ML models.

6.1. Experiment 1: Feature Importance Using RF

In this experiment, the RF feature selection method is applied to the eight ML algorithms: LR, LDA, KNN, DT, NB, SVM, RF, and XGB. For each conducted ML algorithm, the experiment starts with the first $k = 100$ features where the model is evaluated and the score accuracy is recorded. Based on the FFF-SA algorithm, the experiment is continued by increasing an additional $k + 1 = 100$ to the previous k fold to obtain $k = 200$ as an accumulative feature and then the model is executed again to find the highest accuracy. As explained in Figure 4, the highest accuracy is recorded on the accumulative feature $k = 2400$ with the NB algorithm that achieved an accuracy of 84.4%. The second highest accuracy is recorded on the accumulative feature $k = 2300$ with the same algorithm of NB that achieved an accuracy of 84.17%.

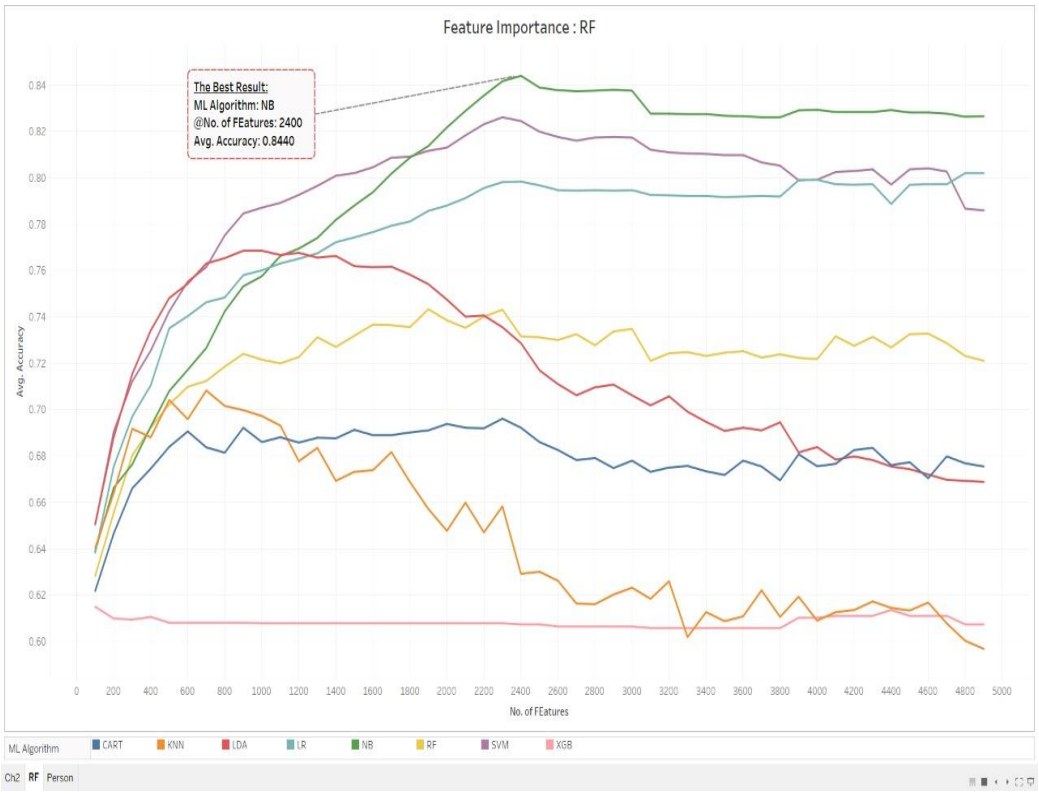

Figure 4. Accuracy of RF feature selection method with ML models.

As stated, the four ML algorithms, NB, LR, SVM, and LDA, start the experiments with a linear increase but the results of the LDA algorithm start decreasing at the accumulative feature $k = 1800$. The NB, LR, and SVM score the best results when compared to the remaining ML algorithms where the highest scored accuracy is recorded with NB algorithm.

