Intelligent Hybrid Feature Selection for Textual Sentiment Classification

JAWAD KHAN 1, AFTAB ALAM 2, and YOUNGMOON LEE 1 (MEMBER, IEEE)

1Department of Robotics, Hanyang University, Ansan-si 15588, South Korea.
2College of Science and Engineering, Hamad Bin Khalifa University, Education City, Ar Rayyan, Qatar.

Corresponding author: Youngmoon Lee (e-mail: youngmoonlee@hanyang.ac.kr)

ABSTRACT Sentiment Analysis (SA) aims to extract useful information from online Unstructured User-Generated Contents (UUGC) and classify them into positive and negative classes. State-of-the-art techniques for SA suffer a high dimensional feature space because of noisy and irrelevant features from the UUGC. Researchers have also proposed feature extraction and selection techniques to reduce high dimensional feature space, but they fall short in extracting and selecting the most effective sentiment features for sentiment model learning. Effective feature extraction and selection are significant for the SA because they can boost the learning algorithm’s predictive performance while reducing the high-dimensional feature space. To address these concerns, we propose an Intelligent Hybrid Feature Selection for Sentiment Analysis (IHFSSA) based on ensemble learning methods. IHFSSA first identifies sentiment features in the review text utilizing Penn Treebank part-of-speech tagset and integrated Wide Coverage Sentiment Lexicons (WCSL). The sentiment features subset is then selected employing a fast and simple rank-based ensemble of multiple filters feature selection method. The selected sentiment features are further refined by applying a wrapper-based backward feature selection method. Finally, for textual sentiment classification, the well-known classification algorithms Support Vector Machine (SVM), Naive Bayes (NB), Generalized Linear Model (GLM) are trained in the ensemble model on the refined sentiment feature set. The in-depth evaluation using heterogeneous domain benchmark datasets demonstrates that IHFSSA outperforms existing SA techniques.

INDEX TERMS Sentiment classification, Hybrid feature selection, Ensemble learning, Linguistic semantic rules, Wide coverage sentiment lexicons, Natural Language Processing

I. INTRODUCTION

Blogs, discussion forums, shared knowledge-seeking networks, social network platforms, and product and movie review portals [1]–[5] are only a handful of social media platforms that have come up with Web 2.0 [6], [7]. On these social platforms and review websites, users express their thoughts, opinions, and perceptions about a range of topics, activities, people, organizations, goods, and services. These expressions are often beneficial to both businesses and customers.

Such user-generated content must be analyzed to become usable information for business strategy and product purchase decisions. However, due to their unstructured nature and huge volume, user-generated contents incur significant cognitive overload to manually search and evaluate. The challenge in accurately extracting valuable knowledge from the massive volume of UUGC is the major hurdle for product manufacturers and data scientists.

SA, also called opinion mining, is the process to analyze, process, and extract the useful knowledge from UUGC and classify it into positive and negative classes employing text mining and machine learning. Textual SA has been carried out at three levels: document, sentence, and aspect [8]. In this paper, we focus on document-level SA, processing each sentence in the document and convert them to words employing sentence parser and tokenizer, respectively.

In textual SA, text pre-processing/ transformation, and feature extraction/selection are vital process [9]. The former is used to create a feature vector for machine understanding, while the latter process is used...
to extract and select the most important features for SA. Specifically, feature extraction focuses to identify and extract sentiment words/phrases, while, in feature selection, the subset of most important and relevant features are selected. Feature extraction and selection are essential for making a scalable generalized model to improve performance and minimize model complexity and storage requirements [10].

Proposals for sentiment classification acknowledge the importance of feature extraction and selection [9], [11]–[14]. State-of-the-art techniques for sentiment classification mostly depend on the conventional Bag-of-Words (BOW) approaches [4], [15], [16]. These approaches can be useful due to their simplicity but they ignore semantics and word order, suffering high dimensionality. Researchers have also proposed semantic-based SA methods utilizing high order n-grams, Part-of-Speech (POS) patterns, and dependency relation features considering word order [9], [11], [12], [17]–[19]. However, not every word or phrase in high-order n-gram, POS patterns, and dependency relation features shows sentiment. Therefore, extracting and selecting the subset of the most relevant feature by considering semantics, word order, and sentiment information is an important task for robust and intelligent model development and high classification performance. Khan et al. [20] focused on exploring POS patterns and POS n-gram patterns for SA while considering sentiment clue and employing ensemble techniques. However, unlike existing work [20], IHFSSA uses hybrid of filter and wrapper method in the selection and refinement of optimum sentiment features to address the issue of SA and to improve the sentiment classification performance. Moreover, it also tackles the challenge of context-based SA that contains mixed opinions and standardizes WCSL to identify the sentiment features.

This paper addresses these challenges via integrating two main types of feature selection, namely, filter and wrapper methods. The filter-based method employs ranking criteria to assign a score (weight) to each feature based on the algorithm implementation. The top ranks subset features are selected according to filter criteria. Examples are Document Frequency (DF), Information Gain (IG), Chi-square (CHI), and Mutual Information (MI) [21]. The filter method is simple, fast, but neglects the interconnection with the classifier. The wrapper approach evaluates the features and selects the most important feature that has the greatest effect on classification performance. Forward selection, backward selection, and evolutionary search are some examples of wrapper method [22], [23]. The wrapper method is computationally costly than the filter-based method.

In this paper, we propose an intelligent model for textual SA based on a hybrid feature selection with ensemble learning methods. The Ensemble of Multiple Filters Feature Selection (EMFFS) is integrated with a wrapper-based backward feature selection (BFS) method. The goal of integrating filter and wrapper methods is to utilize its benefit to find the most effective/optimal sentiment features while reducing feature space and enhancing sentiment classification performance. In this study, first, the EMFFS method that is based on well-known multiple base filters, such as IG, CHI, Gain Ratio (GR), Standard Deviation (SD), and Gini Index (GI), is employed. Each base filter ranks the feature set of a dataset and then selects top-k features. After that, the subset of top-k common features is selected by individual base filters based on the Majority Voting Threshold (MVT). Next, the selected common features are further refined to get the most effective sentiment features by the wrapper-based BFS method. Finally, we train the classification model on the refined most effective sentiment features for SA. We conducted the experiments on different benchmark review datasets which are widely used for SA. The experimental results show that the proposed intelligent hybrid approach is capable of accurately predicting the true sentiment class, as well as improving the performance of SA efficiently and outperforming baseline methods.

This paper makes the following contributions:

- We propose IHFSSA that integrates filter-based EMFFS with wrapper-based Backward Feature Selection (BFS) to select optimum sentiment features with reduce feature space and improve sentiment classification performance.
- We utilize the integrated WCSL to identify the sentiment features in the review text.
- We employ linguistic rules to determine the correct class of review text that contains both positive and negative sentences/clauses.
- We apply POS-unigram and bigram to extract semantic and context-aware sentiment features.
- We refine and reduce the feature space to obtain the subset of most effective sentiment features for sentiment classification.
- We evaluate IHFSSA using two well-known benchmark datasets and demonstrate the effectiveness of hybrid feature selection and improvement in sentiment classification.

The remaining paper is organized as follows. Sec-
II. RELATED WORK

SA systematically identifies, extracts, selects, and classifies sentiment information from source materials using Natural Language Processing (NLP), text mining, and Machine learning (Machine Learning (ML)) techniques [8]. Before 2000, neither NLP nor linguistics received much coverage. The main reason was that there was only a limited amount of opinionated textual data available in digital format at the time. Since its inception in the year 2000, the field of SA has expanded to become one of the most active research fields in NLP, data mining, web mining, information retrieval, and management sciences [8]. The SA techniques have been broadly used for a variety of purposes, including classification of customer reviews, [24]–[26], tracing political viewpoints [27], [28], and predicting stock market trends [29], [30] etc.

Researchers have studied various feature representation schemes and proposed different techniques to improve sentiment classification performance in several studies on textual SA. For sentiment classification, Pang et al. [13] used a supervised machine learning technique. For movie review sentiment classification, they used three machine learning classifiers (NB, ME, and SVM). They utilized various n-gram feature sets, such as unigrams, bigrams, unigrams bignrams together, and unigrams with POS tags, to classify sentiment. They achieved 81%, 82.9%, and 80.4% accuracy for NB, maximum entropy, and SVM classifiers respectively. They also explained that either NB or SVM classifier using unigram features performed well. Similarly, Go et al. [31] used the distant supervision method to conduct SA on Twitter message. They were able to effectively use emoticons as noisy labels to train data for distant supervised learning. They conducted experiments employing NB, SVM and maximum entropy, and achieved an accuracy of 81.3%, 82.2%, and 80.5% respectively. Zhang et al. [32] used NB and SVM for the sentiment classification of restaurant review written in Cantonese. They also examined how feature representation and feature size affect classification accuracy. The NB classifier achieved the highest accuracy 95.67%.

For the movie review classification, Tripathy et al. [11] used four supervised machine learning classifiers: NB, ME, SVM, and Stochastic Gradient Descent. They demonstrated n-gram techniques, unigram, bigram, trigram, and various combinations of these for sentiment feature generation. They had the best results on the SVM classifier by merging unigram and bigram, as well as a unigram, bigram, and trigram. Using n-grams and a dependency relation scheme, Ng et al. [17] used an SVM classifier to classify textual reviews. They addressed two issues: the presence of user reviews in a text document, as well as the polarity of those reviews (positive or negative). M. Rushid et al. [33] presented different feature representation schemes (TFIDF, BO, TO) and n-grams techniques (unigrams, bigrams, and trigrams). They used SVM with a variety of feature sets and evaluated it on a variety of domains. According to their findings, the trigram model has a significant edge over the unigram and bigram models. Moraes et al. [34] compared SVM and NB with an ANN-based technique for sentiment classification. They carried out experiments on four distinct domain datasets: movie dataset, GPS, Books, and Cameras dataset. In all datasets, ANN obtained the highest classification accuracy, but in the unlabeled dataset, both ANN and SVM performed poorly.

