Sarcasm Detection in News Headlines using Supervised Learning

Ashok Kumar Jayaraman
Information Science and Technology
Anna University, Chennai, India
jashokkumar83@auist.net

Tina Esther Trueman
Department of Computer Science
University of the People, United States
tina.trueman@uopeople.edu

Gayathri Ananthakrishnan
Department of Information Technology
VIT University, Vellore, India
gayathri.a@vit.ac.in

Satanik Mitra
Department of ISE
Indian Institute of Technology, Kharagpur
satanikmitra@iitkgp.ac.in

Qian Liu, Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University, Singapore
cambria@ntu.edu.sg

***Abstract*—Nowadays, social media has an enormous amount of news content with a sarcastic message. It is often expressed in the form of verbal and non-verbal. In this paper, the authors aim to identify sarcasm in news headlines using supervised learning. We address this task with the Bag-of-words features, context-independent features, and context-dependent features. Specifically, the authors employ six supervised learning models, namely, Naïve Bayes-support vector machine, logistic regression, bidirectional gated recurrent units, Bidirectional encoders representation from Transformers (BERT), DistilBERT, and RoBERTa. Our experimental results indicate that RoBERTa achieves a better performance than others.**

***Index Terms*—Sarcasm Detection, Supervised learning, News Headlines Data, Transformers, BERT.**

I. Introduction

Today, the digital world generates a huge volume of news content about an individual, an organization, a service, or a product in various formats such as images, videos, audio, and text. Therefore, it becomes one of the powerful resources for people to learn about current events or things in the world. Specifically, readers like to read only the headlines rather than the entire news content. The news headlines influence the readers' understanding, reasoning, and deceiving towards the news statement [1]. In this case, sentiment analysis plays a vital role in identifying the news headlines without any misconception. Specifically, sentiment analysis computes a semantic orientation or sentiment polarity of a given news headline into either positive, negative, or neutral [2], [3]. However, it fails to detect a nuanced form of language from the spoken or given news headline [4]. For instance, the sentiment analysis determines the given news headline 'Voters shocked Christie botched such an easy political cover-up' as positive [5]. On the other hand, researchers used sarcasm or humor detection tasks to identify the nuanced form of languages in the given text [6]. The sarcasm detection task identifies whether the given text or news headline is sarcastic or not sarcastic. For instance, the sarcasm detection task determines the given news headline 'Voters shocked Christie botched such an easy political cover-up' as sarcasm.

Similarly, it determines the given news headline 'Obama visits Arlington National Cemetery to honor veterans' as not sarcastic [7], [8]. Therefore, sarcasm detection is a challenging task in natural language processing (NLP) [9]. It is widely used in various NLP applications such as marketing research and information categorization [10]. Researchers studied the sarcasm detection task using different techniques, namely, rule-based techniques, machine learning-based techniques, and deep learning-based techniques [11], [12]. First, rule-based techniques identify sarcasm in a text through user-specific rules. These rules are designed to capture human knowledge of a text in a specialized domain. Second, machine learning techniques identify sarcasm in a text using feature-engineering methods. Third, deep learning-based techniques use a semantic representation of a text to identify sarcasm. Both machine learning and deep learning techniques are broadly studied into supervised, unsupervised, and semi-supervised learning. The supervised learning models use the labeled data to map an input text to the desired output. In unsupervised learning, models use unlabeled data to group or cluster similar texts together. The semi-supervised learning models use a large amount of unlabeled data and their part of labeled data to predict the desired output [13]. Recently, transformers-based models achieved a better result in various NLP tasks such as text classification, sentiment analysis, dialogue systems, and recommendation systems. In this paper, the authors mainly focus on the sarcasm detection task in news headlines. Specifically, the authors employ supervised learning techniques such as NBSVM (Naïve Bayes ' Support Vector Machine) [14], LR (Logistic regression) [15], BiGRU (Bidirectional Gated Recurrent Unit) [16], [17], BERT (Bidirectional Encoder Representation from Transformers) [18], DistilBERT [19], RoBERTa [20], and XLNet [21] models for the task of sarcasm detection. Both the NBSVM and LR learn bag-of-words (BoW) features. BiGRU learns unidirectional semantic (context-independent) features from either left to right or right to left, and BERT and XLNet-based models learns bidirectional semantic (context-dependent) features.