6.2. Experiment 2: Feature Importance Using Chi²

In this experiment, the Chi² feature selection method is applied to the eight ML algorithms. The experiments are conducted again based on the FFF-SA algorithm where each experiment starts by modeling the ML algorithm with the accumulative set of features k. As explained in Figure 5, the experimental results show dispersed results from most ML algorithms except the NB and SVM algorithms that achieved the best results. The NB algorithm recorded the best accuracy of 83.76% with the accumulative feature $k = 2300$. With the accumulative feature $k = 2400$, the accuracy decreased very slightly and recorded 83.62%. The remaining accumulative k features continued in slight decreases, forming a straight line to the end of the accumulative feature with $k = 4900$ that recorded 82.91%. The second-best results are achieved on the SVM algorithm with the accumulative feature $k = 2200$ that recorded an accuracy of 81.49%. On the accumulative feature $s\ k = 2100$ and $k = 2300$, the SVM recorded accuracies of 81.21% and 80.98%, respectively.

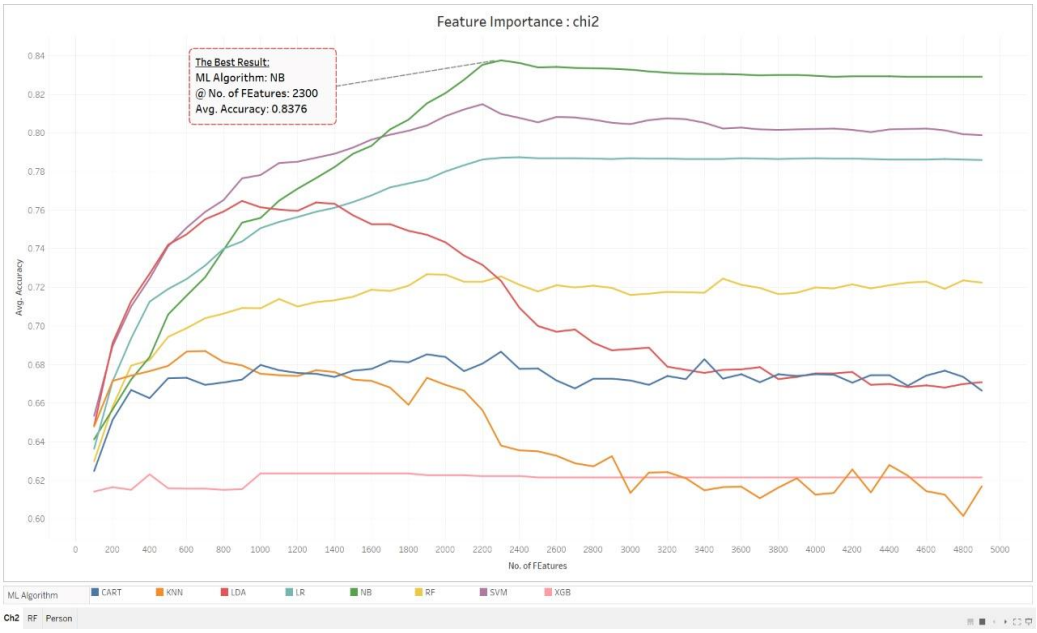

Figure 5. Accuracy of Chi² feature selection method with ML models.

6.3. Experiment 3: Feature Importance Using Pearson Correlation

As presented in Figure 6, the experimental results are executed again using the Pearson coefficient for measuring the correlation between the input features and the target features. As explained, all conducted ML algorithms achieved low results, and both LR and SVM algorithms achieved an accuracy of 58.73% on the accumulative feature $k = 700$. Starting from $k = 700$ to the end of the accumulative features where $k = 4900$, the accuracy remains the same score, forming a straight line. The LDA algorithm achieved the second-best results of 58.51% with the accumulative feature $k = 700$, and the stated results continued in a straight line to the end of the accumulative features where $k = 4900$. As noticed from the figure, most ML algorithms start increasing with k =100 until the accumulative feature $k = 700$, where the results remain without change to the end of the features.

