Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis

Arwa Diwali, Kawther Saeedi, Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain

Abstract—Sentiment analysis can be used to derive knowledge that is connected to emotions and opinions from textual data generated by people. As computer power has grown, and the availability of benchmark datasets has increased, deep learning models based on deep neural networks have emerged as the dominant approach for sentiment analysis. While these models offer significant advantages, their lack of interpretability poses a major challenge in comprehending the rationale behind their reasoning and prediction processes, leading to complications in the models' explainability. Further, only limited research has been carried out into developing deep learning models that describe their internal functionality and behaviours. In this timely study we carry out a first of its kind overview of key sentiment analysis techniques and explainable artificial intelligence (AI) methodologies that are currently in use. Furthermore, we provide a comprehensive review of sentiment analysis explainability.

Index Terms—Sentiment analysis, Deep Learning, Explainability, Interpretability

- *A. Diwali, K. Dashtipour, M. Gogate and A. Hussain are with the School of Computing, Edinburgh Napier University, Edinburgh, UK. E-mail: {arwa.diwali, k.dashitpour, m.gogate, a.hussain}@napier.ac.uk.*
- *A. Diwali and K. Saeedi are with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, KSA. E-mail: {adiwali, ksaeedi}@kau.edu.sa.*
- *E. Cambria is with Nanyang Technological University, Singapore, cambria@ntu.edu.sg*

(Corresponding author: Arwa Diwali)

1 Introduction

THE growth of social media platforms has significantly increased the quantity of written data that is accessible over the internet. This creates a critical need for sentiment analysis (SA) to discover what feelings or views are being discussed about products, brands, or services. The aim of sentiment analysis, or more specifically polarity detection, is to identify positive, negative, or neutral polarities in a piece of written text. Cambria et al. [1] argued that identifying the sentiment of a written text is a challenging task for humans and will be even more difficult for computers. This is due to the fact that the written data contains a variety of qualities, such as discussing the product as a whole or focusing on a particular aspect. In addition, recognising polarity can be made more difficult by the use of sarcasm, dialects, or multilingual features.

Of the approaches currently in use for sentiment analysis, there are three primary categories: lexicon, machine learning and hybrid-based approaches. The most prevalent of these are machine-learning-based approaches, and in particular,e Deep Neural Network (DNN) models, which include, but are not limited to, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [2]. The level of performance in sentiment analysis research is considered state-of-the-art based on DNN models. However, according to [3], a significant drawback of employing deep learning models for Natural Language Processing (NLP) is a lack of transparency in the reasoning used by these models. So, while they offer precise predictions, human understanding, or interpretation, of their reasoning is extremely limited. Increasing the explainability of these models requires improvement in the transparency of the internal activities they undertake for better understanding of their decisions.

This need for explainability in machine learning models has generated an offshoot of interpretable Artificial Intelligence, recently named eXplainable Artificial Intelligence (XAI) [4]. Deep learning comes under the umbrella of machine learning, which is a subset of Artificial Intelligence (AI). XAI can also be said to come under the AI umbrella. Fig. 1 describes the current models landscape.

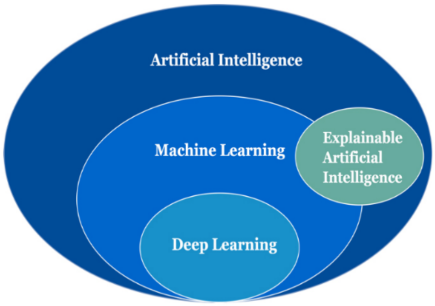


Fig. 1. The current models landscape [5].

As it relates to machine learning, interpretability is described as *"the ability to explain or to present in understandable terms to a human"* [6]. It should be noted that the term "explainability" is linked to the term "interpretability," and many researchers do not differentiate between them, as both terms are frequently used interchangeably. Rudin [7] however, has distinguished between explainable and interpretable machine learning: explainability attempts to explain black-box models by post-hoc justifications, whereas interpretability is concerned with the development of models that are intrinsically capable of being interpreted. Other researchers have proposed different definitions of "explainability".

According to [8], *"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"*. Guidotti et al. [9] have also described explainability as *"an interface between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans."* A recent survey by [10] stated, *"given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand."*

When performance (i.e., prediction accuracy) is the only consideration, the model will become more difficult to understand. This is the case when interpretability of a model is diminished in favor of its performance, as shown in Fig. 2. In contrast, better understanding of a model's decisions may eliminate some of the shortcomings in that model's functionality. High interpretability models include classical regression algorithms, rule-based learning, and decision trees, while deep learning and ensemble methods are examples of low interpretability models. These types of models have weak explainability due to the black-box feature engineering [10].

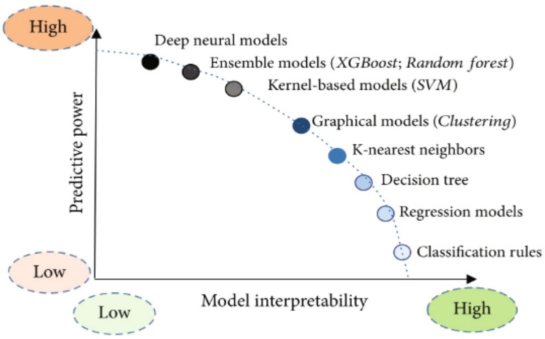


Fig. 2. Trade-off between performance vs. interpretability [11].

There are arguments that can be made in favor of the requirement for XAI. Based on the findings of [12], there is a requirement to justify, control, improve, and discover. It is important that the judgments reached by the models can be explained in order to ensure that they are justifiable. The requirement for explanations raises the level of transparency of a model, making it easier to detect and prevent errors. In addition, explanations contribute to some improvements in terms of the model's accuracy as well as its overall efficiency. Finally, obtaining explanations allows the acquisition of further knowledge and new perspectives regarding the problem at hand [12].

The utilization of XAI approaches can enhance human comprehension of predictions, which can subsequently be used to justify the decisions that are generated by a model. Specifically, when it comes to sentiment analysis, we are able to employ XAI approaches to determine how a given term or phrase that contains several terms influences the sentiment of a particular review. Implementing XAI can increase the model's accountability for its decisions, thereby reducing the likelihood of bias or errors in its outcomes. It can also provide more insights about the underlying factors that contribute to sentiment. Furthermore, certain industries, such as finance or healthcare, may be governed by regulations that require detailed documentation and explanations of model decisions.

