Received: April 13, 2024. Revised: June 13, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

GRAINY-XAI: A Domain-wise Granular Analysis of Sarcastic Text Using Explainable-AI Techniques

Jatinderkumar R. Saini[1] Shraddha Vaidya[1]* Isha Dhulekar[2] Kaushika Pal[3]

[1]*Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India*
[2]*Engineering Department Business Development Office, Heavy Handa Industrial Co. Ltd., Japan*
[3]*Sarvajanik College of Engineering & Technology, Surat, India*
* Corresponding author's Email: phdyashoda.barve@gmail.com

Abstract: Automatic sarcasm detection is essential as sarcastically written text produces noisy hidden information, which is detrimental to data-dependent systems. According to the literature, research focuses on single domain datasets. However, cross-domain or mix-domain sarcasm detection is necessary to understand the limitations and strengths of algorithms and also to accelerate the development of new solutions. In addition, understanding the underlying features of the model's outcome is significant to understand the reasons behind model's behaviour. Nevertheless, providing such explanations for models output across domains is less explored area in the literature. Thus, to fill these gaps, authors propose a model named GRAINY-XAI. The proposed model works in two phases. In phase-1, ensemble learning based models are developed across-domains on baseline datasets related to sarcasm detection namely Sem-Eval-2022 and MUStARD dataset and their performance is evaluated. In phase-2, authors check the robustness of these models using Explainable AI (XAI) techniques with granular explanations having several test cases. Further, the explanations provided by post-hoc XAI techniques, Local Interpretable Model-agnostic Explanations (LIME) for local explanations and Shapley additive explanations (SHAP) for local and global explanations were tested for quality using stability, fidelity, and coverage for LIME results and additivity and consistency for SHAP results. The results obtained through cross-domain experimentation (phase-1) provide the benchmark for further analysis as such type of experimentation is carried out for the first time in the literature. In addition, model's showed improved performance of 4.88% and 8.74%, for cross-domain and mix-domain analysis. Further, models' explanations are provided based on essential features identified using LIME and SHAP. Also, it was observed that SVC shows consistently poor performance across domains, while LGBM, CatBoost, and XGB have performed better across domains compared to other classifiers. The authors' validation test confirmed that the LIME and SHAP model's performance matches with the model's outcome.

Keywords: XAI, Sarcasm, Cross-domain, Mix-domain, LIME, SHAP.

1. Introduction

Sarcasm is a form of figurative language that's frequently utilized in written and spoken texts on microblogs like Twitter. In sarcastic sentiment, people express their bad feelings by revealing their sarcasm using a positive term in the text. There are many different forms and sequences of sarcasm, including written and spoken sarcasm [1]. On the one hand, because sarcasm has distinct implicit and explicit meanings in sentences, it can be challenging to reliably determine when it is utilized in communication when using data mining techniques. Conversely, when sarcasm is represented in utterances, it is harder for the average individual to recognize it because there isn't any tune or gesture. To identify sarcasm in a sentence, an effective Natural Language Processing (NLP) method for text classification is needed, given the presence of sardonic qualities and properties [2].

In addition, because sarcasm is a complex mode of communication that can deceive and mislead analytical systems, it's equally critical to attain superior prediction accuracy and, decision

understanding and action traceability [3]. Explainability can consider context and aid in understanding the components incorporated in the decision-making process, which allows one to modify predictions based on additional factors. This is because models are unable to account for every factor that could influence a decision. Popular techniques such as Explainable Artificial Intelligence (XAI), which aims to explain the behavior of the model and defend the actions, should be incorporated into sarcasm detection [4].

The XAI has emerged as a prerequisite for developing and deploying AI systems, along with security and justice, among other considerations. The concept of explainability has been explored by social scientists, ethical experts, and technical researchers who have already offered multiple solutions in various fields that take uncertainty into account [4]. XAI is an ethical method that offers an extensive understanding of the outcomes derived from black-box models, such as deep learning and ensemble models, contributing to the development of confidence and trust in the model's output. Thus, reducing the risks associated with production AI regarding compliance, law, security, and reputation. Nonetheless, it aids the engineers, data scientists, organizations, and end users in the comprehension of the model's operation and the results. Real-time XAI models can be implemented and utilized across multiple domains. Authors also noted that few studies in the literature provide explanations for models that detect sarcasm across different domains [3].

Thus, examining how technology and domains affect various sectors is crucial as they become more intertwined. Cross-domain or mix-domain analysis is one of the most significant elements that may be considered while examining the generalizability and adaptability of these methods[5]. It permits existing frameworks to be optimized and aids in determining the unique requirements and difficulties of the given domain. Certain changes or alterations are required depending on the goals and features of each domain. Customized solutions can be developed by identifying trends through a comprehensive analysis of applications across many domains [6]. Moreover, cross-domain analysis enables academics and practitioners to assess the efficacy and applicability of diverse models in diverse domains. By merging news headlines with Twitter, the authors assert that they have made the first attempts to identify sarcasm in multi-domain datasets. However, according to the latest literature survey, there is a paucity of studies on cross-domain sarcasm detection and XAI techniques in the field of sarcasm detection [7].

To overcome the above-mentioned challenges, the authors in this research consider the following research questions:

RQ1: How can XAI techniques offer the clarifications of the output of the ensemble learning models at a granular level (local/global explanations) while considering the domain-wise analysis of sarcasm detection?

RQ2: How are the results obtained from XAI models useful in making decisions about the model's accuracy?

To address the concerns above, the authors of this study present a novel to offer the foundation for highly localized, clear, understandable, and realistic explanations with post-hoc and model-agnostic techniques. Authors have used Local Interpretable Model-agnostic Explanations (LIME) for local explanations and Shapley additive explanations (SHAP) for local and global explanations. The explanations are provided at the granular level by considering various test cases consisting of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) per machine learning and ensemble learning model across domains.

The following are the research contributions:

1. Provided the benchmark results of ensemble models across domains of sarcasm detection.
2. Studied the behaviour of models across domains and the reasons behind their outcome using XAI techniques by identifying significant features.
3. Identification of best and worst performing models based on explanations provided with XAI techniques across domains which can assist the researchers in future to make decisions in choosing right models for sarcasm detection across domains.
4. Proposed the methods to validate the performance of LIME and SHAP techniques.

This is the structure of the remaining portion of the paper. We begin with the analysis of relevant work, and then we go on to the section describing the suggested framework. The set-ups that were employed in the experiments are then also described in detail. We wrap the paper with the experiments, findings, closing thoughts, and recommendations.

2. Literature review

This section elaborates on the review of literature in two different aspects: sarcasm detection techniques across domains and XAI in domain-wise sarcasm detection.