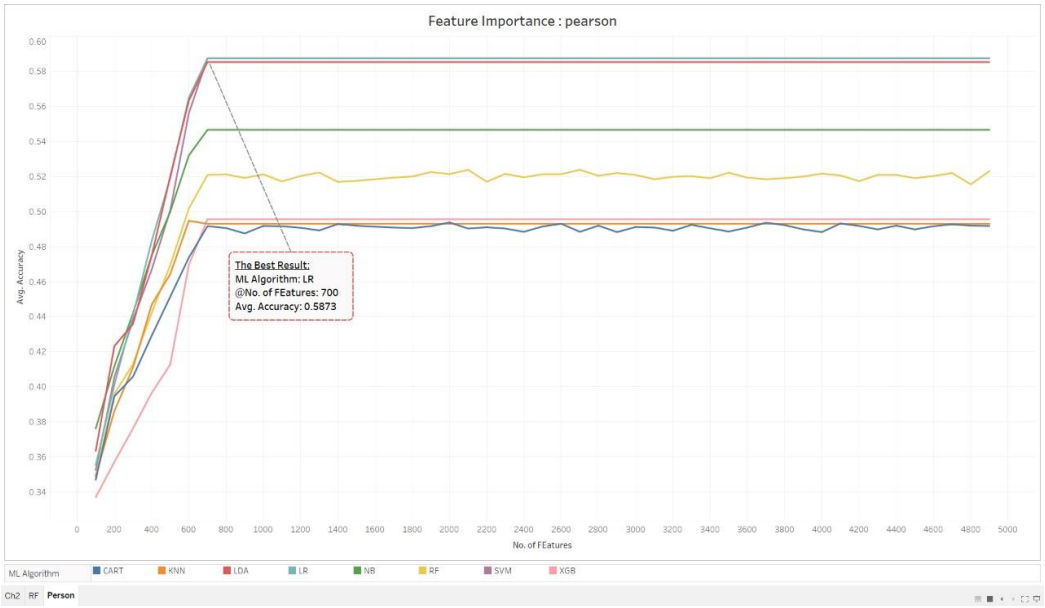

Figure 6. Accuracy of Pearson feature selection method with ML models.

6.4. Feature Importance Comparison Analysis

Based on the previous experimental results, the FFF-SA algorithm is based on a forward chain of processes for performing a feature selection on the split features. As mentioned before, the sentiment analysis documents and sentiments are divided into k features for training and testing. The correlation methods are applied with different ML models for measuring and scoring each set of k features. On the subsequent stages, the features are accumulated with additional k features, while the ML models perform the scoring at each step. Table 4 presents a summary of the performance analysis for each feature importance method with its corresponding ML models. The table showed that the highest accuracy on all experiments and feature importance stages recorded 84.4% for the RF method with the NB algorithm. The second-best accuracy with the NB algorithm was 83.8% for the Chi^2 method. The SVM algorithm also achieved good results, with 82.6% and 81.5% on RF and Chi^2, respectively. The provided accuracy provides promising results based on the challenge of pure Arabic sentiment analysis documents and terms collected from different domains [1,51].

Table 4. Analysis of accuracy for different feature importance methods with ML models.

RF							
NB	SVM	LR	LDA	RF	KNN	CART	XGB
84.4%	82.6%	80.2%	76.9%	74.3%	70.8%	69.6%	61.6%
Chi^2							
NB	SVM	LR	LDA	RF	KNN	CART	XGB
83.3%	81.5%	78.7%	76.8%	72.7%	69%	68.7%	62.5%
Pearson							
NB	SVM	LR	LDA	RF	KNN	CART	XGB
54.6%	58.7%	58.7%	58.5%	52.4%	49.5%	49.4%	49.6%

In Table 5, the number of features and experiments that are conducted on the dataset with different feature selection methods are explained. The table shows more than 3 million extracted features during the testing of data with about 4423 experiments on all feature selection methods.

Table 5. Analysis for the number of features and experiments.

Experiment / Feature Selection	No of Features	No of Experiments
RF	1,073,700	1455
Chi^2	1,012,800	1596
Pearson	980,000	1372
Total	3,066,500	4423

In Table 6, the top 10 accuracies for the experimental results that are executed on all feature importance methods with ML models are provided. The explanation of these results was intended to study the efficiency for feature importance models in all sentiment analysis documents and terms. For each feature $k = 100$, the orientation of polarities were mapped into two classes whether they were positive or negative. The feature importance method was applied to score the accuracy for detecting positive or negative polarities using different ML models. After scoring the first batch of k features, an additional $k + 1$ feature is added to form an accumulative feature where the FFF-SA algorithm is executed again. As stated in the table, the top 10 accuracies are recorded with the NB model with both RF and Chi^2 feature importance methods. The best seven results are recorded with the RF method with a different number of accumulative k features. Based on these experiments, it is clear that the NB model with the RF feature importance method scored the best results on Arabic sentiment analysis terms.