In this context, using an explainable sentiment analysis model may help organizations meet these regulatory requirements. As a result, the purpose of explainable sentiment analysis studies is to predict sentiment while also explaining why a given sentiment is assigned to a specific review. These studies have benefited decision makers and researchers by improving sentiment analysis methodologies in sentiment analysis applications and by evaluating current explainable AI technologies.

Considerable research is still required into making sentiment analysis models explainable and into boosting the decision making explainability of those models. The purpose of this paper is to describe explainable models and to provide some understanding of the current implementation of those models in constructing sentiment analysis applications.

This paper has the following structure: Section 2 discusses methodologies, strategies, and algorithms used in sentiment analysis. Section 3 reports the various techniques and methods that are utilized in explainable sentiment analysis. Final thoughts are presented in Section 4.

2 SENTIMENT ANALYSIS

The word sentiment has two distinct meanings in everyday speech: sentiment as feelings or emotions and sentiment as thoughts or opinions. If sentiment perceived as emotion can be retrieved from body language, facial expression, vocal intonation as well as written or spoken text, sentiment perceived as opinion is considerably more allied to written or spoken text. Thus, sentiment analysis of a general written text includes issues with identifying the emotions and opinions stated in that text [13], [14], [15].

According to [3], sentiment analysis is a large suitcase of NLP subproblems and subtasks. One of the more typical tasks associated with sentiment analysis is polarity detection. It can be thought of as the task of evaluating whether the opinion in a piece of writing is positive, negative or neutral. Additionally, at a more granular level, classification based on polarity can be carried out on a text by calculating scores for opinion strength levels with the values of the scores falling within a real range. Within polarity detection, there remains a wide range of sentiment analysis problems to be investigated, such as sentiment reasoning, sarcasm analysis, multimodal sentiment analysis, and so on [16]. This paper focuses mostly on identifying the opinions expressed in a piece of writing.

2.1 Level of Analysis

Three levels of sentiment analysis are possible: document, sentence, and aspect [17]. The following sections go into more detail on these.

2.1.1 Document-Level

At this stage of the process, a document as a whole can be classified as being positive, negative or neutral. The classification of each text is based on how strongly the author feels about a specific thing (e.g., a specific product or service). Documents that evaluate or compare several things are not appropriate for document-level classification. Also, document-level acts best when the document is created by a single author [18].

2.1.2 Sentence-Level

The sentence is the primary concern at this level. Identifying whether a sentence conveys a positive, negative, or neutral sentiment is the primary objective of this level. The sentence must be defined as objective or subjective in order to accomplish this goal [18].

2.1.3 Aspect-Level

Fine-grained analysis is performed at this level in order to uncover feelings pertaining to certain aspects of a thing. This can be seen in the statement: "The Mercedes engine is powerful", where the "engine" is an aspect of "Mercedes", and the review states that it is positive. In order to reach an opinion at this level, it is necessary to identify aspects of the entity. Tasks at this level are beneficial for pinpointing precisely what individuals like or dislike regarding certain aspects of the entity. This is because the entity's features are analyzed rather than the sentiment of the sentences. Sentiment analysis at aspect level relies on the extraction of aspects, and these may be either implicit or explicit [15], [18]

2.2 Sentiment Analysis Approaches

The three main types of approaches in sentiment analysis are: lexicon-based, machine learning-based and their hybrid approaches [19].

2.2.1 Lexicon-Based Approaches

The traditional technique used in sentiment analysis is lexicon-based. A polarity score is generated for each word in lists of words that have been manually classified as having positive or negative polarity using sentiment lexicon methodologies. The overall sentiment score of a particular post is calculated using this newly created lexicon. As part of the process of classifying a sentence, the sentence is broken down into individual words and each is assigned a sentiment score. The total and average scores of a particular sentence can be used to determine its overall sentiment. Since no training data is required for lexicon-based algorithms, they can be categorized as unsupervised approaches [20]. However, an issue with the lexicon-based approach is domain dependency. It occurs when words have multiple meanings and requires that the lexicon used for sentiment analysis is tailored to the domain of interest. Otherwise, the algorithm may ascribe positive polarity to a word that is not positive in the particular domain of the sentence being analyzed and vice versa.

Sentiment lexicons can be developed in two main ways, one based on dictionaries, the other based on corpora. Using the dictionary-based method, a modest collection of sentiment terms is gradually expanded with terms and alternative terms from published dictionaries. This method typically serves broad needs well. Corpus-based lexicons can be customized for particular fields. The starting point for the corpus-based lexicon is a collection of general purpose sentiment terms which is expanded via searches of a domain corpus for additional sentiment terms using co-occurring term patterns [19].

2.2.2 Machine Learning-Based Approaches

In general, after preprocessing, a review post is vectorized into a suitable representation in machine learning-based approaches, either using the traditional bag-of-words representation or more advanced representations such as word embeddings [21].

The three types of machine learning techniques for sentiment analysis tasks are: supervised, unsupervised and semi-supervised learning. Supervised learning algorithms learn from a set of features that are labeled. During training, the correct labels are known, and the process ends when the algorithm performs well. Unsupervised learning, on the other hand, focuses on discovering patterns and structures in the input data. Because of this, the algorithm receives input data that is unlabeled. The semi-supervised learning technique, as suggested by its name, falls between supervised and unsupervised learning. The input data contains labeled and unlabeled examples [22]. Since most sentiment analysis tasks are modelled as classification problems, the most common technique used is supervised learning.

On the basis of the algorithms employed, machine learning algorithms for sentiment analysis can be categorized as either shallow algorithms or deep learning algorithms. During the early development stages of sentiment analysis, shallow machine learning algorithms were the primary tools utilised. Across the history of the development of the field, a variety of algorithms such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB), have been utilized in combination with a range of features like bag-of-words and lexicons as well as syntactic features like Parts Of Speech (POS) [16].