2.1 Sarcasm detection across domains

Sarcasm is a sentimental expression technique used to humorously criticize something or inflict emotional harm on someone by expressing implicit information that is typically contrary to the meaning of the message content [8]. Various machine learning and ensemble learning models [9] are built in the literature to detect sarcasm in text, mainly on single domain. Cross-domain analysis is one of the most significant variables that may be considered when evaluating the generalizability and adaptability of machine learning and ensemble learning techniques [10]. It facilitates the exchange of information across domains, which can shorten the time and effort required to produce innovative solutions and accelerate their development. Further, it assists the researchers to spot trends and best practices that can be used to create specialized solutions. Nevertheless, the researchers are able to ascertain which algorithms are good at performing consistently and which ones are good at becoming generalized [11]. For example, authors in [12] developed a hybrid model by assimilating content and context features and further expounded these features across domains for adversarial learning. Their work encountered the necessities of each domain in terms of features, which further helped in understanding the suitability and performance of machine learning models. In another research [13], authors have detected sarcasm by merging two datasets of Twitter and news headlines. Domain-wise analysis, allows researchers to transfer knowledge across domains and accelerate the development of new solutions while requiring less time and effort [5]. However, domain-wise, sarcasm detection is still an under-researched area and a challenging task. Moreover, the researchers in the literature have not explored the exhaustive research in the area of domain-wise sarcasm detection with an explainability approach.

2.2 XAI in domain-wise sarcasm detection

In the context of literature, the terms explainability and interpretability are synonymous. Explainability measures how much the inner workings of any machine learning or deep learning model can be clarified in terms that humans can understand, while interpretability describes what the algorithm is doing, allowing us to anticipate what will occur if certain parameters or inputs are changed [3]. A particularly significant division of interpretability techniques may occur depending on the kinds of algorithms that may be used. These techniques are referred to as model-specific if their use is limited to a certain family of algorithms; otherwise, they are called model-agnostic. Based on the scale of interpretation, the interpretable model can offer explanations at the local or global level. It is said to be local if an explanation is given for each occurrence and global if the explanation covers the entire model [4, 14]. The two well-known interpretability techniques are LIME and SHAP.

LIME is a powerful technique that can provide meaningful explanations for every single prediction of the black-box model. LIME provides a cutting-edge explanation technique that offers an interpretable model locally surrounding a prediction to faithfully and interpretably explain any classifier's predictions. LIME works with text, pictures, and tabular data and is incredibly simple to use and computationally quick [15]. Instead of using the training set of data, LIME generates a prediction by testing it against several data sets [14]. On the other hand, SHAP explains the prediction of an instance by computing the contribution of each feature to the forecast, illuminating the black-box model's output. SHAP calculates SHAP values at the local level for feature importance and then sums the absolute SHAP values for every single prediction to create a global feature importance, thereby offering both local and global interpretability [14], [16]. LIME models work based on Eq. (1), wherein LIME provides explanation for instance x, for model g, and computes loss L, by mapping with original model f, the model complexity (Ω) is kept low, and generates all possible solutions G. Eq. (2) computes the SHAP values for subset of features S.

$$explanation\,(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

$$\phi_j(val) = \sum_{S \subseteq \{1,..,p\}\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!} (val(S \cup \{j\}) - val(S)) \tag{2}$$

In the recent literature, LIME and SHAP techniques have been employed. In [17], authors have developed an approach to perform style transfer in sarcastic text by using SHAP to delete irrelevant features and generate new sentences. It was observed that with SHAP, the proposed architecture could generate correct sarcastic and non-sarcastic text. In another research, authors have performed counterfactual analysis using a multiset permutation library from python to analyse the model's performance for sarcasm detection in the political

Table 1. Comparative analysis of existing literature

Ref.	Domain-wise	Ensemble Models	Quality metric/ Assessment	Lime/Shap	Best/Worst Model
[12]	√	×	×	×	×
[13]	√	×	×	×	×
[14]	×	×	×	√	×
Proposed approach	√	√	√	√	√

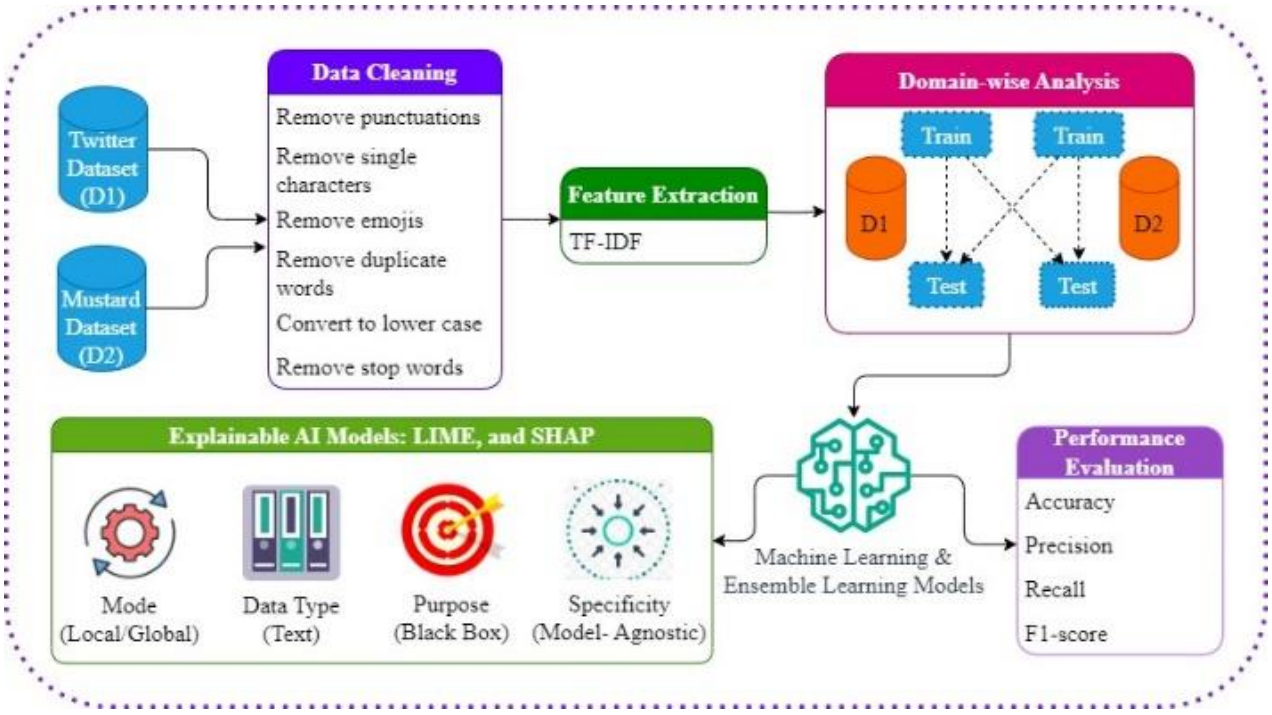

Figure. 1 Overall architecture of the proposed work

domain [18]. Authors in [19] have developed sarcasm detection and classification models for Bangla text and further analysed the important features contributing to the outcome using LIME. However, it was observed that none of the studies related to sarcasm have provided explanations for each model, compared the performance, and suggested the best suitable model, which can be achieved by performing the granular analysis such as identifying the performance of the model based on TP, TN, FP, and FN. Table 1 displays the comparative analysis of relevant existing studies and their limitations.