Table 6. Top 10 accuracies for feature importance methods.

No.	Vectorization	Feature Importance	No. of Features	Classes	ML Model	Accuracy
1	TF–IDF: ngram_range (1,1)	RF	2400	Pos.–Neg.	NB	84.40%
2	TF–IDF: ngram_range (1,1)	RF	2300	Pos.–Neg.	NB	84.17%
3	TF–IDF: ngram_range (1,1)	RF	2500	Pos.–Neg.	NB	83.89%
4	TF–IDF: ngram_range (1,1)	RF	2900	Pos.–Neg.	NB	83.81%
5	TF–IDF: ngram_range (1,1)	RF	2600	Pos.–Neg.	NB	83.78%
6	TF–IDF: ngram_range (1,1)	RF	2800	Pos.–Neg.	NB	83.76%
7	TF–IDF: ngram_range (1,1)	RF	3000	Pos.–Neg.	NB	83.76%
8	TF–IDF: ngram_range (1,1)	Chi^2	2300	Pos.–Neg.	NB	83.76%
9	TF–IDF: ngram_range (1,1)	RF	2700	Pos.–Neg.	NB	83.73%
10	TF–IDF: ngram_range (1,1)	Chi^2	2400	Pos.–Neg.	NB	83.62%

The conducted results in this paper showed promising results on Arabic sentiment analysis documents and terms that are considered a challenge in data preprocessing, feature extraction, and identification of sentiment polarities. In addition, the Arabic language contains several linguistic terms and colloquial terminologies that are difficult to preprocess and analyze. Figure 7 shows the confusion matrix measured for the last experiments for the final winning model.

In Table 7, a comparison is performed with recent methodologies and frameworks conducted on the Arabic dataset to measure the performance of each model and the annotated algorithm. Furthermore, the research methodologies are applied using ML algorithms or lexicon-based that have different methods for processing and analyzing sentiment polarities with binary classification. As shown, the proposed FFF-SA shows high results compared to related papers.

		Predicted	
		Pos.	Neg.
Actual	Pos.	**578**	**112**
	Neg.	**107**	**583**

Figure 7. Confusion Matrix for the final winning model.

Table 7. Comparison with recent sentiment analysis performance.

Ref	Year	No. of Classes	Dialect	Language	Annotation Algorithm	Performance
[17]	2020	Pos.–Neg.	Single	Arabic	Lexicon-based	77%
		Pos.–Neg.	Single	Arabic	Lexicon-based	86%
[19]	2022	Pos.–Neg.	Single	Arabic	Lexicon-based	76.48%
[21]	2023	Pos.–Neg.	Single	Arabic	Lexicon-based	Senti. Score
[23]	2016	Pos.–Neg.	Single	Arabic	Lexicon-based	84%
[40]	2021	Pos.–Neg.	Single	Arabic	ML-based	76.9%
Proposed FFF-SA Algorithm		Pos.–Neg.	Single	Arabic	ML-based	84.4%