Deep learning has evolved under the umbrella of machine learning as a subfield. It identifies features and learns representations from input data by employing numerous stacked layers of artificial neurons. The goal of deep learning is to tackle complicated problems. Deep learning algorithms derive their functionality from the structure and operation of the human brain. They are able to process massive amounts of raw data, something that was previously impossible with shallow machine learning algorithms [23]. As a result, neural networks are replacing, or at least improving, shallow machine learning algorithms [24].

The most popular deep learning technique used for sentiment analysis is supervised learning, in which each instance is given a label with a polarity or other defining category for learning. In sentiment analysis, numerous DNN architectures have been designed and have had satisfactory results, including Simple Recurrent Network (SRN), Gated Recurrent Unit (GRU), Long-Short Term Memory (LSTM), Bidirectional Long-Short Term Memory (Bi-LSTM), and Convolutional Neural Network (CNN) [2], [25].

As computational power has developed, as deep models have been introduced and as training capabilities have continuously improved, the architecture of Pre-Trained Models (PTMs) has evolved from surface to deep. Universal language representations can be learned in PTMs by using a large corpus. These representations can save time by avoiding the need to start from scratch when performing NLP tasks like sentiment analysis [26]. Pre-trained representations are divided into contextual and context-free categories. First-generation PTMs typically convert each word into a vector known as a word embedding [26] such as GloVe [27] and Skip-Gram [28]. These word embeddings are capable of learning the words' meanings. However, they are context-independent and unable to understand syntactic structure or semantic roles.

For instance, the word embedding for "bank" is represented as a single vector for bank deposit and riverbank. In contrast, second-generation PTMs like ELMo [29], OpenAI [30], and BERT [31], try to learn contextual word embeddings, which build representations for all the words in a sentence based on the rest of the words in that sentence.

BERT, Bidirectional Encoder Representations from Transformers, a cutting-edge language model for natural language preprocessing, was created at Google AI Language [31]. BERT has shown breakthrough results on sentiment analysis, natural language inference, question analysis and other NLP tasks.

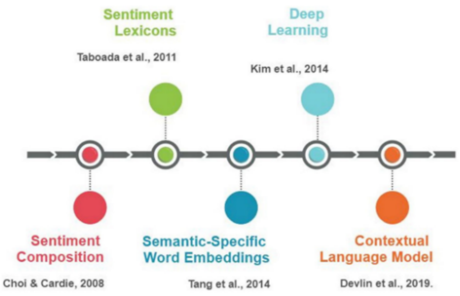


Fig. 3. Sentiment analysis research timeline [19].

The BERT framework has a pre-training phase and a fine-tuning phase. The pre-training phase is unsupervised, and model training is conducted using a huge unlabeled corpus in which the model masks words at random and predicts the masked input. The fine-tuning phase requires the model to be configured with the pre-training parameters before the fine-tuning is carried out using the labeled corpus of a downstream task such as sentiment analysis.

However, while the performance of BERT in the English language achieved superior results, the performance of multilingual BERT (mBERT) in other languages did not get satisfactory results. As a consequence of this, numerous models based on BERT have been constructed from scratch for a variety of languages: TWilBert for Spanish [32], FlauBERT for French [33], FinBERT for Finnish [34], and AraBERT [35], ARBERT and MARBERT [36] for Arabic. These models have been used to get cutting-edge results for sentiment analysis on a wide range of datasets and benchmarks.

The Generative Pre-Trained Transformer (GPT), a large-scale language model developed by OpenAI, uses deep learning techniques to produce natural language text. The GPT model was built on a transformer neural network architecture that was trained on huge volumes of text data to learn patterns and relationships between words and phrases. It can help with a variety of natural language processing tasks, including text completion, question answering, and language translation. ChatGPT is a user interface that was trained using GPT-3.5 by aligning the model with human preferences, i.e., reinforcement learning from human feedback [37]. ChatGPT was used for several sentiment analysis tasks by [38] and compared with fine-tuned BERT. In terms of polarity detection, the results showed that ChatGPT wasn't quite as accurate as the fine-tuned BERT.

Prompting-based methods have been employed in conjunction with pre-trained language models for various NLP tasks, including text classification and sentiment analysis. Prompting involves incorporating human-generated text, typically in the form of brief phrases, into input or output data with the aim of guiding pre-trained models to perform targeted tasks. Utilizing prompts in natural language processing offers several advantages. One of these is that prompting can reduce computational requirements since it may not necessitate updates to the pre-trained language model's parameters, in contrast to fine-tuning methods. In addition, prompts can facilitate a more effective alignment of the new task formulation with the pre-training objective, leading to enhanced utilization of knowledge acquired during pre-training. Lastly, the closer match between the task and the pre-training can enable a few-shot approach which is particularly useful for tasks with limited training data, as a well-crafted prompt can be as valuable as hundreds of labeled data points. Prompt-based learning encompasses three distinct approaches: learning from instructions, template-based learning, and learning from proxy tasks [39].

2.2.3 Hybrid Approaches

The literature has provided a variety of hybrid approaches. Some of these are a combination of lexicon and machine learning-based approaches [40]. A key reason for employing this type of hybrid strategy is to gain stability from a lexicon-based method and high accuracy from machine learning. Another hybrid strategy is an ensemble model employing DNN models [41], [42]. The overall performance of their ensemble model surpassed the individual models. Cambria et al. [43] used both symbolic and sub-symbolic AI in a combination of top-down and bottom-up learning to detect polarity from text. Their SenticNet's sentiment analysis version incorporated logical reasoning into deep learning to generate a common-sense knowledge base. Fig. 3 depicts the evolution of sentiment analysis approaches over the last ten years.

3 EXPLAINABLE AI MODELS

A considerable number of authors reviewed research publications on explainability in artificial intelligence, as well as recent breakthroughs in XAI techniques and future research directions. As a result, based on recent literature, machine learning interpretability methods are frequently categorized according to several criteria [10], [44], [45], [46], [47], [48]. These criteria are typically based on their interpretability scope and their design capabilities for achieving explainability, which can be categorized as either ante-hoc or post-hoc approaches.

Based on their scope, interpretability methods are classified as global or local. The model is said to be interpretable if you can understand it all at once. Interpreting the output of the global model requires the trained model, algorithm knowledge, and data. The goal of local interpretability, on the other hand, is to justify a single prediction.