Following are the research gaps identified.

1. No studies from the literature have developed ensemble-based models across domains, especially with tweets and utterances combinations to detect sarcasm [7].
2. Although, studies have provided explanations for results obtained using XAI techniques, no studies have provided granular explanations for domain-wise results of the models [7].
3. No research focuses on identifying best or worst performing models based on explanations provided with XAI techniques, which could assist in finding the robustness of the models.

Lastly, although, researchers have provided explanations using XAI techniques, the outcome of these techniques is not validated through experiments.

3. Methodology

This section elaborates on the methodology adopted to perform sarcasm detection and provide explanations. Here, the authors intend to give insights into datasets used, data pre-processing and exploration techniques adopted, a list of features extracted from the data, model development on in-domain, cross-domain, and mix-domain datasets, followed by XAI techniques used, and motivation behind choosing these techniques. Fig.1 depicts the overall architecture of the proposed work.

Table 2. The dimensionality of the datasets

Dataset Name	Size	Sarcastic (S)	Non-sarcastic (NS)	Train	Test
SemEval-2022 (D1) [1]	6934	1734	5200	4853	2081
MUStARD (D2)	690	345	345	483	207
Mix-dataset (MD)	5336	1513	3823	3735	1601

[1]https://www.kaggle.com/datasets/adityaraghuvanshi999/semeval22-sarcasm-detection

3.1 Baseline datasets

In this research, authors have used two gold standard datasets. First, the SemEval 2022 sarcasm dataset (D1) from Kaggle[1], having size 6934 records with 1734 sarcastic tweets and 5200 non-sarcastic tweets with two features, namely tweet and sarcastic. The second dataset is the MUStARD (MUltimodal SARcasm Dataset), which has 690 records with an equal number (345) of sarcastic and non-sarcastic utterances. The MUStARD dataset [20] (D2) consists of utterance, speaker, context, show, and sarcasm features. The dataset includes dialogues from popular TV shows. The utterance is speech with pauses, while context provides additional information for the utterances. In this research, authors have considered only two fields, namely utterances and sarcasm, for sarcasm detection tasks to avoid further complexity when dealing with cross-domain and mix-domain analysis. The dataset sizes for cross-domain and mix-domain is discussed in experimental set-up section of this paper. The authors named these datasets as D1, D2, and MD for dataset 1, dataset 2, and mix dataset and will be using these abbreviations throughout the paper for simplicity. Table 2 depicts the details of the datasets used in this research.

3.2 Data cleaning and feature extraction

The authors initially performed exploratory data analysis to understand patterns, identify errors and find the anomalies in the data. The authors identified average length of words, characters, and number of words in sarcastic and non-sarcastic category. Further, the text from D1 and D2 is pre-processed to remove emojis consisting of emoticons, symbols and pictographs, transport and map symbols, and flags. Upon converting the text into lowercase, the text was cleaned by removing punctuation. In addition, the tweet-text is tokenized to generate set of words from each tweet. Additionally, stop-words were removed using stop words list from nltk library of python, however few words which can assist in sarcasm detection such as ‘oh’, ‘us’, ‘go’, etc. have been skipped during this process. Authors have used TF-IDF to generate terms for both tweets in D1 and utterances in D2 which act as the feature set.

3.3 Domain-wise analysis

The present research has performed exhaustive granular experimentation on two different domains consisting of general tweets from Twitter and dialogue speeches from TV shows. Initially, authors have explicitly conducted experiments for in-domain analysis, to study the performance of the models on individual datasets. In addition, the cross-domain analysis is performed to understand the model behavior on different domains. Lastly, to perform the multi-domain analysis the authors have mixed the two datasets and analysed the performance of the models on this dataset. The details of these experiments are explained in the experimental setup section.

3.4 XAI with LIME and SHAP

In this research, authors have followed a model-agnostic approach to understand the output of black-box models. Experiments are performed with textual approaches for LIME and SHAP. Both local and global explanations are provided using LIME and SHAP.

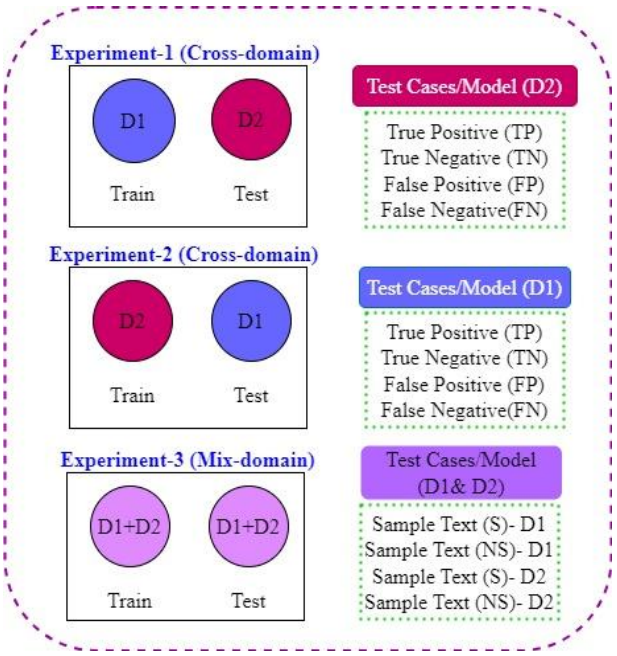

Figure. 2 Experimental set-ups

3.5 XAI with LIME and SHAP

In this research, authors have proposed cross-domain and mix-domain strategies that need datasets from different domains. The Fig. 2 depicts the experimental set-up with three experiments. For Experiment-1 (E-1), which is conducted on cross-domain dataset having train samples from D1 (4853) and test samples from D2 (207). Experiment-2 (E-2), consists of train samples from D2 (483) and 20 percent of test samples from D1 (416). Further, in experiment-3 (E-3) to understand the performance of models on mixed domain samples, the authors have combined samples from D1 and D2. Thus, making the size of MD as 5336 which is a mixture of 4853 samples from D1 and 483 samples from D2 for training set and 2288 (2081 from D1 and 207 from D2) for testing. Five different models are experimented for each experiment namely Support Vector Classifier (SV), eXtreme Gradient Boosting (XGBoost), CatBoost, Light Gradient Boosting Machine (LGBM), and Random Forest Classifier (RFC). The performance of each model is evaluated based on accuracy, precision, recall and F1-score. In addition, to provide explanations for results obtained by each model a granular analysis in terms of test cases is performed by picking one sample of type True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) and computing local explanations for each with LIME and SHAP. For E-3, authors have fetched random samples each of type sarcastic and non-sarcastic and recorded the probability (prob) values obtained with LIME and SHAP for each model. In addition, for each experiment, global explanations are provided by SHAP.