On product review and movie review datasets, Singh et al. [35] analyzed and compared the performance of Multi-layer Perceptron, NB, and SVM. In both the movie and product review datasets, SVM worked best. The presumption of attributes in NB is being independent, which may not be necessary in any case. MLP needs more execution time compared to other techniques. Xue Bai et al. [36] utilized ME, NB, SVM, and Markov Blanket Model algorithms for sentiment classification. They evaluated their models using datasets from movie reviews and news articles. They suggested a heuristic search-enhanced Markov blanket model to grab word dependencies and provide the vocabulary for sentiment extraction. In comparison to other classifiers, their proposed model yielded good results. Kalaivani et al. [37] compared SVM, NB, and KNN for the movie reviews sentiment classification. The SVM approach outperformed NB and KNN approaches, according to their experimental results. The accuracy of SVM was reported for more than 80%. Bilal et al. [38] researched Urdu and English blogs opinions. They employed NB, Decision Tree, and KNN to classify Urdu and English blogs’ opinions. According to their results NB outperformed Decision Tree and KNN.

Turney [14] presented an unsupervised approach.
for review document sentiment classification (recommended or not recommended). In Turney’s work, the review document is classified by computing the average semantic orientation of phrases containing adjectives or adverbs. Another unsupervised learning approach presented by Xiangua et al. [39] to automatically determine the aspects addressed in Chinese reviews as well as the sentiments expressed in different aspects. They found multi-aspect global topics using the Latent Dirichlet Allocation (LDA) model, then extracted local topics and related sentiments using a sliding window context over the review document. The local topic was determined by the trained LDA model, and the polarity of related sentiment was classified by the HowNet lexicon. The results of their experiments indicate that their method performed well in topic exploration and sentiment analysis. Mullen et al. [40] assigned semantic values to words and phrases then, incorporated them to create a classification model. Mastumoto et al. [41] defined the syntactic relationship between words as the basis for review text sentiment classification. They mined frequent word sub-sequence and dependency sub-trees from sentences and used them as features for sentiment classification using an SVM classifier.

Feature selection is an essential task in SA. Feature selection helps to minimize the dimensionality of feature space by choosing the subset of the most useful and effective features without losing too much information. Feature selection methods are mainly divided into two groups, namely filters and wrappers. The filter method chooses a subset of relevant features that are independent of the classifier. The wrapper method for feature selection uses classifier performance to select a subset of relevant features. Tan et al. [42] utilized four statistical filters for feature selection namely Mutual Information, IG, CHI, and Document Frequency with five ML algorithms (K-nearest neighbor, Centroid Classifier, Window Classifier, NB Classifier, and SVM) for Chinese corpus sentiment classification. Their results conducted on Chinese corpus show that IG and SVM perform the best among all the feature selection methods and ML algorithms for sentiment classification. Sharma et al. [43] employed five filter-based feature selection techniques namely IG, CHI, Document Frequency, GR, and Relief for SA on online movie review dataset. Their study shows that features selection can boost the sentiment classification performance but it is dependent on the number of features selected and the technique used to select them. The results of their experiments indicate that GR outperformed all other techniques.

Table 1: An overview of textual sentiment analysis techniques

| Author/Reference | Year | Approach | Algorithm | Dataset | Best Accuracy (%) |
|------------------|------|----------|-----------|---------|-------------------|
| Pang et al. [13] | 2002 | Sentiment classification using machine-learning techniques and n-gram model | Naive Bayes, Maximum Entropy, Support Vector Machine, Maximum Entropy | Movie Reviews | 82.9 |
| Mullen et al. [39] | 2004 | Assigned semantic values to selected words and phrases then, incorporated them to create a classification model | Hybrid of Support Vector Machine | Movie Reviews | 86.0 |
| Ng et al. [17] | 2006 | Review identification and polarity classification employing n-grams and dependency relations | Support Vector Machine | Movie Reviews | 90.5 |
| Matsunoto et al. [40] | 2005 | Sentiment classification using syntactic relation between words in sentences | Support Vector Machine | Movie Reviews | 92.9 |
| Tripathy et al. [11] | 2016 | Transforming text reviews into numeric matrices utilizing countvectorizer and TF-IDF with n-gram model to machine learning algorithm for classification | Naive Bayes, Maximum Entropy, Support Vector Machine, Maximum Entropy, Stochastic Gradient Descent | Large Movie Reviews | 88.94 |
| Xia et al. [12] | 2011 | Designed part-of-speech based feature sets and the word-relation based feature sets with ensemble strategies for sentiment classification | Ensemble classifier | Movie, Book, Electronic, Kitchen, DVD | 87.70, 81.80, 85.95, 88.65, 83.80 |
| Alireza et al. [19] | 2017 | Original-based and frequency-based integration of feature selection for sentiment analysis | Support Vector Machine, Naive Bayes, Maximum Entropy, Linear discriminant Function | Movie, Book, Electronic, Kitchen, Music | 92.92, 85.13, 85.97, 86.62, 85.44 |
| Osman et al. [4] | 2020 | Wrapper feature selection algorithm based on Iterated Greedy (IG) metaheuristic for sentiment classification | Bayesian Logistic Regression, Naive Bayes, Multinomial Naive bayes | Book, Electronics , Kitchen, DVD | 88.00, 91.63, 93.09, 90.25 |
| Jing et al. [15] | 2020 | Modified categorical proportional difference (MCBD) which improves conventional CPD method, and balance category feature (BCF) strategy that equally selects attributes from both positive and negative examples for sentiment classification | Suport Vector Machine | Movie Reviews | 93.50 |
subsets (FIFS). They used unigrams and POS patterns for the sentiment features identification along with their proposed feature selection techniques. Their experimental results show that POS-based features are more effective than unigram-based features. Osman et al. [4] presented a wrapper feature selection algorithm for sentiment classification based on the Iterated Greedy meta-heuristic. They used Multinomial NB as a classifier and GR filter scores as heuristic information for greedy selection. Their experimental results show that their proposed algorithm outperforms traditional filter-based feature selection techniques as well as the Genetic Algorithm-based feature selection algorithm. Jing et al. [15] proposed two feature selection methods called modified categorical proportional difference (MCPD) and balance category feature (BCF) that equally selects attributes from text reviews. Their experimental results showed that the combination of BCF and MCPD methods can not only reduce feature space but also improve the sentiment classification performance.

Ensemble techniques have reported successful results and improved classification performance for many domains [12], [44]–[48]. Kalaivani et al. [49] proposed machine learning-based feature selection method utilizing IG and Genetic Algorithm. They applied NB, logistic regression, SVM, and ensemble techniques on multi-domain datasets and movie review datasets for evaluation. According to their experimental results, IG and genetic algorithm with ensemble technique performed better. Xia et al. [12] investigated the effectiveness of ensemble techniques on different feature sets for sentiment classification. Khan et al. [20] proposed a novel POS and n-gram based ensemble method for SA while considering semantics, sentiment clue, and order between words called EnSWF. In this method appropriate features for SA were extracted and selected based on the POS patterns and POS n-gram patterns employing ensemble techniques. Onan et al. [16] proposed an ensemble approach to feature selection, in which several feature ranking lists obtained by different feature selection methods were aggregated. They used the genetic algorithm for the aggregation of ranked feature lists. They tested their proposed ensemble approach on a variety of domain datasets and found that it outperforms individual filter-based feature selection methods. Table 1 provides an overview of different approaches proposed by different authors for textual sentiment classification.

Recently, deep learning-based algorithms such as long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), gated recurrent unit (GRU), bidirectional gated recurrent unit (BiGRU), convolutional neural network (CNN) [50], capsule network [51] and Transformer based language model (BERT) [52] has also been widely used for SA and challenging natural language processing applications that yield state-of-the-art prediction results. The scope of this work is traditional feature-based approaches. Such techniques are still useful for relevant features selection and dimensionality reduction. Also these methods have their own advantages in interpretability and time complexity. Furthermore, the main objective of this study is to select the most optimum sentiment features with reduced feature space and to improve the sentiment classification performance.

According to the literature review, different feature extraction or/and selection strategies, as well as ensemble learning methods for sentiment classification, have been introduced by researchers. They also improved the sentiment classification results. However, the majority of them overlooked the successful use of hidden significant information that contains more sentiment information than other vocabularies. Likewise, they did not consider the most effective sentiment features. The existence of noisy and irrelevant features in the text is also ignored. Besides, state-of-the-art feature extraction or/and selection techniques have their own statistical biases and also required an educated guess to identify appropriate features. For the evaluation of the most appropriate sentiment features in SA, existing sentiment lexicons and domain knowledge may be used. The combination of the filter-based feature selection method with the wrapper-based backward feature selection method employing the most effective sentiment features for SA is the focus of this paper.

III. METHODOLOGY

Figure 1 depicts the framework of our proposed hybrid approach for textual sentiment classification. There are four major phases in the framework: (1) Feature representation (2) Sentiment features extraction (3) Sentiment features selection (4) Training and applying classification model.

We, first, split the review text into sentences, then transform it into words and assign POS tags to recognize suitable words such as an adjective, adverb, verb, and noun. After this, the identified POS words are matched with the WCSRL to extract suitable sentiment words. Second, we apply the n-gram model (unigram and bigram) to the POS fea-
features and generate POS tagged unigram and bigram features. Besides, we utilize dimensionality reduction techniques (Filter and wrapper-based features selection) to reduce the feature set and select the most effective refine sentiment features for accurate sentiment classification. The filter method is simple, fast, and scales well with the dimensions and numbers of samples. The wrapper method takes care of interaction among features and increases classification performance. Finally, we utilize three distinct supervised ML algorithms such as SVM, NB, and GLM as base learners in conjunction with classifier ensemble (CE) employing majority voting approach for sentiment classification. The main goal of this study is to select the most effective sentiment features with reduce feature space and to improve the sentiment classification performance. The technical details of the proposed methodology are elaborated in the following sub-sections.