The basic concept is to focus on a specific example to clarify how the model's prediction was arrived at [49]. When trying to explain model outcomes, many researchers focus on local explanations instead of trying to find the model's global explanation.

A further categorization of interpretability methods focuses on the fact that explanations can be classified differently depending on whether they are generated automatically as part of the process of prediction (Ante-hoc) or whether they require post-processing after the model has already produced a prediction (post-hoc). Directly interpretable, or self-explaining, methods generate the explanation simultaneously with the prediction based on information from the model's output during the prediction process. Global self-explaining models include decision trees and rule-based models, whereas local self-explaining models include attention-based models [50].

3.1 General Explainable AI Methods Used for Sentiment Analysis

The following methods are general explainability AI strategies that are used in sentiment analysis.

3.1.1 Intrinsically – Ante-Hoc Interpretable Methods

The terms intrinsically [45] and [51] ante-hoc are used interchangeably in the literature. These methods focus on making a model understandable from the start and throughout its training, while still striving to achieve optimal accuracy. So, the simplest way to get to interpretable models is to use a subset of algorithms that develop such models. Linear models and decision trees are widely accepted in the literature as being more transparent inferences than neural networks since they are intrinsically self-interpretable. As a result, models produced using these simpler procedures typically have a lower level of accuracy than those developed using more complex black-box models [46]. Hence, the disadvantage of these models is a loss in performance with natural explainability [52].

3.1.2 Attention Mechanism Method

Attention mechanisms play a crucial role in explaining the model, and their weights are used as a proxy to explain model decisions [53].

Using attention mechanisms with deep learning models to generate explanations can identify the informative terms in a given sentence. Each term is represented by the weighted sum of the representations of the other terms. Naturally, the important terms in the sentence are assigned a high weight value, which is considered as the contribution of that term.

Classically, the architecture of the attention mechanism depends on the scaled dot-product attention applied to a query Q, a key K of dimension d_k, and a value V. The output is calculated as follows [54]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

The use of attention mechanisms to explain deep learning models for English sentiment analysis has been investigated by a few researchers.

There are several deep learning models that are combined with attention mechanisms to highlight the terms that have higher importance in a given input when making a prediction. Yang et al. [55] used two levels of attention mechanisms: sentence level and word level, based on GRU. Thus, this model can consider multiple terms and sentences informatively in different ways, and these terms and sentences are very contextually dependent. The evaluation of this model in six datasets was based on two classification tasks: sentiment analysis and topic classification. The model outperformed previous methods, and the visualization of the attention weights showed the quality of the terms and sentences selected by the model.

The Gated Convolutional Neural Network (GCNN) and the mechanism of self-attention were used in a study [56] to classify the sentiment polarities of a particular review. This study's goal was to understand and visualize the neural network's internal representations. The weights were represented by a heat map, and the results clearly showed the relationship between the term and the weight. According to another study by [57], Amazon product reviews can be classified using the Bi-LSTM model and the attention mechanism. Sentiment analysis at the sentence and aspect levels was carried out using attention weights to investigate explainability. According to the findings, sentence terms received less attention than aspect terms.

For Arabic sentiment analysis at sentence level, [58] integrated the attention mechanism with the Simple Recurrent Unit (SRU). The SRU was chosen because it allows parallel light computations that improve the accuracy and speed of the training process, while enabling adaptation of the attention mechanism to emphasize key words in the sentence. The experiments performed on the Large-Scale Arabic Book Reviews (LABR) dataset outperformed previous deep learning models by being faster and more accurate.

Similarly, [59] used the attention mechanism to identify the most informative words for classifying sentiment polarity at review level using the LABR dataset, but they used the GRU model. Furthermore, transfer learning was examined using [60] and [61] pre-trained word embeddings. The evaluation results showed the benefit of using the pre-trained word embedding in relation to accuracy and visualization of the most important words, and in terms of explanation to highlight the salient features of this prediction. Another study by [62] used the Bi-LSTM model in combination with the attention mechanism to classify three Arabic benchmark datasets. The study also investigated the benefits of preprocessing Arabic tweets and pre-trained word embedding of AraVec [63]. The evaluation results outperformed the latest deep learning models used in Arabic sentiment analysis.

3.1.3 Model-Agnostic Methods

Model-agnostic methods do not make use of any specific variety of machine learning algorithm, so they separate prediction from explanation. This separation has the great advantage of allowing flexibility in representation. Text classification, for example, often uses word embeddings for prediction and exact words for explanation.

These methods are usually post-hoc and are used to explain artificial neural networks. The differences between intrinsic (see section 3.1) and post-hoc interpretability methods are shown in Fig. 4.

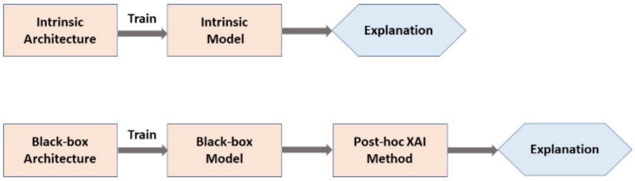
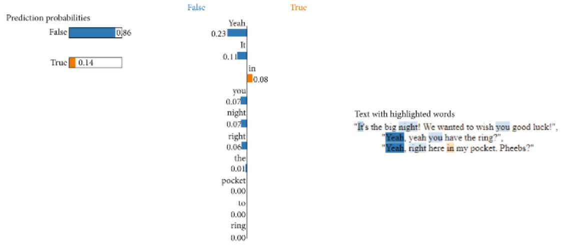


Fig. 4. Intrinsic vs. post-hoc interpretability methods.

To make AI models explainable, several model-agnostic methodologies have been created with various techniques. Local Interpretable Model-Agnostic Explanations (LIME), Feature Interaction, Individual Conditional Expectation (ICE), Feature Importance, and Shapley Values (SHAP) are examples of these techniques [52]. LIME [64] and SHAP [65] are two popular explanation methods used in various studies on explainable AI generally and on explainable sentiment analysis specifically to explain black-box model predictions. LIME was developed in a post-hoc manner, which means that explanations were supplied after the model had been created. This method can only provide a local understanding of the selected black-box model, not a global explanation. The goal of this strategy is to train intrinsically interpretable models, such as linear models or decision trees, on a new dataset obtained by randomly sampling a single instance. For example, in the text dataset, the algorithm randomly eliminates terms from the original text and employs a black-box model to compute the probability of each term to make a prediction. Following this, it attempts to predict the same result as the black box model using the selected inherently interpretable model. Further, LIME can specify the contribution of each feature to the decision. It is worth mentioning that LIME can work with tabular data, images, and text. Fig. 5 illustrates the LIME explanation.