4. Experimental results and Ddiscussion

This section explores the results obtained with grainy experimentation. Section A describes the performance of the models with experiments 1 to 5 in terms of accuracy, precision, recall and F1-score. Section B, describes the results obtained using XAI techniques of LIME and SHAP on in-domain, cross-domain, and mix-domain datasets.

4.1 Performance evaluation

The performance of machine learning and ensemble models on features extracted is elaborated in this part of the paper. In E-1, Fig. 3, LGBM model showed better performance compared to other models with an accuracy, precision, recall, F1-score of 74.88, 74.04, 75.49, and 74.76 percent respectively.

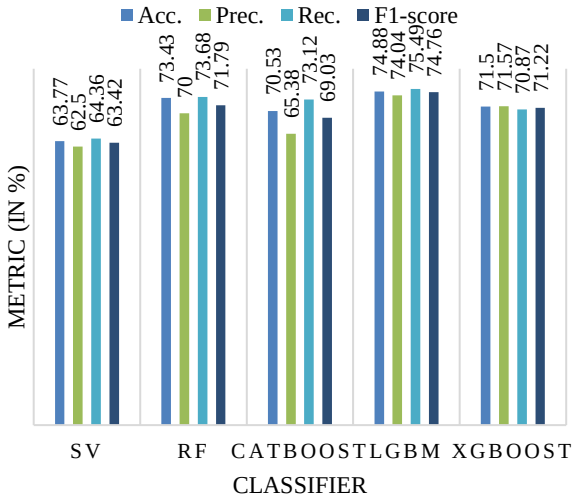

Figure. 3 Cross-domain experimental analysis (E-1)

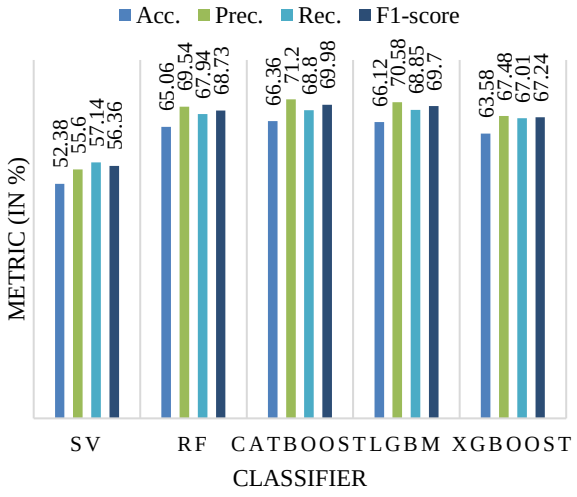

Figure. 4 Cross-domain experimental analysis (E-2)

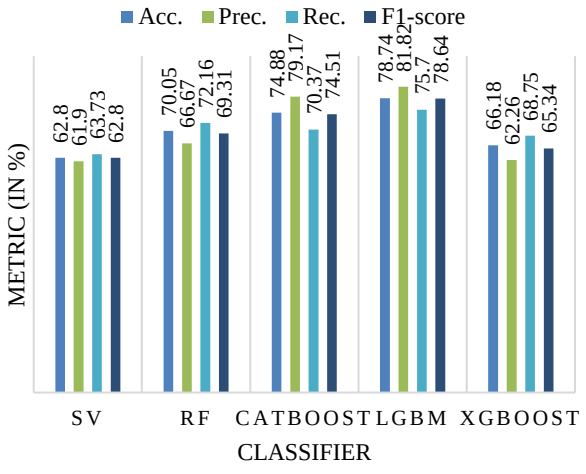

Figure. 5 Mix-domain experimental analysis (E-3)

SVC has lower performance with 63.77, 62.5, 64.36, and 63.42 percent of accuracy, precision, recall, and F1-score.

In E-2, Fig. 4, the overall performance of the models has dropped compared to previous experiments, this may be because the models were trained on D2 having smaller sizes of utterances with dialogues without context. When tested on D1, CatBoost algorithm outperformed other classifiers with accuracy, precision, recall, and F1-score of 66.36, 71.2, 68.8, and 69.98 percent respectively. The SVC classifier showed poor performance with 52.38, 55.6, 57.14, and 56.36 percent of accuracy, precision, recall, and F1-score respectively.

In E-3, Fig. 5, with mix-domain analysis, LGBM model showed best performance with 78.74, 81.82, 75.7, and 78.64 percent of accuracy, precision, recall, and F1-score. The SVC classifier showed poor performance with 62.8, 61.9, 63.73, and 62.8 percent of accuracy, precision, recall, and F1-score. It can be noted that SVC has shown consistently poor performance in all the experiments.

Therefore, in cross-domain analysis with E-1, the models showed average performance with accuracy oscillating from 63% to 74% approximately. In this case, the models had completely different test cases in the test dataset, resulting in a drop-in accuracy compared to in-domain models. However, the E-1 experiments have succeeded more than E-2 experiments, wherein, the accuracy was oscillating between 52% and 66% approximately. This may be because the models were trained only on small text of utterances without having complete context and thus were unable to make correct predictions of tweets. When models were trained on mix-domain datasets, the performance increased compared to cross-domain as the models have seen both (tweets and utterance) types of instances while training. The performance of the models increased by 4% when tested with a mix-domain dataset.

4.2 Explanations of results obtained through LIME and SHAP

In this research, authors have shown results of LIME and SHAP values for test cases for best and worst

performing models. It's difficult to provide results pictographically of all the models per experiment per test case due to length constraints of the paper and clumsy representation. Thus, authors have provided detailed explanations with weightage of each word/feature for random samples from the test cases. In the case of SHAP, the model showed F(x), the predicted value in log odds, which is converted into probability using Eq.(3), which is the exponential value of log odds.

$$Probability = \frac{exp(log\ odds)}{1+exp\ (log\ odds)} \quad (3)$$

Table 3. LIME & SHAP explanation, LGBM model (E-3)

In-domain (D2) predictions	Predicted (cross-domain)	LIME (Prob.)		SHAP (Prob.)		
		S	NS	F(x)	S	NS
TP	TP	***0.85***	0.15	0.81	0.69	0.31
TN	TN	0.42	0.58	0.76	0.32	***0.68***
FP	TP	0.60	0.40	0.64	***0.66***	0.34
FN	TN	0.31	***0.69***	0.69	0.33	0.67