A. FEATURE REPRESENTATION

The identification of suitable features from unstructured textual data is a challenging task in sentiment classification [9]. In literature, various feature representation methods have been used for textual sentiment classification. These methods consists: BOW, POS patterns, linguistic patterns, n-grams, dependency relation features, semantic and contextual features [19], [53], [54]. In this study, we use different modules for feature representation and construction. These modules are sentence parser, linguistic semantic rules, tokenizer, noise remover, lower case transformer, POS tagger, and Term Frequency-Inverse Document Frequency (TFIDF) scheme.

The product review dataset is first loaded, after which the review text is split into words using the sentence parser and tokenizer. The noise remover module is then used to eliminate all noise from the text, and the case transformer module is used to convert the text to lowercase. Further, the POS tags are assigned to the input words by the POS tagger to identify and extract likely words like adjectives, adverbs, verbs, and nouns. After that, these words are searched in the integrated WCSL to extract the suitable sentiment words and remove non-sentiment words. Finally, the feature vector is created and selected using the TFIDF and pruned words below ‘absolute = 3.0 and above ‘absolute = 30000 from the feature vector space. In this study, We employed the Penn Treebank annotation scheme [55] for POS tagging as shown in Table 4.

On product review datasets, the feature representation process also applies linguistic rules. Linguistic rules provide help to the context-based SA that contains mixed opinions. For example, in the sentence, "the director is popular but the movie is boring", in this sentence using linguistic rules, only the clause after "but" is considered and the clause before "but" is ignored. It contains specific words, like 'but', 'despite', 'while', 'unless' etc., that can switch the polarity of a sentence. Many review documents contain mixed opinions. The use of linguistic rules could positively affect the reviews that contain mixed opinions. Based on the assump-
tion of previous researches, each review document or sentence has a single polarity. Following the work [56], [57], we apply the same linguistic rules as shown in Table 2.

B. INTEGRATED WIDE COVERAGE SENTIMENT LEXICONS

Sentiment lexicon is a collection of terms used to express positive or negative sentiments [8]. Sentiment lexicons plays a vital role in sentiment words detection. In literature, various sentiment lexicons such as SentiWordNet [58], WordNet-Affect [59], SenticNet5 [60], NRC Hashtag Sentiment lexicon [61], MPQA Subjectivity Lexicon [62], SentiSense [63], GI [64], SO-CAL [65], OL [24], AFFIN [66] with different sizes have been built. There is no optimum general sentiment lexicon to be used for SA, besides, the format of the sentiment lexicons are different from each other.

We standardize and integrate the aforementioned ten state-of-the-art sentiment lexicons to have one of three different scores, +1,-1, 0. We compute the average sentiment score of the overlapping words in these lexicons to get more sentiment words and create a huge sentiment lexicon. The size and format of the state-of-the-art sentiment lexicons for word coverage are shown in Table 3. First, we standardized them by assigning scores, +1,-1, 0 to positive, negative, and neutral words respectively. Then calculated the sentiment score of every word by the average score of the overlapping words in the sentiment lexicons.

In this study, sentiment words in the review documents are matched against integrated wide coverage sentiment lexicons and then used for sentiment classification.

C. SENTIMENT FEATURES EXTRACTION

Feature extraction is a method of identifying and extracting text features that are useful for classification. The primary issue in sentiment classification is the extraction of appropriate features that perform better than simple features [9]. Identification of appropriate features is critical for reliable model learning. BOW, higher-order n-grams, POS, and negation are some of the feature extraction methods used in SA. In this step, we identify and extract suitable features that express sentiment clues. In literature, four types of POS words, i.e. adjectives and adverbs, accompanied by verbs and nouns are considered as the sentiment features for sentiment classification [12], [19]. We use the same POS words for an initial set of feature extraction in this study, based on the aforementioned literature. For the initial set of sentiment features extraction, first, the extracted POS words are matched with the integrated WCSL. Then the n-gram model (unigram and bigram) is utilized to generate POS tagged unigram and bigram sentiment features. The sentiment features are usually single words, i.e., unigrams (beautiful, nice, bad, awful, etc.) or bigrams (very nice, much better, not bad). As a result, we extracted sentiment features using unigram and bigram. The list example set for the N-gram model is shown in Table 5.

D. FEATURES SELECTION

The process of selecting the most important and applicable features in order to minimize the high-dimensional feature space and make an accurate prediction is known as feature selection. Feature selection has the potential to improve classification performance. Feature selection techniques can be mainly classified into filter and wrapper methods. The filter method assigns a score (weight) to each feature based on the relevant algorithm implementation using ranking criteria. The filter method is simple, fast, and scales well with the dimensions and the number of samples. A classifier is used in the wrapper approach to evaluate a subset of features and select the most significant feature that has the greatest impact on classification performance. The
Table 2: Linguistic semantic rules

| Rule | Example | Linguistic Semantic Rules |
|------|---------|--------------------------|
| R1   | I have never like that director, but I love this movie. | If a sentence contains “but”, ignore previous sentiment and only consider the sentiment after “but” part. |
| R2   | I love this movie, despite the fact that I hate the director. | If a sentence contains “despite”, only consider the sentiment before the “despite” part. |
| R3   | Everyone like this video unless he is a sociopath. | If sentence contains “unless” and the “unless” is followed by a negative clause, ignore the “unless” clause. |
| R4   | While they did their best, the team played a horrible game. | If a sentence contains “while” ignore the sentence following the “while” and consider the sentiment only of the sentence that follows the one after the “while”. |
| R5   | The film counted with good actors, however, the plot was very poor. | If a sentence contains “however” ignore the sentence before “however” and consider the sentiment of the sentence after “however”. |

Table 3: State-of-the-art sentiment lexicons

| Lexical Resources (Sentiment Lexicon) | Size | Category | Sentiment Score |
|--------------------------------------|------|----------|-----------------|
| AFFIN                                | words:2477 | N/A | [5, +5] |
| GI                                   | Words: 11,789 | Positive, Negative, pstv, ptgv, strong, weak etc. | N/A |
| OL                                   | Words: 6786 | Positive, Negative | N/A |
| SentiWordNet                         | 117,659 Synsets, Words: 155,287 | Positive, negative and objective | [0, 1] |
| SO-CAL                               | Words: 6306 | N/A | [5, +5] |
| Subjectivity Lexicons                | Words: 8,221 | Positive, negative and neutral | N/A |
| WordNet-Affect                        | 2,872 Synsets, Words: 4,785 | Positive, Negative, Ambiguous, Neutral | N/A |
| NRC Hashtag Sentiment lexicon        | Words: 54,129 | Positive, Negative | [7, 1] |
| SenticNet                            | Concepts: 100,000 | Positive, Negative | [+1, +1] |
| SentiSense                           | 2,190 Synsets, Words: 5,496, | Like, Love, Hope, Hate etc. | N/A |

Table 4: POS Tags

| S.No. | Tag | Description |
|-------|-----|-------------|
| 1     | JJ  | Adjective   |
| 2     | JJR | Adjective, comparative |
| 3     | JJS | Adjective, superlative |
| 4     | RB  | Adverb      |
| 5     | RBR | Adverb, comparative |
| 6     | RBS | Adverb, superlative |
| 7     | VB  | Verb, base form |
| 8     | VBD | Verb, past tense |
| 9     | VBG | Verb, gerund or present participle |
| 10    | VBN | Verb, past participle |
| 11    | VBP | Verb, non-3rd person singular present |
| 12    | VBZ | Verb, 3rd person singular present |
| 13    | NN  | Noun, singular or mass |
| 14    | NNS | Noun, plural |
| 15    | NNP | Proper noun, singular |
| 16    | NNPS| Proper noun, plural |

Table 5: Example of N-gram model

| N-Gram | Unigram | Bigram |
|--------|---------|--------|
| This is amazing film | This is, is amazing, amazing film |

Greedy search algorithm is used for the wrapper process. Even though it is computationally costly, the interaction between features is observed. The five well known statistical filters in the EMFFS for sentiment features selection are used in the proposed method, which is described as follows:

**Information gain (IG):** The relevance of features in the documents for the given category prediction is calculated by information gain. IG assigns weights to features. The higher a feature weight, the more relevant it is considered to be [21]. Based on information gain value (weight), we pick the top-ranked most appropriate N features. IG is used to reduce the size of features by retaining the highest-ranked features and excluding those with information gains smaller than a predefined threshold. Equation 1 (1) [67] defines the IG for the feature ‘X’.

\[ IG(X, Y) = H(X) - H(X | Y) \]  \hspace{1cm} (1)

Where X is the feature, and Y is class.

**Standard deviation (SD):** The standard deviation indicates how much divergence or dispersion occurs in the feature space from the average (mean, or expected value). A low standard deviation means that the features are fairly similar to the mean, while a high standard deviation indicates that the features are distributed over a wide range of values. The formula is straightforward: it is equal to the square root of the variance. The standard deviation is measured as follows [68].

\[ SD(f_i) = |Stdev(f_i, C_1) - Stdev(f_i, C_2)| \]  \hspace{1cm} (2)

\[ Stdev(f_i, C_k) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{ij} - mean_k(f_i))^2} \]  \hspace{1cm} (3)
where \( f \) is the feature, \( N \) is the number of instances in a class, \( x_{ij} \) is the weight of \( j^{th} \) feature, \( mean_k \) is the mean and \( Stdev(f, C_k) \) is the standard deviation of the \( j^{th} \) feature to the \( k^{th} \) class.