In particular, this paper contributes to the following:

- Addresses the sarcasm detection task using the BoW, unidirectional semantic, and bidirectional semantic context features.
- Employs and compares six supervised learning models on the sarcasm detection task.
- The RoBERTa model with bidirectional semantic context features achieves a better performance.

The rest of this paper is structured as follows: Section II presents the existing studies of sarcasm detection; Section III illustrates the sarcasm detection task in news headlines using supervised learning techniques; Section IV explains the obtained results and the discussion; finally, Section V offers concluding remarks.

II. Related work

In day-to-day life, everyone uses sarcasm in different situations, where it positively conveys the negative message. Researchers studied sarcasm detection in cognitive sciences, psychology, and linguistics. In this paper, the authors briefly describe the existing research works in sarcasm detection. Amir et al. [22] developed a system to automatically detect sarcasm in social media using a deep neural network. Specifically, the authors applied user embeddings with word embeddings for recognizing sarcasm. Their study indicated that modeling of authors' information significantly improves the performance. Hazarika et al. [23] designed a contextual sarcasm detection system to identify sarcasm in online social media content. This system adopts both user embeddings (context-driven) and content embeddings for boosting classification performance. Kolchinski et al. [24] explored two data-driven methods for sarcasm detection in social media. In particular, the authors used a Bayesian approach to represent the authors' behavior to be sarcastic and a dense embedding approach to learning the author and text interactions, respectively. Then, they employed an augmented BiRNN (Bidirectional Recurrent Neural Networks) on these representations to improve the classification performance. Castro et al. [25] introduced a multimodal sarcasm detection dataset based on popular TV shows. The labels are annotated based on the conversation of audiovisual utterances. Their results indicated that the SVM reduces the error rate up to 12.9% in F1-score. Jena et al. [26] implemented BERT-based C-Net (Contextual-Network) architecture for sarcasm detection. This method uses an SSE (Simple exponential smoothing) in the fusion layer of the proposed C-Net architecture. Their results indicated that the C-Net with SSE achieves 75.0% in the Twitter dataset and 66.2% in the Reddit dataset. Nayel et al. [27] developed the SVM model to detect sentiment and sarcasm in Arabic Twitter. They achieved 85.55% accuracy for sentiment detection and 84.22% accuracy for sarcasm detection. Bouazizi and Ohtsuki [28] used random forest (RF) model to implement a pattern-based approach for detecting sarcasm. The author defined four sets of features sets such as sentiment features, punctuation features, syntactic and semantic features, and pattern-related features. They achieved a 81.3% accuracy for sarcasm detection.

Moreover, Bamman and Smith [29] adopted the binary LR method to implement the contextualized sarcasm detection. The authors divided the features into four classes, namely, tweet features, author features, audience features, and response features. Their results indicate that the proposed LR achieves 85.1% by including all features. Zhang et al. [30] developed a BiGRU with a pooling layer to detect sarcasm in Twitter. The BiGRU model captures the syntactic and semantic features and the pooling layer automatically extracts features from tweets history. They achieved a 90.74% F1 score for balanced data and a 90.26% F1 score for imbalanced data. Liu et al. [31] implemented an A2Text-Net (a deep neural network) for sarcasm detection. This network combines multiple auxiliary data such as part-of-speech (POS), punctuations, emoji, numeral, etc. The authors have shown a better performance using multiple auxiliary data. Mukherjee and Bala [32] proposed a Naïve Bayes and fuzzy clustering model to detect sarcasm. They practically tested this model with different feature sets, namely, content words, function words, POS tags, POS n-grams, content and function words, function and POS n-grams, and content, function words, and POS n-grams. Their study indicated that the NB and fuzzy clustering model achieves a better result with content and function words. Joshi et al. [33] developed an SVM with a radial basis kernel to detect sarcasm in tweets and posts. They mainly used context incongruity and sarcastic to define the linguistic features such as implicit congruity, explicit congruity, lexical, and pragmatic features. Their results indicated that the proposed SVM model improves 40% accuracy for rule-based algorithms and 5% for statistical classifiers. Overall, the existing researchers studied sarcasm detection tasks using frequency and unidirectional features in tweets and posts. In this paper, the authors explore sarcasm detection in news headlines using supervised learning methods. Specifically, the authors study various feature techniques such as frequency, unidirectional, and bidirectional.