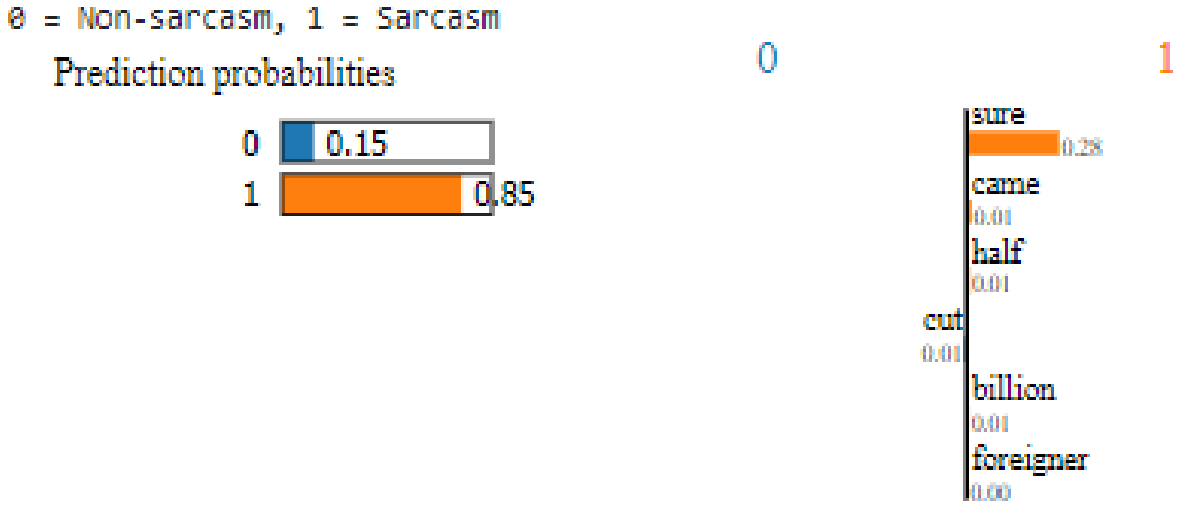

Figure. 6 LIME explanation for TP (E-3)

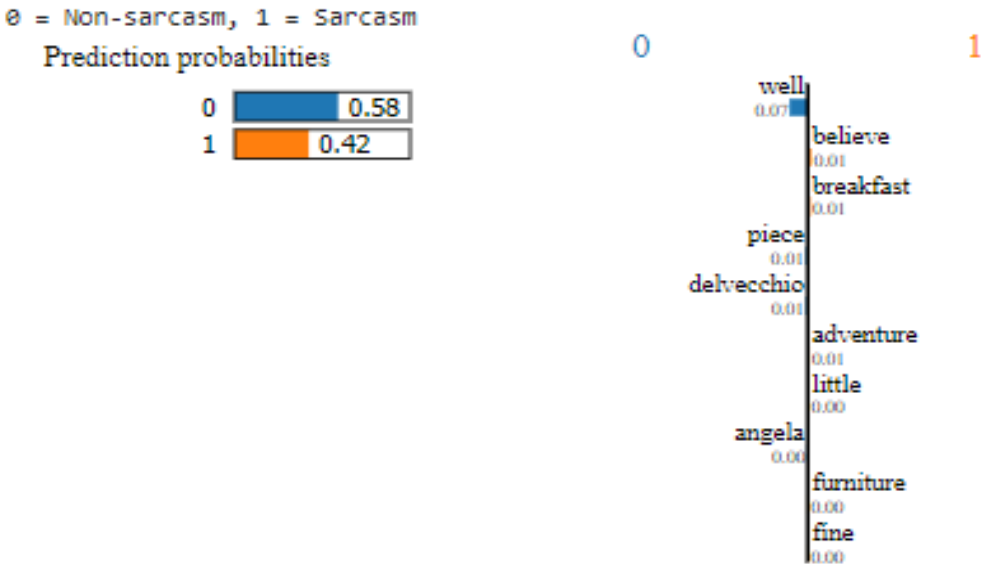

Figure. 7 LIME explanation for TN (E-3)

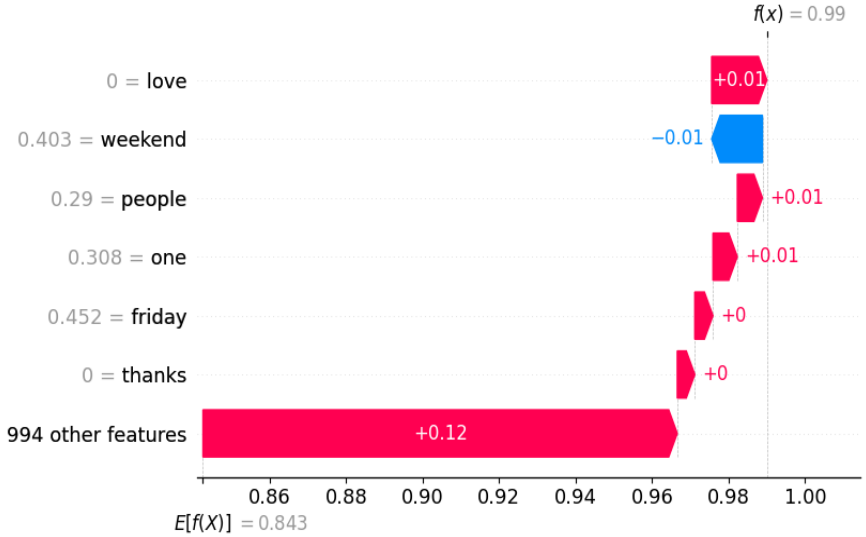

Figure. 8 SHAP explanations for Sample 1 (NS)

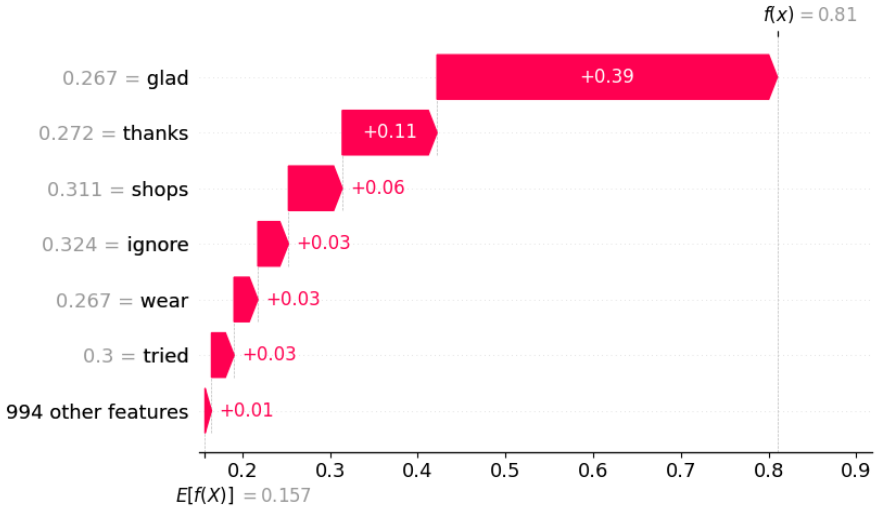

Figure. 9 SHAP explanations for Sample 2 (S)