**Chi-square (CHI):** The chi-square [21] is used to compute the degree of association between feature \( f \) and its category \( c \), and choose the desired number of features with the best chi-square score. The intuition is that if the category \( c \) is independent of the feature \( f \) then the feature is removed. Using Chi-square, the worth of a feature is measured by computing the value of the chi-square statistic with respect to the target class. The formula for the Chi-square is given in equation 4 [21].

\[
X^2(f, c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{4}
\]

Where \( f \) is the feature, \( c \) is its category, \( N \) is the number of total documents, \( A \) is the number of times \( f \) and \( c \) co occur, \( B \) is the number of times \( f \) occurs without \( c \), \( C \) is the number of times \( c \) occurs without \( f \), and \( D \) is the number of times neither \( c \) nor \( f \) occurs.

**Gain Ratio (GR):** The Gain Ratio, also known as the Information Gain Ratio, is a modification of information gain that decreases its bias against high-branch features. It is the ratio of information gain to intrinsic information gain [69]. The GR formula is given in Equation 5.

\[
GR = \frac{IG(X)}{IntrinsicInf_{0}(X)} \tag{5}
\]

Where \( X \) represents the feature.

**Gini Index (GI):** The original GI algorithm is used to assess the impurity of features in terms of classification. The feature’s GI is then the smallest at all judgment thresholds \( f \) in [0,1] [48]. The lower the value, the lower the impurity and the better the feature. On the other hand, when calculating the purity of a feature for categorization, the higher the value reflects the purity and the greater the feature. The calculation of purity has been used in several experiments on the GI theory. The improved GI formula [48] is given in the below equation 6.

\[
GiniIndTxt(W) = \sum_{i=1}^{M} P(W|C_i)^{\alpha} P(C_i|W)^{2} \tag{6}
\]

**E. THE ENSEMBLE OF MULTIPLE FILTERS FEATURE SELECTION**

The EMFFS contains five state-of-the-art different statistical filters, namely IG, SD, CHI, GR, and GI are used as base fillers for features selection. In this method, first, the features are ranked by the aforementioned base filters. Then the subset of top-k common ranked features based on MVT is selected. A MVT ‘\( \lambda \)’ (i.e. \( \lambda \geq t, t = 3 \), where \( t \) is the value of MVT is used to choose common ranked features selected by at least three filters among base filters. After that, the ranks of common features are aggregated using the arithmetic mean. Furthermore, normalization is used to scale the values of these features to fit in a certain range. When dealing with features of various scales, adjusting the value range is important. In feature ranking, each statistical filter assigns a different ranking score to features that may reveal various scales. In order to avoid various scales, we use feature scaling to re-scale the features’ values assigned by different statistical filters in the range between 0 and 1 employing min-max normalization techniques. The equation 7 is used to normalize the values of selected common features. Finally, the scaled features are sorted in descending order for further processing (refinement). Generally, the aim of EMFFS method is to produce the subset of relevant sentiment features and is to decrease the risk of irrelevant features that do not meet the MVT criteria. The workflow of EMFFS is shown in Figure 2.

\[
\text{Min-Max Normalization} = \frac{\text{value} - \text{min}}{\text{max} - \text{min}} \tag{7}
\]

**F. WRAPPER-BASED BACKWARD FEATURE SELECTION**

The wrapper approach generates and evaluates various subsets of features in the feature space. The heuristic search algorithm is used to generate a subset of features, and a classifier evaluates the goodness of a feature subset. Wrapper approaches often result in better classification performance. BFS, also known as backward elimination, is the wrapper approach used in this analysis. We use BFS to eliminate features that do not have a noticeable impact on the prediction performance. The BFS is an iterative process. It begins with the complete set of features/attributes and eliminates each remaining attribute of the given example set in each round. Cross-validation is used to estimate the performance for each removed attribute. Only the attribute giving the smallest decrease of performance is finally omitted from the selection. Then a new round is started with the modified selection. This process is repeated as long as there is a performance improvement. When there is no substantial improvement in per-
et al.: Effective Scholarly Search

Figure 2: Ensemble of multiple filters feature selection.

Figure 3: Backward feature selection.

G. EFFECTIVE FEATURE SELECTION USING HYBRID APPROACH

This research proposes a hybrid feature selection approach for effective sentiment feature selection SA that includes both filter-based EMFFS and wrapper-based BFS methods. For appropriate feature selection, filter methods use statistical and mathematical criteria. Filter methods are computationally simple, fast, and scalable, but they ignore the classifier’s interconnection. Furthermore, since this approach is univariate, treat each feature individually and relinquish it to establish the relationship between features. Wrapper methods use the classifier to pick the best features. In high-dimensional feature space, these approaches (Wrappers) are computationally costly. Although wrapper methods provide better classification results in general. As a result, using a hybrid solution is the intelligent one. It makes a trade-off between filter and wrapper method. Algorithm 2 describes the detailed method of the proposed solution for best/effective feature selection using EMFFS and BFS.

Algorithm 2 begins by examining the input dataset D, which contains n typical features. First, the EMFFS which consists of five separate base filters: IG, SD, CHI, GR, and GI is used to rank and pick the top-k features. Based on MVT, these top-k features are then integrated and common features are obtained. These common features are then fed into the BFS method for refinement. The BFS begins with all of these features and removes the least significant feature at each iteration, improving the classifier’s accuracy. As the optimal features, the feature subset that increases classification performance in each iteration is chosen. This procedure is repeated until no change in feature elimination is detected. Once we have a final subset of optimum/effective sentiment features, classification algorithms are then used to compute accuracy using 5-fold cross-validation.

H. CLASSIFICATION ALGORITHMS AND ENSEMBLE LEARNING METHOD

In this section, we briefly discuss the classification algorithms with ensemble learning method for sentiment classification.

Classification Algorithms: Supervised learning trains models to predict the optimal outcome using a training set. In a variety of traditional text classification and sentiment classification tasks, supervised ML classifiers have been commonly used. In this study, we use three well-known ML classifiers for
Algorithm 2 Effective sentiment features selection using hybrid of filters and backward feature selection.

Input: // Dataset D with n feat.
17 $D: d_1, d_2, d_3, \ldots, d_n$
   // set of statistical filters F
18 $F: IG, SD, CHI, GR, GI$
19 $\lambda$: threshold
20 $t, k$: // Top K

Output: // Effective Sentiment Features (ESF), Predicted Accuracy (AC)
21 $ESF, AC$;
22 RF ← new Set();
23 foreach $f$ in $F$ do do
   // Apply statistical filter $f$ on dataset $D$
24 $X_F ← rankFeatures(D, F)$;
   // Sort the ranked features in descending order
25 $X_F ← sortDESC(X_F)$;
   // Select the top-k features
26 $T_F ← selectTopKFeatures(X_F, k)$;
   // Add the top-K features to the feature set RF
27 $RF.add(T_F)$;
   // Use majority voting to choose common features with $\lambda \geq t$
28 $C_F ← majorityVoting(R_F, \lambda \geq t)$
   // Select highest rank features
29 $HR_C_F ← selectFeatures(C_F, HR)$
   // Apply backward elimination and select effective sentiment features with high performance
30 $HR_C_F ← backwardElimination(ESF, CV)$
   // apply classification algorithms on dataset using effective sentiment features set ESF
31 $ESF ← classificationAlgos(SVM, NB, GLM)$
   // Calculate accuracy of classification algorithms using 5-fold cross validation
32 Compute Accuracy (AC);
33 return $AC$

review text sentiment classification: SVM, NB, and GLM.

Support Vector Machine (SVM): SVM [70] is one of the popular supervised ML classifiers that is commonly utilized for classification problems. SVM has been successfully applied in text categorization [71] due to its robustness in high dimensional space and has shown good performance in SA [13], [33], [34]. The aim of SVM is to divide hyperplane in the search space that can best separate the group of instances in one class from another class. Given a training set of $n$ instances in the form $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i$ mean the training instance, labeled with class $y_i$, which is either 1 or -1. We want to determine the "maximum-margin hyperplane" that divides the positive and negative training instances with maximum margin. In this study, we select SVM linear for training and testing.

Naive Bayes classifier (NB): The NB classifier is a basic supervised ML classifier based on the Bayes theorem and naive independence assumptions between features. The NB classifier assumes that the features are conditionally independent of one another provided some class. Given a problem instance represented by a vector $X = (x_1, \ldots, x_n)$, where $n$ denotes the number of features. The probability that an instance $X_i$ belongs to a class $C_k$ is calculated using Bayes’ theorem, as given in below equation 8.

$$P(C_k|X) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i|C_k)}{P(X)}$$ (8)

Where $P(x_i|C_k)$ represents the conditional probability that features $x_i$ belongs to class $C_k$ and $P(C_k)$ represents the prior probability that the instance exists in class $C_k$. Because of its positive results, the NB classifier has been widely used for text classification [72]–[74]. In this study, we employed Multinomial NB classifier [75], [76]. Multinomial NB evaluate the conditional probability of a feature/word given a class as the frequency related to feature $f$ in the documents belongs to class $c$.

Generalized Linear Model (GLM): The GLM [77], [78] is an extension of general linear model. The GLM fits generalized linear models to the data by maximizing the log-likelihood. The GLM is a flexible, robust and highly interpretable model provide useful approach to modeling classification [79].

The ensemble learning method: The ensemble learning approach incorporates the outputs of multiple base classifiers to boost generalization ability/robustness and classification accuracy over a single classifier. It also reduces the rate of error in classification tasks, which may occur with single classifiers. The use of ensemble methodology has been considered by researchers from various disciplines [80]. To perform sentiment classification, we use a combination of classifiers in this process. As base classifiers in the mix, SVM, NB, and GLM are included. The user’s review text is divided into
positive and negative classes based on the majority voting of base classifiers.