7. Conclusions and Future Works

Due to the continuous and increasing use of social networks and E-commerce sites, many platforms depend on the analysis of user opinions to improve the provided services and measure customer satisfaction. One of sentiment analysis's most common problems is the language customers use to express their opinions. The Arabic language is considered one of the most difficult languages in the world because it contains complex linguistic terms and many different dialects that may be used in the same comment or review, making the analysis process more difficult. This paper provides a clear view of recent sentiment analysis approaches and algorithms that depend mainly on ML and lexicon approaches for storing and analyzing a dataset. A comparison of recent research strategies is also provided to compare different methodologies with applied language, including the Arabic language, and their achieved performance. An advanced methodology is proposed using cross-validation that divides the sentiment analysis documents and terms into *k*-folds for training and testing. Data vectorization is applied using the TF–IDF algorithm for counting polarity terms, and an encoder is generated from the training dataset to be applied to the testing dataset. Furthermore, the paper provided an algorithm called FFF-SA based on a forward filter of feature selection that measures and scores the accuracy for each *k*-chuck feature and the following accumulative features. The scoring is processed by executing three feature importance methods, Pearson, Chi^2, and RF, with eight ML models where each feature importance is executed with each ML model. Each experiment measures and scores the recorded accuracy for each *k*-chunk feature and then adds accumulative feature to measure the accuracy again. The results proved that the best accuracy is recorded with RF feature importance with the NB model. Future research directions will be directed to apply the same methodology and algorithm on additional Arabic datasets and apply deep learning models to measure the performance and accuracy.

Author Contributions: Data curation, A.M.M.; formal analysis, A.M.M. and M.E.; investigation, A.M.M., M.A. (Meshrif Alruily) and A.A.; supervision, M.A. (Meshrif Alruily) and A.A.; writing—original draft, M.E. and M.A. (Meeaad Aljasir); writing—review and editing, A.M.M., M.E., M.A. (Meshrif Alruily), and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Furnished on request.

Acknowledgments: The authors acknowledge the Deanship of Scientific research at Jouf University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alamro, H.; Alshehri, M.; Alharbi, B.; Khayyat, Z.; Kalkatawi, M.; Jaber, I.; Zhang, X. Overview of the Arabic Sentiment Analysis 2021 competition at KAUST. *King Abdullah Univ. Sci. Technol.* **2021**, *10754*, 1–9. Available online: http://hdl.handle.net/10754/672089 (accessed on 15 November 2022).
2. Zirikly, A.; Diab, M. Named Entity Recognition for Arabic Social Media. In Proceedings of the1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, CO, USA, 5 June 2015. [CrossRef]
3. Alruily, M. Classification of Arabic Tweets: A Review. *Electronics* **2021**, *10*, 1143. [CrossRef]
4. Oueslati, O.; Cambria, E.; HajHmida, M.; Ounelli, H. A Review of Sentiment Analysis Research in Arabic Language. *Future Gener. Comput. Syst. Elsevier* **2020**, *112*, 408–430. [CrossRef]
5. Hassan, S.; Mubarak, H.; Abdelali, A.; Darwish, K. ASAD: Arabic Social Media Analytics and Understanding. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Kiev, Ukraine, 19–23 April 2021. [CrossRef]
6. Alomari, K.; ElSherif, H.; Shaalan, K. Arabic Tweets Sentimental Analysis Using Machine Learning. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Arras, France, 27–30 June 2017. [CrossRef]
7. Ansari, M.; Aziz, M.; Siddiqui, M.; Mehra, H.; Singh, K. Analysis of Political Sentiment Orientations on Twitter. *Procedia Comput. Sci. Elsevier* **2020**, *167*, 1821–1828. [CrossRef]
8. Vidya, N.; Fanany, M.; Budi, I. Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. *Procedia Comput. Sci. Elsevier* **2015**, *72*, 519–526. [CrossRef]
9. Adilah, M.; Supendar, H.; Ningsih, R.; Muryani, S.; Solecha, K. Sentiment Analysis of Online Transportation Service Using the Naïve Bayes Methods. *J. Phys.* **2020**, *1641*, 012093. [CrossRef]
10. Bakliwal, A.; Foster, J.; van der Puil, J.; O'Brien, R.; Tounsi, L.; Hughes, M. Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. In Proceedings of the Workshop on Language in Social Media, Atlanta, GA, USA, 13 June 2013; Available online: https://aclanthology.org/W13-1106 (accessed on 1 January 2023).
11. Rao, A.; Kanade, V.; Motarwar, C.; Girme, S. Election Result Prediction Using Twitter Analysis. In Proceedings of the International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 19 January 2017. [CrossRef]
12. Patel, R.; Passi, K. Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT* **2020**, *1*, 218–239. [CrossRef]
13. Zhang, X.; Yang, Q.; Albaradei, S.; Lyu, X.; Alamro, H.; Salhi, A.; Ma, C.; Alshehri, M.; Jaber, I.; Tifratene, F.; et al. Rise and Fall of the Global Conversation and Shifting Sentiments during the COVID-19 Pandemic. *Humanit. Soc. Sci. Commun. Nat.* **2021**, *8*, 120. [CrossRef]
14. Wang, Y.; Guo, J.; Yuan, C.; Li, B. Sentiment Analysis of Twitter Data. *Appl. Sci.* **2022**, *12*, 1775. [CrossRef]
15. Gutierrez, E.; Karwowski, W.; Fiok, K.; Davahli, M.; Liciaga, T.; Ahram, T. Analysis of Human Behavior by Mining Textual Data: Current Research Topics and Analytical Techniques. *Symmetry* **2021**, *13*, 1276. [CrossRef]
16. Li, S.; Liu, F.; Zhang, Y.; Zhu, B.; Zhu, H.; Yu, Z. Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. *Mathematics* **2022**, *10*, 3554. [CrossRef]
17. Kwaik, K.; Saad, M.; Chatzikyriakidis, S.; Dobnik, S.; Johansson, R. An Arabic Tweets Sentiment Analysis Dataset (ATSAD) Using Distant Supervision and Self-Training. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, Marseille, France, 12 May 2020. Available online: https://aclanthology.org/2020.osact-1.1 (accessed on 22 November 2022).
18. Li, Q.; Li, Z.; Du, Y.; Fan, Y.; Chen, X. A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis. *Appl. Sci.* **2022**, *12*, 10236. [CrossRef]
19. Chennafi, M.; Bedlaoui, H.; Bedlaoui, A.; Al-qaness, M. Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers. *Knowledge* **2022**, *2*, 388–401. [CrossRef]
20. Alwakid, G.; Osman, T.; El Haj, M.; Alanazi, S.; Humayun, M.; Us Sama, N. MULDASA: Multifactor Lexical Sentiment Analysis of Social-Media Content in Nonstandard Arabic Social Media. *Appl. Sci.* **2022**, *12*, 3806. [CrossRef]
21. Mostafa, A. Enhanced Sentiment Analysis Algorithms for Multi-Weight Polarity Selection on Twitter Dataset. *Intell. Autom. Soft Comput.* **2023**, *35*, 1015–1034. [CrossRef]
22. Alharbi, B.; Alamro, H.; Alshehri, M.; Khayyat, Z.; Kalkatawi, M.; Jaber, I.; Zhang, X. ASAD: A Twitter-Based Benchmark Arabic Sentiment Analysis Dataset. *arXiv* **2022**, arXiv:2011.00578. [CrossRef]