Table 4. LIME & SHAP results, CatBoost model (E-2)

In-domain (D2) predictions	Predicted (cross-domain)	LIME (Prob.)		SHAP (Prob.)		
		S	NS	F(x)	S	NS
TP	TP	***0.74***	0.26	0.88	0.71	0.29
TN	TN	0.46	0.54	0.95	0.38	***0.72***
FP	TP	0.63	0.37	0.66	***0.66***	0.34
FN	TN	0.22	***0.78***	0.99	0.37	0.73

Table 3 shows the cross-domain explanations by LIME and SHAP for LGBM model. In this scenario, a comparison of prediction between in-domain and cross-domain is performed. The peach-colored row depicts that the model prediction for instance changed when executed with cross-domain. It can be seen that LGBM, does wonderful job with cross-domain datasets by predicting all the instances correctly. The LIME and SHAP explanations also support the predictions of LGBM by providing meaningful explanations with features. In this case, the LIME and SHAP have performed equally with LIME showing best classification for TP and FN instance of D2 while SHAP for TN and FP instance of D2. For TP with LIME, the significant features are 'sure, came, half, cut, and billion' which have contributed in detecting the instance as sarcastic. For true negative instance, the words such as 'well, believe, fine, little, and piece' have contributed in non-sarcastic category identification. Figs. 6 and 7, depict the LIME explanations for TP and TN of E-3.

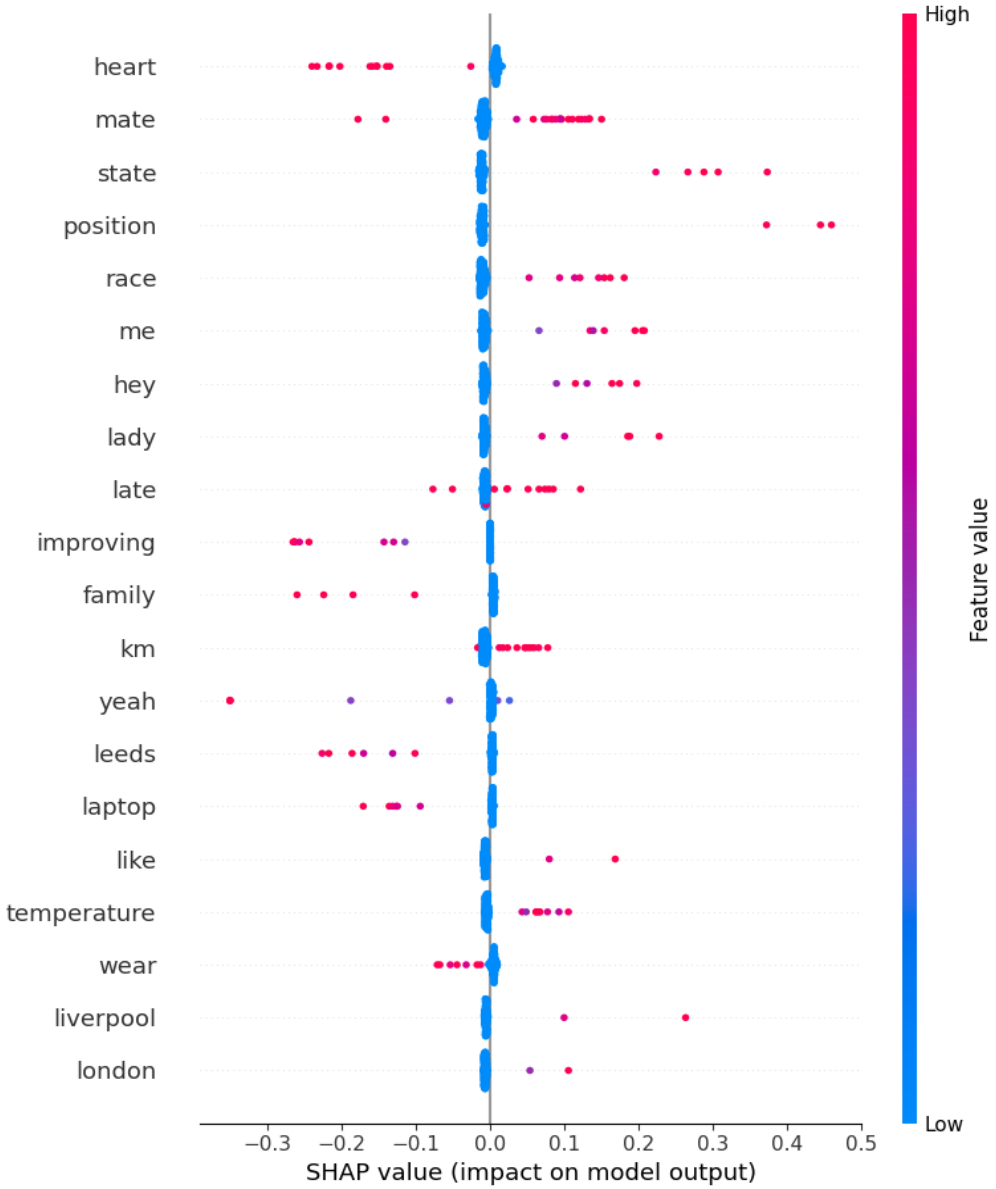

Figure. 10 SHAP summary plot (E-2)

Table 5. LIME & SHAP explanations for mix-domain Of XGB (E-3)

Random Sample	Actual Prediction	LIME		SHAP	
		S	NS	S	NS
Sample 1 (D1)	0 (NS)	0.27	***0.73***	0.3	***0.7***
Sample 2 (D1)	1 (S)	***0.98***	0.02	***0.69***	0.31
Sample 3 (D2)	0 (NS)	0.22	***0.78***	0.32	***0.68***
Sample 4 (D2)	1 (S)	***0.67***	0.33	***0.67***	0.33

Table 6. XAI based performance of models

Domain	Best performing model	Worst performing model
E1	LGBM	SVC
E2	CatBoost	SVC
E3	XGB	SVC

In, E-2, the CatBoost model showed best performance compared to other models. The LIME and SHAP have performed equally across all the test cases. The SHAP base value is 0.13 for sarcastic category and 0.87 for non-sarcastic category. Table 4 display the LIME and SHAP-based probabilities and predicted value of SHAP for the CatBoost. In E-3, mix-domain analysis is conducted, and the sample test cases are studied to understand the explanations given by LIME and SHAP. Table 5 displays the LIME explanations with textual approaches consisting of the words or features with respective weights for best performing XGB model and SVC with worst performance. Figs. 8 and 9, display the SHAP waterfall plot for Sample 1 and Sample 2 text. Further, SHAP global explanations for each experiment is studied to understand the overall important and contributing features. SHAP summary plots are shown in Fig.10. The summary plot integrates feature effects and feature importance. Every point on the summary plot represents a Shapley value for a feature and an instance. The Shapley value determines the position on the y-axis, while the feature determines the position on the x-axis.