IV. EXPERIMENTAL STUDY

A. SENTIMENT ANALYSIS DATASETS

In this paper, we test the efficacy of our proposed method on two commonly used public SA datasets: the Cornell movie review dataset [13] and the Amazon product review datasets [81].

The Cornell movie review dataset contains 2000 ratings, 1000 of which are positive and 1000 of which are negative. The Amazon product reviews datasets are divided into four categories: books, DVDs, electronics, and kitchen, with 1000 positive and 1000 negative reviews in each category. Table 6 lists the dataset’s specifics.

B. EVALUATION MEASURES

To assess the predictive performance of our proposed hybrid method, we used different evaluation measures: Accuracy (ACC), Precision (PRE), Recall (REC), F-measure, receiver operating characteristic (ROC) curve, area under the curve (AUC), and precision-recall curve (PR curve). The ratio of true positive (TP) and true negative (TN) instances to the total number of instances attained by the classification algorithm is known as classification accuracy.

\[
ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{9}
\]

Precision is the ratio of true positives against the total number of true positives and false positives.

\[
PRE = \frac{TP}{TP + FP} \tag{10}
\]

Recall is the ratio of true positives against the total number of true positives and false negatives.

\[
REC = \frac{TP}{TP + FN} \tag{11}
\]

The F-measure is the harmonic mean of precision and recall.

\[
F - \text{Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}
\]

ROC curve is a performance measurement for the classification problems at distinct threshold settings. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis. A precision-recall curve (or PR curve) is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds. These measuring metrics are widespread in machine learning, and they have been well used in a number of different studies [82]–[84].

For all of the experiments in multi-domain datasets, we used 5-fold cross-validation, which is similar to state-of-the-art SA approaches. In 5-fold cross-validation, each domain dataset was randomly split into a training set and a testing set, with one set used for testing and the rest for training. The averaged results of 5-fold cross-validation are used to estimate classification performance. After building the model, we also used an external validation set to assess our model’s predictive performance on unseen data. To calculate the evaluation measures of the proposed solution, we used the paired t-test (P<0.05).

C. EXPERIMENTAL SETTING

All of the experiments were carried out on a PC with an Intel(R) Core(TM) i7 processor, 16.0 GB of RAM, and the Windows 10 64-bit operating system. We used Rapidminer Studio [85] for the proposed algorithm implementation, which is a complete environment for ML, text mining, deep learning, data processing, and predictive analytics. We utilized SVM, NB, GLM, as base classifiers in conjunction with Classifier Ensemble (CE). The hyperparameters along with the optimal values are given in Table 7.

D. EXPERIMENTAL RESULTS AND DISCUSSION

On multi-domain datasets, we evaluate the results obtained by different techniques used in the proposed approach in this section. We also analyze and
discuss how the proposed hybrid approach (IHFFS) could intelligently address the sentiment analysis issue. All of the features were represented using the POS tagged n-gram features known as POS unigram and bigram (Adjective, Adverb, Verb, and Noun). All the experimental results were conducted through 5-fold cross-validation. In the CE, three separate base classifiers (SVM, NB, and GLM) were used. IG, SD, CHI, GR, and GI are the baseline statistical filters of EMFFS. BFS is used as a wrapper for the final feature refinement and feature selection. The different proportions of features are evaluated using average classification accuracies. Besides, the proposed method is compared against baseline filters and EMFFS.

E. PERFORMANCE ANALYSIS OF STATISTICAL FILTERS WITH CLASSIFICATION ALGORITHMS

We compare the predictive performance of various filter methods (IG, SD, CHI, GR, and GI) together with classification algorithms (SVM, NB, GLM, and CE) used for SA in this section. For each domain dataset, we first performed different experiments by selecting the top-k high-ranked score features (300 ~1000) for each filter. These features are then used to train classification algorithms. Finally, based on the input feature space (top-k high ranked score features), the trained algorithms’ accuracy is assessed. Table 11 shows the effects of statistical filters with classification algorithms for various proportions of top-k high-ranked features. The experimental results of these methods gave us the following intuition. In terms of individual filter feature selection methods, IG-based filter feature selection achieves high predictive performance. GR, GI, CHI, and SD all have acceptable predictive performance. In terms of classification performance on the selected input feature space (top-k high ranked features), SVM outperformed NB, particularly on the Movie dataset. On other domain datasets (Book, DVD, Electronics, and Kitchen), however, NB outperforms SVM in terms of predictive performance. In comparison to SVM and NB, GLM has a low performance score. The CE outperformed among base classifiers (SVM, NB, and GLM). The experimental results show that using ensemble learning methods (Ensemble of filters and classifiers) to extract and pick appropriate and suitable features has a statistically significant effect on predictive performance.

G. PERFORMANCE ANALYSIS OF HYBRID FEATURE SELECTION APPROACH

We examine the overall performance of the proposed IHFSSA approach for textual SA in this section. To obtain the most effective refine sentiment features for SA, we fused the filter-based EMFFS with the wrapper-based BFS. The EMFFS is used to obtain a common high-rank feature set, which is then fed into the BFS for refinement in the second stage of the proposed hybrid approach. All the parameters of BFS are set at their default values in Rapidminer Studio. The results of the proposed IHFSSA approach for textual SA are shown in Table 9. For the domains of Movie, Book, DVD, Electronics, and Kitchen datasets, the final effective refined features generated by the proposed approach are 896, 862, 910, 775, and 824, respectively. SVM achieved a high classification performance on refined features space in multi-domain datasets, especially on the Movie dataset, and NB obtained the second-highest classification performance. However, the predictive performance of NB on other domain datasets (Book, DVD, Electronics, and Kitchen) is high. SVM comes in second. The performance of GLM is low compared to SVM and NB. The CE based on majority voting is also utilized and obtained the best results. Table 9 shows that the feature space has decreased substantially and classification performance has improved significantly.
Table 8: Sentiment analysis results achieved by ensemble of multiple filters feature selection with classification algorithm on each domain

| Method | Dataset | Classifier | Common high ranked features | Precision (%) | Recall (%) | F-measure (%) | Accuracy (%) |
|--------|---------|------------|-----------------------------|--------------|-----------|---------------|--------------|
|        |         | SVM        | 945                         | 94.00        | 93.07     | 93.53         | 93.50        |
|        |         | NB         | 945                         | 92.50        | 91.93     | 92.21         | 92.19        |
|        |         | GLM        | 945                         | 91.88        | 90.74     | 91.30         | 91.25        |
|        |         | CE         | 945                         | 93.88        | 93.76     | 93.82         | 93.81        |
|        |         | SVM        | 950                         | 91.63        | 90.94     | 91.28         | 91.25        |
|        |         | NB         | 950                         | 92.00        | 90.86     | 91.43         | 91.38        |
|        |         | GLM        | 950                         | 91.25        | 90.68     | 90.97         | 90.94        |
|        |         | CE         | 950                         | 91.63        | 91.51     | 91.57         | 91.51        |
|        |         | SVM        | 968                         | 90.13        | 90.01     | 90.07         | 90.06        |
|        |         | NB         | 968                         | 90.38        | 90.26     | 90.32         | 90.31        |
|        |         | GLM        | 968                         | 89.88        | 89.76     | 89.82         | 89.81        |
|        |         | CE         | 968                         | 90.63        | 90.57     | 90.56         | 90.56        |
|        |         | SVM        | 916                         | 90.00        | 88.67     | 89.33         | 89.25        |
|        |         | NB         | 916                         | 90.63        | 88.41     | 89.51         | 89.38        |
|        |         | GLM        | 916                         | 88.13        | 88.01     | 88.07         | 88.06        |
|        |         | CE         | 916                         | 90.75        | 89.97     | 89.85         | 89.75        |
|        |         | SVM        | 920                         | 92.50        | 91.13     | 91.81         | 91.75        |
|        |         | NB         | 920                         | 92.50        | 92.04     | 92.27         | 92.25        |
|        |         | GLM        | 920                         | 90.13        | 90.01     | 90.07         | 90.06        |
|        |         | CE         | 920                         | 92.50        | 92.50     | 92.50         | 92.50        |

Table 9: Sentiment analysis results achieved by the proposed Hybrid Feature Selection approach

| Method | Dataset | Classifier | Best Features | Precision (%) | Recall (%) | F-measure (%) | Accuracy (%) |
|--------|---------|------------|---------------|--------------|-----------|---------------|--------------|
|        |         | SVM        | 896           | 95.00        | 94.06     | 94.53         | 94.50        |
|        |         | NB         | 896           | 94.88        | 93.82     | 94.24         | 94.31        |
|        |         | GLM        | 896           | 93.13        | 92.66     | 92.89         | 92.88        |
|        |         | CE         | 896           | 95.00        | 94.29     | 94.65         | 94.63        |
|        |         | SVM        | 862           | 92.00        | 91.89     | 91.94         | 91.94        |
|        |         | NB         | 862           | 92.50        | 91.25     | 91.75         | 91.81        |
|        |         | GLM        | 862           | 93.13        | 91.98     | 92.55         | 92.50        |
|        |         | CE         | 862           | 92.00        | 90.98     | 91.49         | 91.44        |
|        |         | SVM        | 910           | 92.75        | 91.54     | 91.79         | 91.75        |
|        |         | NB         | 910           | 91.50        | 90.82     | 91.16         | 91.13        |
|        |         | GLM        | 910           | 92.63        | 91.59     | 92.11         | 92.06        |
|        |         | CE         | 910           | 91.25        | 91.14     | 91.19         | 91.19        |
|        |         | SVM        | 775           | 91.25        | 91.36     | 91.93         | 91.88        |
|        |         | NB         | 775           | 91.63        | 90.16     | 90.89         | 90.81        |
|        |         | GLM        | 775           | 92.50        | 91.70     | 92.10         | 92.06        |
|        |         | CE         | 775           | 93.65        | 92.01     | 92.81         | 92.75        |
|        |         | SVM        | 824           | 94.38        | 92.19     | 93.27         | 93.19        |
|        |         | NB         | 824           | 94.38        | 92.19     | 93.27         | 93.19        |
|        |         | GLM        | 824           | 92.50        | 91.47     | 91.98         | 91.94        |
|        |         | CE         | 824           | 94.38        | 92.64     | 93.50         | 93.44        |