23. Aldayel, H.; Azmi, A. Arabic Tweets Sentiment Analysis—A Hybrid Scheme. *J. Inf. Sci.* **2016**, *42*, 782–797. [CrossRef]
24. Mostafa, A. An Automatic Lexicon with Exceptional-Negation Algorithm for Arabic Sentiments Using Supervised Classification. *J. Theor. Appl. Inf. Technol.* **2017**, *95*, 3662–3671. Available online: http://www.jatit.org/volumes/Vol95No15/25Vol95No15.pdf (accessed on 1 January 2023).
25. Mostafa, A. Advanced Automatic Lexicon with Sentiment Analysis Algorithms for Arabic Reviews. *Am. J. Appl. Sci.* **2017**, *14*, 754–765. [CrossRef]
26. Banjar, A.; Ahmed, Z.; Daud, A.; Abbasi, R.; Dawood, H. Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter. *Comput. Mater. Contin.* **2021**, *67*, 2203–2225. [CrossRef]
27. Mehmood, S.; Ahmad, I.; Khan, M.; Khan, F.; Whangbo, T. Sentiment Analysis in Social Media for Competitive Environment using Content Analysis. *Comput. Mater. Contin.* **2022**, *71*, 5603–5618. [CrossRef]
28. Ibrahim, A.; Hassaballah, M.; Ali, A.; Nam, Y.; Ibrahim, I. COVID19 Outbreak: A Hierarchical Framework for User Sentiment Analysis. *Comput. Mater. Contin.* **2022**, *70*, 2507–2524. [CrossRef]
29. Oglah, M.; Baniata, A.; Asghar, S. Sentiment Analytics: Extraction of Challenging Influencing Factors from COVID-19 Pandemics. *Intell. Autom. Soft Comput.* **2021**, *30*, 821–836. [CrossRef]
30. Abdukhamidov, E.; Juraev, F.; Abuhamad, M.; El-Sappagh, S.; AbuHmed, T. Sentiment Analysis of Users' Reactions on Social Media during the Pandemic. *Electronics* **2022**, *11*, 1648. [CrossRef]
31. Deniz, E.; Deniz, E.; Cosar, M. Multi-Label Classification of e-Commerce Customer Reviews via Machine Learning. *Axioms* **2022**, *11*, 436. [CrossRef]
32. Mutanov, G.; Karyukin, V.; Mamykova, Z. Multi-Class Sentiment Analysis of Social Media Data with Machine Learning Algorithms. *Comput. Mater. Contin.* **2021**, *69*, 913–930. [CrossRef]
33. Saranya, S.; Usha, G. A Machine Learning-Based Technique with Intelligent Word-Net Lemmatize for Twitter Sentiment Analysis. *Intell. Autom. Soft Comput.* **2023**, *36*, 339–352. [CrossRef]
34. Iqbal, M.; Latha, K. A Parallel Approach for Sentiment Analysis on Social Networks Using Spark. *Intell. Autom. Soft Comput.* **2023**, *35*, 1831–1842. [CrossRef]
35. Hnaif, A.; Kanan, E.; Kanan, T. Sentiment Analysis for Arabic Social Media News Polarity. *Intell. Autom. Soft Comput.* **2021**, *28*, 107–119. [CrossRef]
36. Grande-Ramírez, J.; Roldán-Reyes, E.; Aguilar-Lasserre, A.; Juárez-Martínez, U. Integration of Sentiment Analysis of Social Media in the Strategic Planning Process to Generate the Balanced Scorecard. *Appl. Sci.* **2022**, *12*, 12307. [CrossRef]
37. Al-Absi, A.; Kang, D.; Al-Absi, M. Sentiment Analysis and Classification Using Deep Semantic Information and Contextual Knowledge. *Comput. Mater. Contin.* **2023**, *74*, 671–691. [CrossRef]