The colour indicates the feature value from low to high. The Shapley values for overlapping points are jittered in the direction of the y-axis, providing us with an understanding of the distribution of the Shapley values for each feature. The features are ranked in order of importance. Further, Table 6 shows the best and worst performing models based on predictions of LIME and SHAP. In E1 and E2, RFC and SVC showed the best and worst performance; these predictions match the performance metric of the models. For E1, E2, and E3, LIME and SHAP predicted LGBM, CatBoost, and XGB as the best-performing models, while SVC showed poor performance for all the experiments. The XGB model was effective in selecting significant features when experimented with mix domain dataset. For cross-domain analysis, LGBM and CatBoost can be utilized. The colour indicates the feature value from low to high. The Shapley values for overlapping points are jittered in the direction of the y-axis, providing us with an understanding of the distribution of the Shapley values for each feature. The features are ranked in order of importance.

Further, Table 6 shows the best and worst performing models based on predictions of LIME and SHAP. In E1 and E2, RFC and SVC showed the best and worst performance; these predictions match the performance metric of the models. For E1, E2, and E3, LIME and SHAP predicted LGBM, CatBoost, and XGB as the best-performing models, while SVC showed poor performance for all the experiments. The XGB model was effective in selecting significant features when experimented with mix domain dataset. For cross-domain analysis, LGBM and CatBoost can be utilized.

4.3 Assessing insights of XAI models

Essentially, XAI models should be evaluated for evaluating the stability, consistency, reliability, and efficiency of the explanations provided by the XAI models [4]. This section discusses the parameters used to evaluate the performance of XAI models in providing interpretable and trustworthy explanations. Authors in this research have performed rigorous analysis of both the XAI techniques implemented in this research across all the experiments (E-1 to E-3). Section A and B discusses the LIME and SHAP assessment techniques and results obtained respectively.

A. Assessment of LIME Insights

LIME insights were measured with three standard parameters from the literature fidelity, stability, and coverage. These parameters are described as follows.

1. Stability

The stability parameter tests whether the LIME explanations are steady over number of executes or small perturbations. In this research, authors have performed perturbations by making small changes to the feature values of LIME over a perturbation scale of 0.01 and randomly adding noise to the feature values. The perturbations were tested on 30% size of the datasets in each experiment and similarity is

Table 7. LIME stability check with perturbation

Algorithm	E1	E2	E3
SVC	0.88	0.87	0.89
RFC	0.87	0.86	0.86
LGBM	0.91	0.93	0.93
XGBoost	0.92	0.91	0.97
CatBoost	0.86	0.88	0.94

Table 8. LIME coverage of subset of instances

Algorithm	E1	E2	E3
SVC	91	87	88
RFC	95	96	95
LGBM	92	95	97
XGBoost	92	91	97
CatBoost	94	90	94

Table 9. SHAP additivity difference

Algorithm	E1	E2	E3
SVC	0.05	0.03	0.02
RFC	0.01	0.01	0.01
LGBM	0.03	0.03	0.02
XGBoost	0.03	0.01	0.03
CatBoost	0.02	0.04	0.01

Table 10. SHAP stability check with perturbations

Algorithm	E1	E2	E3
SVC	0.76	0.79	0.88
RFC	0.93	0.96	0.99
LGBM	0.94	0.93	0.93
XGBoost	0.92	0.91	0.98
CatBoost	0.88	0.91	0.93

calculated between original LIME prediction and perturbed instance-based LIME prediction using three similarity measures namely Jaccard, Euclidean, and Cosine. It was observed that LIME has shown more stable performance with Cosine measures, while average to poor performance by Jaccard and Euclidean based models. Table 7 displays the average of similarity between original and perturbed explanations of instances. Value close to 1 indicate that the explanations are very much similar while 0 indicates difference in the similarity.

2. Coverage

The LIME coverage parameter indicates the contribution of features with higher weights contributing in predicting the LIME outcomes. In this scenario, LIME explanations are generated per experiment for each model and summation of the maximum weights above a threshold value is chosen which indicates covered instances. The covered instances are the cases for which LIME has provided meaningful and correct explanations similar to the underlying complex models based on the parameter of threshold value. Therefore, to compute the coverage in percentage, authors divided the covered instances by total number of instances multiplied by hundred. Table 8 shows that experiments 1 to 5 have shown coverage of 90 percent to 99 percent, which means that the LIME explanations have provided meaningful explanations for all the cases considered in this set.

B. Assessment of SHAP Insights

SHAP insights were examined with two major aspects namely additivity and consistency. This section elaborates the experimentation conducted and results obtained.

1. Additivity

The additivity property of SHAP indicates that the model output is equivalent to addition of expected model output or baseline output and summation of SHAP values. In other words, sum of the SHAP values equals the difference between model output and expected output, in a case where these values are different, it is called as additivity difference. Ideally, the additivity difference must be zero indicating the calibrated and correct contribution of SHAP features. Table 9, shows the additivity difference for experiments from E-1 to E-5, which is close to zero indicating the good contribution of features in SHAP predictions.

2. Consistency

SHAP consistency can be measured by perturbing the instances and comparing the similarity between original explanation and the perturbed explanations. Authors have followed similar approach as that of LIME stability computation to calculate the consistency of SHAP across the experiments. Table 10 shows the similarity values per model across experiments using Jaccard, Euclidean, and Cosine similarities. It can be seen that SHAP explanations were more consistent with Cosine similarity measure, compared to other similarity measures.

Table 11. Comparative analysis

Sr. No.	Ref.	Acc.	Prec.	Rec.	F1-Score
1	E-1	74.88	74.04	75.49	74.76
2	E-2	66.36	71.20	68.80	69.98
3	E-3	78.74	81.82	75.70	78.64
4	[10]	0.70	0.71	0.67	0.67

4.4 Comparative analysis with existing studies

According to the literature survey, authors have found only two studies detecting sarcasm across domains [12, 13]. Nevertheless, these two studies are using different datasets and deep learning-based techniques, which are not the part of this research. Thus, authors have tried to compare the performance of the proposed approach of E-1, E-2, and E-3 with results from [13] as it uses TF-IDF features and ensemble models. Table 11 displays the comparative analysis, wherein the proposed E-1 and E-3, have outperformed existing technique by 4.88% and 8.74% respectively, this certainly differs due to data cleaning process and not excluding sarcasm related words from the data as mentioned in methodology section. These words contribute in detecting sarcasm, which was not considered by authors in [13]. They have focused on merely capturing term frequency from the contents. Thus, proposed approach of data pre-processing has helped in boosting the performance of the models.