Table 10: Performance comparison of proposed approach with various state-of-the-art sentiment analysis approaches using movie and product review datasets

| Author/Year | Method | Dataset | Best Accuracy (%) |
|-------------|--------|---------|-------------------|
| Pang et al. [13] (2002) | Sentiment classification using n-grams | Movie Reviews | 82.9 |
| Mullen et al. [39] (2004) | Hybrid SVMs with unigrams | Movie Reviews | 86.0 |
| Ng et al. [17] (2006) | Learning-based polarity classification | Movie Reviews | 90.3 |
| Matsumoto et al. [40] (2005) | Unigrams with SVM | Movie Reviews | 92.9 |
| Tripathy et al. [11] (2016) | Sentiment classification using n-grams | Large Movie Reviews | 88.94 |
| Xia et al. [12] (2011) | Ensemble for sentiment classification | Movie, Book, Electronic, Kitchen, DVD | 87.70, 81.80, 85.95, 88.65, 83.80 |
| Alireza et al. [19] (2017) | Hybrid filter and wrapper | Movie, Book, Electronic, Kitchen, Music | 92.92, 85.13, 85.97, 86.62, 85.64 |
| Osman et al. [4] (2020) | Iterated greedy metaheuristic | Book, Electronics, Kitchen, DVD | 88.00, 91.63, 93.09, 90.25 |
| Jing et al. [15] (2020) | Modified categorical proportional difference | Movie | 93.59, 92.13, 91.59, 91.65, 93.21 |
| Proposed Approach | IHFSSA | Movie, Book, Electronic, Kitchen, DVD | 94.55, 92.13, 91.59, 91.65, 93.21 |
proposed hybrid feature selection method, which
MVT, outperformed IG and other base filters. The
and SD. The EMFFS method, which combines
Figure 4), the IG outperformed the CHI, GR, GI,
and classification accuracy of each filter (as seen in
Figure 4), the IG outperformed the CHI, GR, GI,
and SD. The EMFFS method, which combines
Figure 4 summarize the experimental results (Ta-
ble 8, 9, and 11) of various feature selection tech-
niques on multi-domain datasets. To compare the
classification accuracy of each filter (as seen in
Figure 4), the IG outperformed the CHI, GR, GI,
and SD. The EMFFS method, which combines
the aforementioned individual base filters based on
MVT, outperformed IG and other base filters. The
proposed hybrid feature selection method, which
combines EMFFS and BFS, has the highest classifi-
cation accuracy of all techniques.

Regarding dimensionality reduction, the pro-
posed hybrid feature selection approach highly re-
duced the feature space while obtained high accu-
...
for the domains of Book, DVD, Electronics, and Kitchen, the IHFSSA reduced and selected the best feature set. The reduced feature space for Book, DVD, Electronics, and Kitchen datasets are 862, 910, 775, and 824, respectively. In terms of high dimensional feature space reduction, IHFSSA significantly reduced the feature space and achieved the highest classification accuracy as compared to state-of-the-art methods [13], [19]. In [19], the two-hybrid feature selection methods, namely FIFS and OIFV have been proposed for SA. The FIFS method reduced the average feature space to 10,179 and 2325 on the movie and kitchen datasets, respectively. The OIFV method reduced the average feature space to 4201 and 3427 for Book and electronics dataset, respectively. In [13] the reduced feature space on the movie domain dataset is 2633.

I. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART APPROACHES

Over five benchmark datasets, we compare the classification performance of our proposed solution IHFSSA in terms of best accuracy with state-of-the-art feature selection sentiment classification approaches. Table 10 and Figure 5 provide a quantitative performance analysis across benchmark datasets. Figure 5 along with the table presents the accuracy rate in percentage for each method. For sentiment classification, Xia et al. [12] presented two kinds of feature sets: POS-based feature sets and word relation-based feature sets. They used three different ensembles methods, namely Fixed combination, weighted combination, and meta classifier combination. Word-relation based ensemble method was found best in terms of classification accuracy. Alireza et al. [19] proposed two hybrids of filter and wrapper methods for SA: frequency-based integration of different feature subsets (FIFS) and ordinal-based integration of different feature vectors (OIFV). FIFS was obtained high accuracy on Movie and Kitchen datasets, while OIFV was achieved high accuracy on Book and Electronics datasets. Osman et al. [4] proposed a wrapper feature selection algorithm based on Iterated Greedy metaheuristic for sentiment classification. They reported the high accuracy for Electronics and Kitchen datasets and obtained comparable results for the Book and DVD datasets. On five separate datasets, we tested our proposed solution IHFSSA against state-of-the-art approaches. As seen in Table 10 and Figure 5, the proposed IHFSSA approach outperformed state-of-the-art feature selection sentiment classification methods. There are several explanations why the proposed solution IHFSSA achieves the best and comparable results. The first explanation is that during text pre-processing, noisy and meaningless features are removed from the textual data. The extraction and selection of effective sentiment features is the second explanation. The third explanation is to classify mixed-opinionated texts using linguistic rules and semantic information. The inclusion of WCSL for sentiment features identification is the fourth reason. The fifth reason is the integration of EMFFS with BFS. The sixth explanation is the prediction of final sentiment orientation based on CE.

We conducted a statistical test based on accuracy results to compare the predictive performance between the two proposed feature selection methods EMFFS and IHFSSA. We ran the IHFSSA in comparison to EMFFS. We employed the paired t-test (P<0.05) to evaluate the performance difference between IHFSSA and EMFFS. Table 11 shows the statistically significant difference between IHFSSA and EMFFS. IHFSSA significantly outperformed than EMFFS in term of classification accuracy on 5 different domain datasets.

J. VALIDATION RESULTS

We used an external validation dataset (400 review documents for each domain i.e., Movie, Book, DVD, Electronics, and Kitchen) to evaluate the predictive performance of our trained intelligent hybrid model on these unseen domains datasets to confirm its efficiency. The accuracy performance of
### Table 11: Classification accuracy for different proportion of features (300 ~ 1000) obtained by statistical filters in conjunction with classification algorithms.

| Methods | No. of features | Movie Dataset | Book Dataset | DVD Dataset | Electronics Dataset | Kitchen |
|---------|----------------|---------------|--------------|-------------|---------------------|---------|
| SVM     | 300            | 89.90         | 88.44        | 87.75       | 81.81               | 81.87   | 82.31 | 84.50       | 81.69               | 81.23               | 79.30               | 80.30               | 81.50               | 82.00               | 83.25               | 86.10               | 87.01               | 85.13               | 88.50               |
| NB      | 500            | 88.69         | 89.06        | 87.31       | 88.84       | 85.56               | 87.18   | 83.44   | 86.60       | 85.38               | 84.20               | 87.50               | 84.50               | 84.50               | 85.10               | 82.80               | 84.76               | 88.35               | 89.20               | 86.75               | 89.40               |
| GLM     | 600            | 89.50         | 89.98        | 87.62       | 89.80       | 88.12               | 86.83   | 86.00   | 89.20       | 85.80               | 84.70               | 83.30               | 86.90               | 85.70               | 86.95               | 84.00               | 87.10               | 89.52               | 90.80               | 87.06               | 89.90               |
| IG      | 700            | 90.75         | 91.81        | 88.25       | 90.93       | 93.50               | 90.94   | 87.56   | 90.60       | 87.87               | 86.50               | 85.80               | 88.70               | 87.30               | 88.20               | 86.80               | 88.60               | 90.50               | 91.60               | 90.90               | 91.50               |
| CHI     | 800            | 89.80         | 88.85        | 86.90       | 90.93       | 88.57               | 90.79   | 86.61   | 87.84       | 88.22               | 86.87               | 88.47               | 87.84               | 88.22               | 87.84               | 88.47               | 88.22               | 91.52               | 91.62               | 89.84               | 91.57               |
| Average | 89.60         | 89.15        | 87.31        | 88.82       | 87.81       | 86.13               | 85.85   | 83.85   | 85.35       | 85.57               | 84.90               | 82.17               | 86.87               | 85.39               | 86.21               | 83.84               | 86.45               | 89.13               | 90.21               | 89.75               | 90.36               |
| NB      | 900            | 87.84         | 88.98        | 86.57       | 89.97       | 90.55               | 89.97   | 86.64   | 87.87       | 86.57               | 85.67               | 85.87               | 87.35               | 88.27               | 86.53               | 86.56               | 85.56               | 90.14               | 91.64               | 89.53               | 91.71               |

| Average | 86.83       | 89.79        | 84.75        | 87.23       | 85.82               | 85.62               | 85.73   | 86.29   | 85.57       | 82.85               | 81.02               | 83.32               | 84.83               | 84.91               | 82.66               | 85.08               | 88.28               | 88.74               | 87.17               | 89.85               |
| NB      | 1000          | 86.94         | 88.55        | 84.00        | 87.15       | 81.50               | 83.80   | 82.30   | 84.10       | 80.50               | 81.00               | 79.10               | 81.80               | 81.40               | 82.10               | 80.20               | 80.70               | 86.00               | 87.05               | 85.70               | 87.30               |
| GLM     | 1100          | 86.44         | 86.87        | 86.50        | 89.97       | 84.00               | 85.50   | 85.25   | 86.50       | 85.35               | 84.35               | 86.60               | 85.90               | 85.60               | 85.90               | 85.40               | 85.80               | 85.90               | 91.10               | 87.80               | 89.50               |
| IG      | 1200          | 86.91         | 89.40        | 89.70       | 90.20       | 89.60               | 89.70   | 86.90   | 87.60       | 86.50               | 87.20               | 84.25               | 86.70               | 87.30               | 88.05               | 86.40               | 86.70               | 88.40               | 90.49               | 91.50               | 91.60               |
| Average | 86.59       | 89.87        | 87.26        | 90.25       | 87.09       | 87.97               | 85.87   | 88.18   | 84.21       | 84.89               | 82.56               | 85.02               | 85.30               | 85.94               | 86.45               | 86.81               | 89.94               | 90.80               | 88.02               | 90.17               |