38. Hadwan, M.; Al-Hagery, M.; Al-Sarem, M.; Saeed, F. Arabic Sentiment Analysis of Users' Opinions of Governmental Mobile Applications. *Comput. Mater. Contin.* **2022**, *72*, 4675–4689. [CrossRef]
39. Musleh, D.; Alkhales, T.; Almakki, R.; Alnajim, S.; Almarshad, S.; Alhasaniah, R.; Aljameel, S.; Almuqhim, A. Twitter Arabic Sentiment Analysis to Detect Depression Using Machine Learning. *Comput. Mater. Contin.* **2022**, *71*, 3463–3477. [CrossRef]
40. Muhammad, A.; Abdullah, S.; Sani, N. Optimization of Sentiment Analysis Using Teaching-Learning Based Algorithm. *Comput. Mater. Contin.* **2021**, *69*, 1783–1799. [CrossRef]
41. Bhatti, M.; Azhar, S.; Sohail, A.; Hijji, M.; Ayemen, H.; Ramzan, A. Multilingual Sentiment Mining System to Prognosticate Governance. *Comput. Mater. Contin.* **2022**, *71*, 389–406. [CrossRef]
42. Qureshi, M.; Asif, M.; Hassan, M.; Mustafa, G.; Ehsan, M.; Ali, A.; Sajid, U. A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis. *Comput. Mater. Contin.* **2022**, *7*, 4987–5004. [CrossRef]
43. Hilal, A.; Alfurhood, B.; Al-Wesabi, F.; Hamza, M.; Duhayyim, M.; Iskandar, H. Artificial Intelligence Based Sentiment Analysis for Health Crisis Management in Smart Cities. *Comput. Mater. Contin.* **2022**, *71*, 143–157. [CrossRef]
44. Albahli, S.; Algsham, A.; Aeraj, S.; Alsaeed, M.; Alrashed, M.; Rauf, H.; Arif, M.; Mohammed, M. COVID-19 Public Sentiment Insights: A Text Mining Approach to the Gulf Countries. *Comput. Mater. Contin.* **2021**, *67*, 1613–1627. [CrossRef]
45. Qin, Z.; Ronchieri, E. Exploring Pandemics Events on Twitter by Using Sentiment Analysis and Topic Modelling. *Applied Sciences* **2022**, *12*, 11924. [CrossRef]
46. Alanazi, S.; Khaliq, A.; Ahmad, F.; Alshammari, N.; Hussain, I.; Zia, M.; Alruwaili, M.; Alanazi, R.; Alsayat, A.; Afsar, S. Public's Mental Health Monitoring via Sentimental Analysis of Financial Text Using Machine Learning Techniques. *Environ. Res. Public Health* **2022**, *19*, 9695. [CrossRef]
47. Alali, M.; Sharef, N.; Murad, M.; Hamdan, H.; Husin, N. Multitasking Learning Model Based on Hierarchical Attention Network for Arabic Sentiment Analysis Classification. *Electronics* **2022**, *11*, 1193. [CrossRef]
48. Omara, E.; Mosa, M.; Ismail, N. Emotion Analysis in Arabic Language Applying Transfer Learning. In Proceedings of the IEEE International Conference on Computer Engineering, Cairo, Egypt, 9 March 2020. [CrossRef]
49. Alwehaibi, A.; Roy, K. Comparison of Pre-trained Word Vectors for Arabic Text Classification using Deep Learning Approach. In Proceedings of the IEEE International on Machine Learning and Applications, Orlando, FL, USA, 17 January 2019. [CrossRef]
50. Chouikhi, H.; Alsuhaibani, M.; Jarray, F. BERT-Based Joint Model for Aspect Term Extraction and Aspect Polarity Detection in Arabic Text. *Electronics* **2023**, *12*, 515. [CrossRef]
51. Arabic Sentiment Analysis 2021 @ KAUST. Available online: https://kaggle.com/competitions/arabic-sentiment-analysis-2021-kaust (accessed on 28 December 2022).