Another phase of this research is XAI techniques. Although, no direct comparison of such technique is available in existing literature, thus making this research a novel contribution, authors have made comparison with single domain studies. It can be seen that, the research by [14], have used LIME and SHAP to provide explanations and found that XGBoost algorithm was superior and provided interpretations for the same. Taking it further, authors in this research found that the XGBoost algorithm performed better with mix-domain dataset as well. Thus, it can be concluded that XGBoost algorithm is trustworthy and used with other domains as well.

Thus, all in all, the proposed work seems novel in terms of performing domain-wise analysis of sarcastic text and contributing to the benchmark results, providing explanations to the model's outcome based on significant features, validating the quality of explanations through experiments, and finally determining the best and worst performing models across domains using XAI techniques.

5. Conclusion, Limitations and future enhancements

The research in this paper is divided into two parts. Firstly, the performance of the ensemble learning models is evaluated with three experiments (E-1 through E-3) across domains. Secondly, the post-hoc explanations at the local and global levels were provided using LIME and SHAP techniques of XAI. The granular analysis is performed by collecting test cases for each model concerning TP, TN, FP, and FN obtained by the model. In E1, LGBM outperformed with 74.88%, while SVC showed poor performance with 63.77%. In E2 and E3, CatBoost and LGBM performed well with 66.36% and 78.74% respectively while SVC again showed poor performance. Thus, proposed approach outperformed existing studies by 4.88% and 8.74% for E-1 and E-2. Explanations of LIME and SHAP based on significant features have been validated for quality with several experimentation. Additionally, based on test cases XAI-based best performing models include LGBM, CatBoost, and XGBoost while worst-performing model was SVC.

The research certainly has few limitations which can be overcome in future. First, the authors have not considered the context of the MUStARD dataset, which can improve the performance of the model. Secondly, authors have implemented ensemble classifiers, but techniques such as voting and stacking can be adopted to improve the performance and study the difference in results and explanations by LIME and SHAP.

In the future, the authors want to study the topic modelling-based feature extraction across domains to improve the performance of the models.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Shraddha Vaidya; methodology, Shraddha Vaidya; software, Isha Dhulekar; validation, Isha Dhulekar, and Shraddha Vaidya; formal analysis, Jatinderkumar R. Saini; investigation, Jatinderkumar R. Saini; resources, Shraddha Vaidya; data curation, Shraddha Vaidya; writing—Shraddha Vaidya; writing review—Shraddha Vaidya, Jatinderkumar R. Saini; visualization, Shraddha Vaidya; supervision, Jatinderkumar R. Saini; project administration, Isha Dhulekar.

References

[1] C. I. Eke, A. A. Norman, and H. F. Nweke, "Sarcasm identification in textual data: systematic review, research challenges and open directions", *Artif Intell Rev*, Vol. 53, No. 6, pp. 4215-4258, 2020, doi: 10.1007/s10462-019-09791-8.

[2] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, "Sarcasm detection using machine learning algorithms in Twitter: A systematic review", *International Journal of*

Market Research, Vol. 62, No. 5, pp. 578-598, 2020, doi: 10.1177/1470785320921779.

[3] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods", *Entropy*, Vol. 23, No. 1, pp. 1-45, 2021, doi: 10.3390/e23010018.

[4] R. Bagate, A. Saini, K. Sethi, H. Tomar, and A. Singh, "Sarcasm Detection and Explainable AI: A Survey", In: *Proc. of the 3rd International Conference on Communication & Information Processing (ICCIP)*, 2021.

[5] R. K. Singh, M. K. Sachan, and R. B. Patel, "360 degree view of cross-domain opinion classification: a survey", *Artif Intell Rev*, Vol. 54, No. 2, pp. 1385-1506, 2021, doi: 10.1007/s10462-020-09884-9.

[6] V. Khetani, Y. Gandhi, S. Bhattacharya, S. N. Ajani, and S. Limkar, "Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains", *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 11, No. 7s, pp. 253-262, 2023.

[7] W. Chen, F. Lin, G. Li, and B. Liu, "A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities", *Neurocomputing*, Vol. 578, 2024, doi: 10.1016/j.neucom.2024.127428.

[8] H. Liu, R. Wei, G. Tu, J. Lin, C. Liu, and D. Jiang, "Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection", *Information Fusion*, Vol. 108, 2024, doi: 10.1016/j.inffus.2024.102353.

[9] V. Haripriya and P. G. Patil, "An Ensemble Framework with Optimal Features for Sarcasm Detection in Social Media Data", *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 12, No. 1s, pp. 748-760, 2024

[10] N. Singh, U. Chandra Jaiswal, and M. Mohan, "Cross Domain Sentiment Analysis Techniques and Challenges: A Survey", In: *Proc. of 4th International Conference on Communication & Information Processing (ICCIP)*, 2022.

[11] V. Khetani, Y. Gandhi, S. Bhattacharya, S. N. Ajani, and S. Limkar, "Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains", *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 11, No. 7s, pp. 253-262, 2023.

[12] P. Kadli and B. M. Vidyavathi, "Cross Domain Hybrid Feature Fusion based Sarcastic Opinion Recognition Over E-Commerce Reviews Using Adversarial Transfer Learning", *International Journal of Intelligent Engineering and Systems*, Vol. 16, No. 2, pp. 152-165, 2023, doi: 10.22266/ijies2023.0430.13.

[13] R. Jamil, I. Ashraf, F. Rustam, E. Saad, A. Mehmood, and G. S. Choi, "Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model", *PeerJ Comput Sci*, Vol. 7, pp. 1-24, 2021, doi: 10.7717/peerj-cs.645.

[14] 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.

[15] F. Curia, "Cervical cancer risk prediction with robust ensemble and explainable black boxes method", *Health Technol (Berl)*, Vol. 11, No. 4, pp. 875-885, 2021, doi: 10.1007/s12553-021-00554-6.

[16] K. Zahoor and N. Z. Bawany, "Explainable artificial intelligence approach towards classifying educational android app reviews using deep learning", *Interactive Learning Environments*, 2023, doi: 10.1080/10494820.2023.2212708.

[17] X. Yang, "Transferring Styles between Sarcastic and Unsarcastic Text using SHAP, GPT-2 and PPLM", In: *Proc. of - 2022 4th International Conf on Natural Language Processing, ICNLP 2022*, 2022, pp. 390-394. doi: 10.1109/ICNLP55136.2022.00072.

[18] R. A. Bagate and R. Suguna, "Sarcasm Detection on Text for Political Domain— An Explainable Approach", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 10, pp. 255-268, 2022, doi: 10.17762/ijritcc.v10i2s.5942.

[19] R. Anan, T. S. Apon, Z. T. Hossain, E. A. Modhu, S. Mondal, and D. G. R. Alam, "Interpretable Bangla Sarcasm Detection using BERT and Explainable AI", In*: Proc. of 2023 IEEE 13th Annual Computing and Communication Workshop and Conf, CCWC 2023*, pp. 1272-1278, 2023, doi: 10.1109/CCWC57344.2023.10099331.

[20] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper)", In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, 2019