### Table 12: The P-values of paired t-test that compares IHFSSA method with baseline method EMFFS.

| Dataset | testFeatures | Classification Algorithm | Accuracy (%) | EMFFS | Accuracy (%) | IHFSSA | T-test P-values |
|---------|--------------|--------------------------|--------------|-------|--------------|--------|----------------|
| Movie   | 896          | GLM                      | 89.90        | 0.020 | 89.88        |        |                |
| Book    | 862          | SVM                      | 90.10        |       | 94.50        |        |                |
| DVD     | 910          | NB                       | 91.75        |       | 0.030        |        |                |
| Electronics | 775        | CE                       | 88.20        |       | 0.006        |        |                |
| Kitchen | 824          | CE                       | 93.44        |       | 0.022        |        |                |

the training and validation datasets is then compared in Figure 6, which indicates that our model was reliable.

### K. ROC CURVE AND PRECISION-RECALL CURVE ANALYSIS

Area under the ROC Curve (or AUC) and PR curve are the well known measurement metrics in machine learning used for evaluating a classifier’s performance. We present these metrics curves of our intelligent hybrid model on each domain dataset. Figure 7 shows the area under ROC of the proposed intelligent hybrid model on the datasets Movie, Book, DVD, Electronics and Kitchen domains. The corresponding Area Under Curve (AUC) values obtained by the proposed model for these domain datasets are 0.955, 0.941, 0.947, 0.946, and 0.945 respectively. While Figure 8 shows the PR curves and corresponding Area Under PR Curve values obtained by proposed model on the datasets Movie, Book, DVD, Electronics and Kitchen domains are 0.952, 0.927, 0.944, 0.941, and 0.932 respectively. The improved area of the ROC curve and PR curves proves that the proposed intelligent model works well and improves performance with the minimum optimum feature set.

### V. CONCLUSION

A gigantic amount of unstructured user-generated data in the form of textual comments and reviews

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/ACCESS.2021.3118982, IEEE Access
exist online. This unstructured data contains valuable information for business analysis and product purchase decisions. SA is the fastest growing research area of text mining and web mining, which goals to classify online unstructured user-generated data into positive and negative classes. Due to the huge amount and unstructured nature of textual data, state-of-the-art techniques for SA suffers from noisy and irrelevant features. Likewise, these techniques could not identify the hidden significant information in the unstructured data efficiently. Moreover, the extraction and selection of the most effective sen-

Figure 7: ROC curve and area Under the ROC curve obtained by the proposed model for (a) Movie (b) Book (c) DVD (d) Electronics and (e) Kitchen datasets.
sentiment features are not handled successfully. In this respect, the extraction and selection of suitable and relevant sentiment features are significant for SA. Also to deal with high dimensional feature space, an intelligent feature selection technique is essential for accurate sentiment classification. In this study, an intelligent hybrid feature selection approach, which we called IHFSSA, is proposed for textual SA. The proposed approach enables us to identify, refine and select the most effective sentiment features with the goals of decreasing high dimensional feature space, taking into account sentiment information, semantic aspects, and word order for textual sentiment classification. The results of the experimental study show that the proposed hybrid approach outperformed state-of-the-art baseline approaches. In future work, we will incorporate deep learning, pre-trained NLP models such as BERT, GPT3, ELMo, and XLNet, etc. for testing IHFSSA.

Figure 8: PR curve and area under PR curve obtained by the proposed model for (a) Movie (b) Book (c) DVD (d) Electronics and (d) Kitchen datasets.
References

[1] M. M. Skoric, J. Liu, and K. Jaidka, “Electoral and public opinion forecasts with social media data: A meta-analysis,” Information, vol. 11, no. 4, p. 187, 2020.

[2] C. GUO and K. CAINE, “Anonymity, user engagement, quality, and trolling on q&a sites,” 2021.

[3] M. K. Khan and N. Rahman, “Study of users attitude and classification of comments and likes from facebook using rapidminer,” in Data Intelligence and Cognitive Informatics, Springer, 2021, pp. 673–687.

[4] O. Gokalp, E. Tasci, and A. Ugur, “A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification,” Expert Systems with Applications, vol. 146, p. 113176, 2020.

[5] Z. Shaukat, A. A. Zulfikar, C. Xiao, M. Azeem, and T. Mahmod, “Sentiment analysis on imdb using lexicon and neural networks,” SN Applied Sciences, vol. 2, no. 2, pp. 1–10, 2020.

[6] E. E. Stömmez and H. Çakir, “Effect of web 2.0 technologies on academic performance: A meta-analysis study,” International Journal of Technology in Education and Science, vol. 5, no. 1, pp. 108–127, 2021.

[7] M. Mariani, “Web 2.0 and destination marketing: Current trends and future directions,” Sustainability, vol. 12, no. 9, p. 3771, 2020.

[8] B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.

[9] Z. Zhai, H. Xu, B. Kang, and P. Jia, “Exploiting effective features for chinese sentiment classification,” Expert Systems with Applications, vol. 38, no. 8, pp. 9139–9140, 2011.

[10] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams engineering journal, vol. 5, no. 4, pp. 1093–1113, 2014.

[11] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of sentiment reviews using n-gram machine learning approach,” Expert Systems with Applications, vol. 57, pp. 117–126, 2016.

[12] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences, vol. 181, no. 6, pp. 1138–1152, 2011.

[13] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[14] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, pp. 437–441.

[15] J.-R. Chang, H.-Y. Liang, L.-S. Chen, and C.-W. Chang, “Novel feature selection approaches for improving the performance of sentiment classification,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–14, 2020.

[16] A. Onan and S. Koruküçü, “A feature selection model based on genetic rank aggregation for text sentiment classification,” Journal of Information Science, vol. 43, no. 1, pp. 25–38, 2017.

[17] V. Ng, S. Dasgupta, and S. Aafroz, “Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews,” in Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006, pp. 611–618.

[18] V. S. Subrahmanian and D. Reforgiato, “Ava: Adjective-verb-adverb combinations for sentiment analysis,” IEEE Intelligent Systems, vol. 23, no. 4, pp. 43–50, 2008.

[19] A. Yousefpour, R. Ibrahim, and H. N. A. Hamed, “Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis,” Expert Systems with Applications, vol. 75, pp. 80–93, 2017.

[20] J. Khan, A. Alam, J. Hussain, and Y.-K. Lee, “EnSWF: effective features extraction and selection in conjunction with ensemble learning methods for document sentiment classification,” Applied Intelligence, vol. 49, no. 8, pp. 3123–3145, 2019.

[21] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization.”

[22] X. Deng, Y. Li, J. Weng, and J. Zhang, “Feature selection for text classification: A review,” Multimedia Tools and Applications, vol. 78, no. 3, pp. 3797–3816, 2019.

[23] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zienjarfa, “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization,” Procedia Engineering, vol. 53, pp. 453–462, 2013.

[24] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, pp. 168–177.

[25] J. Khan and B. S. Jeong, “Summarizing customer review based on product feature and opinion,” in Machine Learning and Cybernetics (ICMLC), 2016 International Conference on, vol. 1. IEEE, 2016, pp. 158–165.

[26] R. V. A. Ogutu, R. Rimuru, and C. Otieno, “Target sentiment analysis model with naive bayes and support vector machine for product review classification,” Int. J. Comput. Sci. Inf. Secur.(IJCSIS), vol. 17, no. 7, pp. 1–17, 2019.

[27] A. Jones, J. Ellman, and N. Jin, “An application of sentiment analysis techniques to determine public opinion in social media.” International Conference on Information Society, 2019.

[28] S. Kannagara and W. Wobcke, “Weakly supervised joint entity-sentiment-issue model for political opinion mining,” in Pacific Rim International Conference on Artificial Intelligence. Springer, 2019, pp. 568–581.

[29] R. Ren, D. D. Wu, and T. Liu, “Forecasting stock market movement direction using sentiment analysis and support vector machine,” IEEE Systems Journal, vol. 13, no. 1, pp. 760–770, 2018.

[30] S. Bharathi and A. Geetha, “Sentiment analysis for effective stock market prediction,” International Journal of Intelligent Engineering and Systems, vol. 10, no. 3, pp. 145–154, 2017.

[31] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N project report, Stanford, vol. 1, no. 12, p. 2009, 2009.

[32] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, “Sentiment classification of internet restaurant reviews written in cantonese,” Expert Systems with Applications, vol. 38, no. 6, pp. 7674–7682, 2011.

[33] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, “Experiments with svm to classify opinions in different domains,” Expert Systems with Applications, vol. 38, no. 12, pp. 14799–14804, 2011.

[34] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between svm and ann,” Expert Systems with Applications, vol. 40, no. 2, pp. 621–633, 2013.

[35] P. K. Singh and M. S. Hussain, “Methodological study of opinion mining and sentiment analysis techniques,” International Journal on Soft Computing, vol. 5, no. 1, p. 11, 2014.

[36] X. Bai, “Predicting consumer sentiments from online text,” Decision Support Systems, vol. 50, no. 4, pp. 732–742, 2011.

[37] P. Kalmivuori and K. Shunnugananthan, “Sentiment classification of movie reviews by supervised machine learning approaches,” Indian Journal of Computer Science and Engineering, vol. 4, no. 4, pp. 283–292, 2013.