52. Zhang, X.; Liu, C. Model Averaging Prediction by *K*-Fold Cross-Validation. *J. Econom.* 2022, in press. [CrossRef]
53. Criminisi, A.; Shotton, J.; Konukoglu, E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Found. Trends Comput. Graph. Vis.* **2012**, *7*, 81–227. [CrossRef]
54. Gupta, R.; Agrawalla, R.; Naik, B.; Evuri, J.; Thapa, A.; Singh, T. Prediction of Research Trends Using LDA Based Topic Modeling. *Glob. Transit. Proc.* **2022**, *3*, 298–304. [CrossRef]
55. Balaji, T.; Annavarapu, C.; Bablani, A. Machine Learning Algorithms for Social Media Analysis: A Survey. *Comput. Sci. Rev.* **2021**, *40*, 100395. [CrossRef]
56. Jordan, M.; Mitchell, T. Machine learning: Trends, perspectives, and prospects, Science. *Science* **2015**, *349*, 255–260. [CrossRef]
57. Saritas, M.; Yasar, A. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *Int. J. Intell. Syst. Appl. Eng.* **2019**, *7*, 88–91. [CrossRef]
58. Istia, S.; Purnomo, H. Sentiment Analysis of Law Enforcement Performance Using Support Vector Machine and K-Nearest Neighbor. In Proceedings of the 3rd IEEE International Conference on Information Technology, Information System and Electrical Engineering, Yogyakarta, Indonesia, 13–14 November 2018. [CrossRef]
59. Chen, J.; Li, K.; Tang, Z.; Bilal, K.; Yu, S.; Weng, C.; Li, K. A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment. *IEEE Trans. Parallel Distrib. Syst.* **2016**, *28*, 919–933. [CrossRef]
60. Zhou, J.; Qiu, Y.; Armaghani, D.; Zhang, W.; Li, C.; Zhu, S.; Tarinejad, R. Predicting TBM Penetration Rate in Hard Rock Condition: A Comparative Study among Six XGB-Based Metaheuristic Techniques. *Geosci. Front.* **2021**, *12*, 101091. [CrossRef]