Ensemble methods combine multiple models with distinctive features, leading to more robust and accurate predictions. However, this approach can make it difficult to provide selective explanations for the predictions, as there are multiple contributing factors. To address this challenge, interpretable machine learning models should provide concise explanations that prioritize the most salient features, even in complex scenarios. One promising method that achieves this goal is LIME, which offers a concise explanation of the model's predictions by highlighting the most influential features [45].



"It's the big night! We wanted to wish you good luck!",
"Yeah, yeah you have the ring?",
"Yeah, right here in my pocket. Pheebs?"

Fig. 5. Lime explanation [11].

The LIME explanations are calculated mathematically as follows:

$$\xi(X) = argmin_{g\epsilon G}\ L(f,g,\pi_X) + \Omega(g), \qquad (2)$$

where L is the fidelity function, π_X is the proximity measure and $\Omega(g)$ is the measure of complexity.

The black-box models can also be explained in terms of Shapley values, a cooperative game theory dating back to 1935. The concept of Shapley values is to distribute the payoff equitably among the participants in the game based on their contribution. The assumption of this theory is that the feature values of a given example (i.e., the review) are game participants, and the prediction is the payoff distributed among the game participants (i.e., the features) based on their contribution. In 2017, a unified architecture called SHapley Additive exPlanations (SHAP) was proposed by [65]. This is a post-hoc locally explainable model based on Shapley values to explain various complicated models, such as deep learning or ensemble models. This approach may also illustrate the importance of features for each prediction, which are useful for making decisions. The SHAP approach, unlike LIME, is based on a solid theoretical foundation. Fig. 6 illustrates the SHAP explanation.

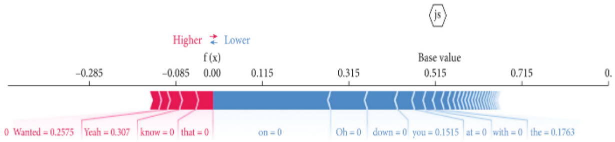


Fig. 6. SHAP explanation [11].

SHAP calculates the explanation of a specific prediction mathematically as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \quad \phi_i z'_i\ , \qquad (3)$$

where $z'\epsilon\ \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \epsilon\ R$.

In general, the following are key characteristics of the SHAP method:

1. Local accuracy, i.e., the explanation model matches the actual model.
2. Missingness, i.e., if the feature is missing, the attribution value is 0.
3. Consistency, i.e., the values change depending on how much the feature values of the model contribute.

The adaptation of existing model-agnostic explanation methods has been investigated by the research community for topics modelling in tweets, sentiment analysis, and sarcasm detection. One example is the Ex-Twit approach, which combined a model-agnostic method (LIME) and topic modelling to predict and explain health-related topics on Twitter. Evaluation results showed the effectiveness of Ex-Twit [66]. In [67], an Explainable Sentiment Analysis application for Twitter (XSA) was developed, focusing on the evaluation of this application in crisis management. The XSA helped to identify the needs of the customers and increased the marketing analysts' trust in this application during the decision-making process. LIME and SHAP are used to provide the explanations within their XSA. The results showed that the XSA can be valuable in identifying key terms that appeal to customer sentiment in textual tweets.

Another study by [68], examined three XAI explanation techniques for the application of sentiment analysis using the BERT model, namely attention-based technique, LIME, and the Integrated Gradients (IntGrad) of [69]. The Stanford Sentiment Treebank (SST) dataset of movie reviews was used in this study. According to the findings, the attention-based technique extracted explanation scores at a substantially lower computational cost than LIME and IntGrad. Sangani et al. [70] conducted a comparison study using Amazon Fine Food Reviews and LSTM and DDN models, which had similar prediction results for accuracy and F1 measure but differed in their internal architectures. The goal of the study was to get user feedback on a model that was acceptable based on LIME's explanations. The findings of the evaluation revealed that LIME was unable to capture phrase features in its explanations.

To detect sarcasm, the study by [11] used an ensemble model with LIME and SHAP to make the detection of sarcasm interpretable. The MUStARD dataset was used to evaluate the model. The researchers conducted two trials, one with utterances and the other with utterances and context. The explainable models revealed the terms that influence the model's decisions and assisted the user in determining how the model recognises sarcasm in a sentence.

3.2 Specific Explainable Methods for Sentiment Analysis

In recent years, several concepts have been explored to make the task of sentiment analysis more explainable without compromising performance. The Data Augmentation Method, which teaches the model to make predictions based on sentiment terms, is one method for increasing model explainability. Chen & Ji [71] offered two data augmentation strategies to increase the model's explainability by giving more training instances: one using an external sentiment word list and the other with adversarial examples. Three benchmark sentiment datasets were used to assess the suggested approaches for CNN and RNN classifiers. The model's explainability was evaluated using human evaluators and a simple automatic evaluation measurement.

A new neural network model, the CSNN model (Contextual Sentiment Neural Network) developed by [72], used four layers to describe a prediction technique for sentiment analysis that is natural and appeals to the human mind. It offered Initialization propagation (IP) as a new method of interpretability. The effectiveness of IP learning in enhancing the interpretability of each CSNN layer was tested using real-text datasets. Based on the findings, the CSNN had a satisfactory level of predictability and explanation ability. Two different techniques, both of which utilised transformer architecture on the IMDB dataset, were proposed by [73]. These techniques resulted in the generation of extractive summaries which offered an explanation for the decisions that the system made.

Another piece of work titled SenticNet 7, conducted by [74], involved the use of neurosymbolic techniques for sentiment analysis. In order to successfully carry out polarity detection from text, this sort of NLP task aims to identify, extract, quantify, and investigate subjective information and affective states. The explanation generated by this framework results from the fact that the classifications were explicitly tied to feelings as well as the input concepts that represent these feelings.