[38] M. Bilal, H. Israr, M. Shahid, and A. Khan, “Sentiment classification of roman-urdu opinions using naive bayesian, decision tree and knn classification techniques,” Journal of King Saud University-Computer and Information Sciences, vol. 28, no. 3, pp. 330–344, 2016.

[39] F. X. Yang, L. Guo, G. Yanyan, and W. Zhiqiang, “Multi-aspect sentiment analysis for chinese online social re-
views based on topic modeling and hownet lexicon,” Knowledge-Based Systems, vol. 37, pp. 186–195, 2013.

[40] T. Mullen and N. Collier, “Sentiment analysis using support vector machines with diverse information sources,” in Proceedings of the 2004 conference on empirical methods in natural language processing, 2004.

[41] S. Matsamoto, H. Takamura, and M. Okumura, “Sentiment classification using word sub-sequences and dependency sub-trees,” in Pacific-Asia conference on knowledge discovery and data mining, Springer, 2005, pp. 301–311.

[42] S. Tan and J. Zhang, “An empirical study of sentiment analysis for chinese documents,” Expert Systems with applications, vol. 34, no. 4, pp. 2622–2629, 2008.

[43] A. Sharma and S. Dey, “A comparative study of feature selection and machine learning techniques for sentiment analysis,” in Proceedings of the 2012 ACM research in applied computation symposium, 2012, pp. 1–7.

[44] C. Catal and M. Nangir, “A sentiment classification model based on multiple classifiers,” Applied Soft Computing, vol. 50, pp. 135–141, 2017.

[45] S. Li, C. Zong, and X. Wang, “Sentiment classification through combining classifiers with multiple feature sets,” in Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on, IEEE, 2007, pp. 135–140.

[46] K. Tsutsui, K. Shimada, and T. Endo, “Movie review classification based on a multiple classifier,” in Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation, 2007, pp. 481–488.

[47] Y. Yan and Q. Gao, “An ensemble sentiment classification system of twitter data for airline services analysis,” in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, IEEE, 2015, pp. 1318–1325.

[48] H. Wang, T. M. Khoshgoftaar, and J. Van Hulse, “A comparative study of threshold-based feature selection techniques,” in Granular Computing (GrC), 2010 IEEE International Conference on, IEEE, 2010, pp. 499–504.

[49] P. Kalavani and K. Shanmuganathan, “Feature reduction based on genetic algorithm and hybrid model for opinion mining,” Scientific Programming, vol. 2015, 2015.

[50] M. U. Salur and I. Aydin, “A novel hybrid deep learning model for sentiment classification,” IEEE Access, vol. 8, pp. 58 080–58 093, 2020.

[51] W. Zhao, H. Peng, S. Eger, E. Cambria, and M. Yang, “Towards scalable and reliable capsule networks for challenging nlp applications,” arXiv preprint arXiv:1906.02829, 2019.

[52] G. Choi, S. Oh, and H. Kim, “Improving document-level sentiment classification using importance of sentences,” Entropy, vol. 22, no. 12, p. 1336, 2020.

[53] J. Khan, B. S. Jeong, Y.-K. Lee, and A. Alam, “Sentiment analysis at sentence level for heterogeneous datasets,” in Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory, ACM, 2017, pp. 159–168.

[54] Y. Choi, Y. Kim, and S.-H. Myaeng, “Domain-specific sentiment analysis using contextual feature generation,” in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, ACM, 2009, pp. 37–44.

[55] A. Aïbèlîl, Treebanks: Building and using parsed corpora, Springer Science & Business Media, 2012, vol. 20.

[56] O. Appel, F. Chiclana, J. Carter, and H. Fujita, “A hybrid approach to the sentiment analysis problem at the sentence level,” Knowledge-Based Systems, vol. 108, pp. 110–124, 2016.

[57] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Hombo, A. Agrawal, and A. N. Choudhary, “Muses: multilingual sentiment elicitation system for social media data,” IEEE Intelligent Systems, vol. 29, no. 4, pp. 34–42, 2013.

[58] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentimentnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining,” in Lrec, vol. 10, no. 2010, 2010, pp. 2200–2204.

[59] C. Strapparava, A. Valitutti et al., “Wordnet affect: an affective extension of wordnet,” in Lrec, vol. 4, no. 1083-1086. Citeseer, 2004, p. 40.

[60] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, “Sentiment 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[61] S. M. Mohammad, S. Kirtichenko, and X. Zhu, “Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets,” arXiv preprint arXiv:1308.6242, 2013.

[62] E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” in Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 105–112.

[63] J. C. De Albornoz, L. Plaza, and P. Gervás, “Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis,” in LREC, vol. 12. Citeseer, 2012, pp. 3562–3567.

[64] P. J. Stone and E. B. Hunt, “A computer approach to content analysis: studies using the general inquirer system,” in Proceedings of the May 21-23, 1963, spring joint computer conference, 1963, pp. 241–256.

[65] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” Computational linguistics, vol. 37, no. 2, pp. 267–307, 2011.

[66] F. A. Nielsen, “A new anew: Evaluation of a word list for sentiment analysis in microblogs,” arXiv preprint arXiv:1103.2903, 2011.

[67] J. Ding and L. Fu, “A hybrid feature selection algorithm based on information gain and sequential forward floating search,” Journal of Intelligent Computing Volume, vol. 9, no. 3, p. 93, 2018.

[68] G. Ansari, T. Ahmad, and M. N. Doja, “Ensemble of feature ranking methods using hesitant fuzzy sets for sentiment classification,” IEEE Access, 2019.

[69] R. P. Priyadarzini, M. Valarmathi, and S. Sivakumari, “Gain ratio based feature selection method for privacy preservation,” ICTACT J Soft Comput, vol. 1, no. 04, pp. 2229–6956, 2011.

[70] C. C. Aggarwal and C. Zhai, Mining text data, Springer Science & Business Media, 2012.

[71] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in European conference on machine learning, Springer, 1998, pp. 137–142.

[72] E. Frank and R. R. Bouckaert, “Naive bayes for text classification with unbalanced classes,” in European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2006, pp. 503–510.

[73] Y. H. Li and A. K. Jain, “Classification of text documents,” The Computer Journal, vol. 41, no. 8, pp. 537–546, 1998.

[74] D. D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval,” in European conference on machine learning, Springer, 1998, pp. 4–15.

[75] A. McCallum, K. Nigam et al., “A comparison of event models for naive bayes text classification,” in AAAI-98 workshop on learning for text categorization, vol. 752, no. 1. Citeseer, 1998, pp. 41–48.

[76] O. Vechtomova, “Introduction to information retrieval christopher d. Manning, prabhakar raghavan, and hinrich schutze (stanford university, yahoo! research, and university of stuttgart) cambridge: Cambridge university press, 2008, xxii+ 482 pp; hardbound, isbn 978-0-521-86571-5.” 2009.

[77] M. Goldburt, A. Khare, and C. D. Tevet, “Generalized linear models for insurance rating,” in Casualty Actuarial Society, 2016.
[78] P. McCullagh, “Generalized linear models,” European Journal of Operational Research, vol. 16, no. 3, pp. 285–292, 1984.
[79] I. Trofimov and A. Genkin, “Distributed coordinate descent for generalized linear models with regularization,” Pattern Recognition and Image Analysis, vol. 27, no. 2, pp. 349–364, 2017.
[80] L. Rokach, “Ensemble-based classifiers,” Artificial Intelligence Review, vol. 33, no. 1, pp. 1–39, 2010.
[81] J. Blitzer, M. Dredze, and F. Pereira, “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,” in Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 440–447.
[82] N. Q. K. Le, T. N. K. Hung, D. T. Do, L. H. T. Lam, L. H. Dang, and T.-T. Huynh, “Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from mri,” Computers in Biology and Medicine, vol. 132, p. 104320, 2021.
[83] S. Balakrishnan, R. Narayanaswamy, N. Savarimuthu, and R. Saminann, “Svm ranking with backward search for feature selection in type ii diabetes databases,” in 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2008, pp. 2628–2633.
[84] A. R. Pathak, M. Pandey, and S. Rautaray, “Topic-level sentiment analysis of social media data using deep learning,” Applied Soft Computing, vol. 108, p. 107440, 2021.
[85] M. Hofmann and R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2013.

DR. JAWAD KHAN received the master’s degree in computer science from Kohat University of Science and Technology (KUST), Pakistan and PhD degree in computer engineering from Kyung Hee University (Global Campus), South Korea. He is currently working as a PostDoc at the Department of Robotics, Hanyang University (ERICA Campus), South Korea. His research interests include natural language processing, information retrieval, sentiment analysis/opinion mining, text processing, social media mining, machine and deep learning.

DR. AFTAB ALAM received his BS, MS, and Ph.D. degrees in computer science from the University of Malakand, KPK, Pakistan, the University of Peshawar, KPK, Pakistan, and Kyung-Hee University, South Korea, respectively. He was working as a Post-doc at the Department of Computer Science and Engineering, Kyung-Hee University (Global Campus), South Korea. Currently, he is a Postdoc at the College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His research interests include big data analytics in the cloud, distributed computing, information retrieval, social computing, machine/deep learning, information service engineering, knowledge engineering, complex event analysis in the multi-stream environment, and graph mining.

DR. YOUNGMOON LEE (Member, IEEE) received the BS degree in electrical and computer engineering from Seoul National University, Korea, in 2014, the MS degree in computer science and engineering from the University of Michigan, Ann Arbor, in 2016, and the PhD degree in computer science and engineering from the University of Michigan, Ann Arbor, in 2019. He is currently an assistant professor with the Department of Robotics, Hanyang University, Korea. His research interests include cyber-physical systems, embedded systems, and mobile computing.

***