Linguistic characteristics were used by [75] in their framework for explainable sentiment analysis in Arabic, which incorporated dependency-based rules together with deep learning models. Sentiment terms could be mapped to concepts according to the sentence's dependence structure using the rules based on dependency analysis. These rules can be completely explained, and they develop insights into the concepts and dependencies to support each prediction. Therefore, having trust in and transparency of the model can be achieved when there is understanding of how the model reached its conclusions. If the rules fail to categorize the sentiment, the method employs deep neural networks. As a result, the framework is partially explainable. Table 1 summarizes explainable sentiment analysis methods.

TABLE 1
SUMMARY OF EXPLAINABLE SENTIMENT ANALYSIS METHODS

Explainable Model	Type of Model	Scope	Reference
Intrinsic	Ante-hoc	Global	[46]
Attention Mechanism	Ante-hoc	Local	[55], [56], [57] ,[58], [59], [62] ,[68]
Agnostic (LIME)	Post-hoc	Local	[66], [67], [68], [11], [70]
Agnostic (SHAP)	Post-hoc	Local	[67], [11]
Agnostic (IntGrad)	Post-hoc	Local	[68]
Specific (Data Augmentation)	Ante-hoc	Local	[71]
Specific (CSNN)	Ante-hoc	Local	[72]
Specific (Transformer architecture)	Ante-hoc	Local	[73]
Specific (SenticNet7)	Ante-hoc	Local	[74]
Specific (Dependency-based rules)	Ante-hoc	Local	[75]

The aforementioned methods are examples of explainable sentiment analysis methods. It is expected that this trend will continue in the future, resulting in an increasing number of novel approaches to achieving explainability in AI, particularly in the context of sentiment analysis models. One possible strategy to create an explainable model is to develop hybrid models that incorporate the qualities of several approaches. This could involve combining rule-based models and neural network models to generate more interpretable and accurate sentiment analysis methods. Another potential direction for future research in explainable sentiment analysis is the creation of explainable multimodal sentiment analysis methods that use not just text but also images and audio to provide a richer understanding of sentiment across a variety of settings.

This strategy may result in the development of more accurate and robust sentiment analysis models that account for the multimodal nature of human communication.

4. Conclusion

In this survey, the broad approaches of sentiment analysis have been explored. The innovative methods of sentiment analysis today rely heavily on deep neural networks. However, their predictions are completely uninterpretable for humans. One of the most critical issues is coming to terms with explainability and the necessity of doing so in order that a model's predictions can be accepted and justified. However, only a few of the contributions mentioned here are being used to construct models that really explain how the models justify their decisions. The results of this comprehensive survey contribute to the current literature in the field of sentiment analysis explainability by pulling together insights into techniques for sentiment analysis and the explainable methodologies in current use. The insights gained from this study may be of assistance to future research. In the future, we intend to address the issue of multilingual sentences by extending the current lexicon and to further investigate the generalization of current explainable AI approaches to range different languages and emotions. In addition, we plan to exploit Graph Neural Networks to automatically learn dependency-based rules for Arabic language.

Acknowledgment

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions. Hussain would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

References

[1] E. Cambria, A. Kumar, M. Al-Ayyoub, and N. Howard, "Guest Editorial: Explainable artificial intelligence for sentiment analysis," *Knowl Based Syst*, vol. 238, no. C, 2022, doi: 10.1016/j.knosys.2021.107920.

[2] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, 2018, doi: 10.48550/ARXIV.1801.07883.

[3] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment Analysis Is a Big Suitcase," *IEEE Intell Syst*, vol. 32, no. 6, pp. 74–80, 2017, doi: 10.1109/MIS.2017.4531228.

[4] M. Turek, "Explainable artificial intelligence (XAI)," 2018. https://www.darpa.mil/program/explainable-artificial-intelligence

[5] Y. Zhang, Y. Weng, and J. Lund, "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery," *Diagnostics*, vol. 12, no. 2, 2022, doi: 10.3390/diagnostics12020237.

[6] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv: *Machine Learning*, pp. 1–13, 2017, doi: 10.48550/ARXIV.1702.08608.

[7] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5. pp. 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[8] D. Gunning, "Explainable Artificial Intelligence (XAI)," 2017. doi: 10.1111/fct.12208.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput Surv*, vol. 51, no. 5, 2018, doi: 10.1145/3236009.

[10] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[11] A. Kumar, S. Dikshit, and V. H. C. Albuquerque, "Explainable Artificial Intelligence for Sarcasm Detection in Dialogues," *Wirel Commun Mob Comput*, vol. 2021, 2021, doi: 10.1155/2021/2939334.

[12] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[13] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining," *ACM Comput Surv*, vol. 50, no. 2, pp. 1–33, 2017, doi: 10.1145/3057270.

[14] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," *IEEE Trans. Affect Comput*, vol. 13, no. 2, pp. 845–863, 2022, doi: 10.1109/TAFFC.2020.2970399.

[15] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect Comput*, vol. 5, no. 2, pp. 101–111, 2014, doi: 10.1109/TAFFC.2014.2317187.

[16] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research," *IEEE Trans. Affect Comput*, pp. 1–30, 2020, doi: 10.1109/TAFFC.2020.3038167.

[17] B. Liu, Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies. Springer Cham*, 2012.

[18] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl Based Syst*, vol. 226, p. 107134, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107134.

[19] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artif Intell Rev*, vol. 54, no. 7, pp. 4997–5053, 2021, doi: 10.1007/s10462-021-09973-3.

[20] S. Shayaa et al., "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges," *IEEE Access*, vol. 6, pp. 37807–37827, 2018, doi: 10.1109/ACCESS.2018.2851311.

[21] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011, doi: https://arxiv.org/abs/1103.0398.

[22] Y. Bengio, I. Goodfellow, and A. Courville, Deep Learning. 2015.

[23] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[24] D. W. Otter, J. R. Medina, and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Trans. Neural Netw Learn Syst*, 32, no. 2, pp. 604–624, 2021.

[25] N. C. Dang, M. N. Moreno-García, and F. de la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics (Basel),* vol. 9, no. 3, 2020, doi: 10.3390/electronics9030483.

[26] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained Models for Natural Language Processing: A Survey," *Sci China Technol Sci*, vol. 63, no. 10, pp. 1872–1897, 2020, doi: 10.1007/s11431-020-1647-3.

[27] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Proc. of the 26th International Conference on Neural Information Processing Systems*, 2013, vol. 2, pp. 3111–3119.

[29] M. E. Peters et al., "Deep contextualized word representations," *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, vol. 1, pp. 2227–2237. doi: 10.18653/v1/N18-1202.

[30] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018, [Online]. Available: https://gluebenchmark.com/leaderboard

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[32] J. Á. González, L.-F. Hurtado, and F. Pla, "TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter," *Neurocomputing*, vol. 426, pp. 58–69, 2021, doi: 10.1016/j.neucom.2020.09.078.

[33] H. Le et al., "FlauBERT: Unsupervised Language Model Pre-training for French," ArXiv, 2020, doi: 10.48550/ARXIV.1912.05372.

[34] A. Virtanen et al., "Multilingual is not enough: BERT for finnish," ArXiv, 2019, doi: 10.48550/ARXIV.1912.07076.

[35] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," *Proc. of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.

[36] M. Abdul-Mageed, A. Elmadany, E. Moatez, B. Nagoudi, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," 2020, doi: 10.48550/ARXIV.2101.01785.

[37] OpenAI, "GPT-4 Technical Report," Mar. 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[38] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, "Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.04339

[39] B. Min et al., "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.01243

[40] R. Narayan, M. Roy, and S. Dash, "Ensemble based Hybrid Machine Learning Approach for Sentiment Classification- A Review," *Int J Comput Appl*, vol. 146, no. 6, pp. 31–36, 2016, doi: 10.5120/ijca2016910813.

[41] S. Minaee, E. Azimi, and A. Abdolrashidi, "Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models," 2019, doi: 10.48550/ARXIV.1904.04206.

[42] V. Balakrishnan, Z. Shi, C. L. Law, R. Lim, L. L. Teh, and Y. Fan, "A deep learning approach in predicting products' sentiment ratings: a comparative analysis," *Journal of Supercomputing*, vol. 78, no. 5, pp. 7206–7226, 2021, doi: 10.1007/s11227-021-04169-6.

[43] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis," *Proc. of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 105–114. doi: 10.1145/3340531.3412003.

[44] A. Rawal, J. Mccoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives," *IEEE Trans. on Artificial Intelligence*, pp. 1–1, Dec. 2021, doi: 10.1109/tai.2021.3133846.

[45] C. Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[46] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021, doi: 10.1016/j.inffus.2021.05.009.

[47] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021, doi: 10.1613/JAIR.1.12228.

[48] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities," 2021, doi: 10.48550/ARXIV.2111.06420.

[49] D. v. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics (Basel),* vol. 8, no. 8, pp. 1–34, 2019, doi: 10.3390/electronics8080832.

[50] V. Arya et al., "One Explanation Does Not Fit All : A Toolkit and Taxonomy of AI Explainability Techniques," *CoRR*, 2019, doi: 10.48550/ARXIV.1909.03012.

[51] G. Vilone and L. Longo, "Classification of Explainable Artificial Intelligence Methods through Their Output Formats," *Mach Learn Knowl Extr,* vol. 3, no. 3, pp. 615–661, 2021, doi: 10.3390/make3030032.

[52] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," 2021, doi: 10.48550/ARXIV.2101.09429.

[53] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2020, doi: 10.18653/v1/p19-1580.

[54] A. Vaswani et al., "Attention is all you need," *Adv Neural Inf Process Syst,* vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[55] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489. doi: 10.18653/v1/N16-1174.

[56] H. Yanagimto, K. Hashimoto, and M. Okada, "Attention Visualization of Gated Convolutional Neural Networks with Self Attention in Sentiment Analysis," *2018 International Conference on Machine Learning and Data Engineering (iCMLDE),* 2019, pp. 77–82. doi: 10.1109/iCMLDE.2018.00024.

[57] X. Li, X. Sun, Z. Xu, and Y. Zhou, "Explainable Sentence-Level Sentiment Analysis for Amazon Product Reviews," *2021 5th International Conference on Imaging, Signal Processing and Communications (ICISPC),* 2021, pp. 88–94. doi: 10.1109/ICISPC53419.2021.00024.

[58] S. Al-Dabet and S. Tedmori, "Sentiment Analysis for Arabic Language using Attention-Based Simple Recurrent Unit," *2nd International Conference on new Trends in Computing Sciences (ICTCS),* 2019, pp. 1–6. doi: 10.1109/ICTCS.2019.8923072.

[59] N. Almani and L. H. Tang, "Deep attention-based review level sentiment analysis for Arabic reviews," *Proc. 2020 6th Conference on Data Science and Machine Learning Applications*, CDMA 2020, pp. 47–53, 2020, doi: 10.1109/CDMA47397.2020.00014.

[60] A. A. Altowayan and L. Tao, "Word Embeddings for Arabic Sentiment Analysis," *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3820–3825. doi: 10.1109/BigData.2016.7841054.

[61] M. A. Zahran, A. Magooda, A. Y. Mahgoub, H. Raafat, M. Rashwan, and A. Atyia, "Word Representations in Vector Space and their Applications for Arabic," *Computational Linguistics and Intelligent Text Processing, 2015*, pp. 430–443. doi: 10.1007/978-3-319-18111-0_32.

[62] H. Elfaik and E. H. Nfaoui, "Deep Attentional Bidirectional LSTM for Arabic Sentiment Analysis In Twitter," *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2021, pp. 1–8. doi: 10.1109/eSmarTA52612.2021.9515751.

[63] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," *Proc. Computer Science*, 2017, vol. 117, pp. 256–265. doi: 10.1016/j.procs.2017.10.117.

[64] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[65] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. *of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777. doi: 10.48550/ARXIV.1705.07874.

[66] T. Islam, "Ex-Twit: Explainable Twitter Mining on Health Data," 2019, doi: 10.48550/ARXIV.1906.02132.

[67] D. Cirqueira et al., "Explainable Sentiment Analysis Application for Social Media Crisis Management in Retail," *Proc. of the 4th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2020),* 2020, no. Chira, pp. 319–328. doi: 10.5220/0010215303190328.

[68] F. Bodria, A. Panisson, A. Perotti, and S. Piaggesi, "Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis," *CEUR Workshop Proc*, vol. 2646, pp. 100–107, 2020.

[69] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *Proc. of the 34th International Conference on Machine Learning, 2017, vol. 70*, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

[70] R. B. Sangani, A. Shukla, and B. Radhika Selvamani, "Comparing Deep Sentiment Models using Quantified Local Explanations," *2021 Smart Technologies, Communication and Robotics (STCR),* 2021, pp. 1–6. doi: 10.1109/STCR51658.2021.9588834.

[71] H. Chen and Y. Ji, "Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019, pp. 1–11. doi: 10.48550/ARXIV.1909.04225.

[72] T. Ito, K. Tsubouchi, H. Sakaji, K. Izumi, and T. Yamashita, "CSNN: Contextual sentiment neural network," *2019 IEEE International Conference on Data Mining (ICDM),* 2019, pp. 1126–1131. doi: 10.1109/ICDM.2019.00135.

[73] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Extractive Summarization for Explainable Sentiment Analysis using Transformers," *Proc. of International Workshop on Deep Learning meets Ontologies and Natural Language Processing*, 2021, pp. 62–73.

[74] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis," *Proceedings of LREC (2022),* 2022.

[75] A. Diwali, K. Dashtipour, K. Saeedi, M. Gogate, E. Cambria, and A. Hussain, "Arabic sentiment analysis using dependency-based rules and deep neural networks," *Appl Soft Comput*, vol. 127, p. 09377, 2022, doi: 10.1016/j.asoc.2022.109377.

Arwa Diwali received the BS degree in computer science from Taibah University and the MS degree in computer science from King Abdulaziz University, Saudi Arabia. She is currently working towards the PhD degree through a joint programme between King Abdulaziz University, Jeddah, SA, and Edinburgh Napier University, Scotland, UK. Her research interests include the area of natural language processing, particularly detecting sentiments from Arabic text, and explainable AI in domains like sentiment analysis. Her research in this area has resulted in one journal paper. She is working as a lecturer at King Abdulaziz University, Jeddah, SA.

Kawther Saeedi is an Associate Professor in the Department of Information Systems at King Abdulaziz University (KAU) in Jeddah, Saudi Arabia. For the academic year 2021-2022, she is a visiting scholar at Universidad Complutense de Madrid (UCM) in Spain, working on the P2P model's project Amara. Saeedi is interested in applied research to support the adaptation of cutting-edge technologies across a wide range of domains. Her recent research has focused on the use of blockchain and other decentralized technologies to facilitate cooperation and social justice. Saeedi is from Saudi Arabia, where she earned her bachelor's degree in computer science from King Abdulaziz University in 2002. She had the opportunity to study and work in various locations around the world. Saeedi has a Ph.D. and a Master's degree in Computer Science from Manchester University in the United Kingdom. In 2007, she was awarded a JICA scholarship to spend six months in Japan learning about web applications for e-government promotion. She worked as an IT specialist for ING Bank in Amsterdam, the Netherlands, and as a programmer and Solution Engineer in Jeddah, Saudi Arabia, before entering academia.

Kia Dashtipour obtained his Honour Degree from Edinbrugh Napier University, UK, 2011. During 2015–2017 he was doing a Master in Computer Advanced System Development in University West of Scotland. He is currently a full-time Ph.D. researcher in the University of Stirling, Scotland, UK. He is working on sentiment analysis using deep learning.

Mandar Gogate obtained his B.Eng. in Electronics (with the highest 1st Class Honours with distinction) from BITS Pilani, India, in 2016. During 2015-16, he worked as a Research assistant at ENSTA ParisTech - École Nationale SupĘrieure de Techniques AvancĘes, Paris, France where he researched deep learning models for Multimodal Robotics sensor fusion and Incremental learning. He obtained his PhD in 2021 from Edinburgh Napier University, UK, where is currently a senior postdoctoral research fellow at the Centre of AI and Robotics. . He is working on multimodal big data analytics and fusion using deep neural networks in collaboration with global industry partners for solving a number of challenging real-world problems, including multi-talker speech separation, sentiment and opinion mining, cyber security and 5G-IoT applications.

Erik Cambria is the Founder of SenticNet, a Singapore-based company offering B2B sentiment analysis services, and an Associate Professor at NTU, where he also holds the appointment of Provost Chair in Computer Science and Engineering. Prior to joining NTU, he worked at Microsoft Research Asia (Beijing) and HP Labs India

(Bangalore) and earned his PhD through a joint programme between the University of Stirling and MIT Media Lab. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting. He is recipient of several awards, e.g., IEEE Outstanding Career Award, was listed among the AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future. He is an IEEE Fellow, Associate Editor of many top-tier AI journals, e.g., INFFUS and IEEE TAFFC, and is involved in various international conferences as program chair and SPC member.

Amir Hussain obtained his BEng (1st Class Honours with distinction) and PhD from the University of Strathclyde in Glasgow, UK, in 1992 and 1997, respectively. He is founding Director of the Centre of AI and Robotics at Edinburgh Napier University, UK. His research interests are cross-disciplinary and industry-led, and aimed at developing cognitive data science and trustworthy AI technologies to engineer the smart healthcare and industrial systems of tomorrow. He has (co)authored three international patents and around 600 publications, including nearly 300 journal papers and 20 Books/monographs. He has led major national and international projects and supervised over 40 PhD students. He is founding Chief Editor of Springer's Cognitive Computation journal and Springer Book Series on Socio-Affective Computing. He has been invited Editor for various other journals, including the IEEE Transactions on Neural Networks and Learning Systems, Information Fusion (Elsevier), the IEEE Transactions on Systems, Man and Cybernetics: Systems, and the IEEE Transactions on Emerging Topics in Computational Intelligence. Amongst other distinguished roles, he is an executive committee member of the UK Computing Research Committee (UKCRC) - the national expert panel of the IET and the BCS for UK computing research). He has served as General Chair of the IEEE WCCI 2020 (the world’s largest IEEE technical event in computational intelligence, comprising the IJCNN, IEEE CEC and FUZZ-IEEE) and the 2023 IEEE Smart World Congress (featuring six co-located flagship IEEE Conferences). He is Chair of the IEEE UK and Ireland Chapter Chair of the IEEE Industry Applications Society.