III. Supervised learning for sarcasm Detection

The authors present the sarcasm detection task in news headlines using supervised learning methods. Fig. 1 shows the generalized step-by-step process of the proposed supervised models. Each step of the model is described as follows.

A. Dataset

The authors use the benchmarked news headlines dataset for the task of sarcasm detection [7], [8]. It has two versions, namely version1, and version2. The first version of the dataset contains 26709 news headlines that include 11724 sarcastic news headlines and 14985 non-sarcastic news headlines. Similarly, the second version of the dataset contains 28619 news headlines that include 13634 sarcastic news headlines and 14985 non-sarcastic news headlines. It is upgraded with 1910 news headlines from version1. Specifically, the sarcastic news headlines are collected from the TheOnion website, and non-sarcastic news headlines are collected from the HuffPost website.

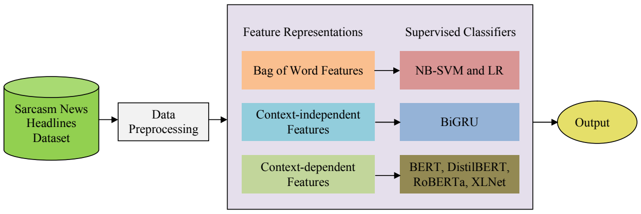
Sarcasm News Headlines Dataset
Data Preprocessing
Feature Representations
Supervised Classifiers
Bag of Word Features
NB-SVM and LR
Context-independent Features
BiGRU
Context-dependent Features
BERT, DistilBERT, RoBERTa, XLNet
Output

Fig. 1. The general framework for supervised learning models for the task of sarcasm detection

B. Text preprocessing

The authors first correct the broken Unicode in the given sarcasm news headlines datasets. Second, the authors expand the shorten tokens like "can't" into "cannot" [34]. Third, the authors use upper case to lower case letters and punctuation removal (except question marks, single and double quotes, and periods) for obtaining the quality of data.

C. Feature Representations

1) BoW Features: BoW describes a list of words or tokens in the matrix form based on their frequencies or occurrences in the given text. The BoW technique completely ignores the grammar, structure, and semantic meaning of the text [35].

2) Context-independent Features: Context-independent features map an input token into a continuous semantic context vector. It includes Word2Vec [36], GloVe [37], and fastText [38] embedding models. In particular, the authors use fastText to generate semantic context vectors in a fixed dimension. The fastText is built with the sub-word information (character n-grams) and skip-gram model with negative sampling. Specifically, these capture the meaning of sub-words and their prefixes and suffixes. Moreover, the authors use the BiGRU to learn unidirectional context information from both forward and backward directions.

3) Context-dependent Features: Context-dependent features learn an input token based on the current and previous input token. In particular, the BERT model uses a self-attention mechanism to capture the context information from both left and right directions [39]. For instance, consider the sentences 'He is reading a book' and 'He is reading a research article.' In both sentences, the context-independent feature generators such as Word2Vec, GloVe, and fastText learn the same context information for the token 'reading.' On the other hand, the BERT model learns different context information for the given token 'reading' based on other input tokens in the sentence.

D. Supervised Learning Methods

1) Logistic Regression: Logistic Regression is a simple discriminative model. It is also known as the maximum-entropy classifier, logit regression, or log-linear model [40], [41]. This model computes the probabilities that are a possible outcome of an event using a sigmoid function. Moreover, the LR handles a binary event with linear and non-linear data, and it selects a high probability value in the case of multiple events or classes.

TABLE I
DATA DISTRIBUTION FOR TRAINING, VALIDATION, AND TESTING

Dataset	Class	Train	Valid	Test
Version1	Not Sarcastic	12137	1349	1499
	Sarcastic	9497	1055	1172
	Total	21634	2404	2671
Version2	Not Sarcastic	12137	1349	1499
	Sarcastic	11044	1227	1363
	Total	23181	2576	2862

2) Naïve Bayes-Support Vector Machine: NBSVM is a text classification approach proposed by Wang and Manning [14]. This approach infuses a linear SVM model with Bayesian probabilities. In particular, it uses the Naïve Bayes log-count ratios instead of the word count features. Therefore, the NBSVM has become one of the fast and powerful approaches in text classification tasks. In addition, it performs well for long documents.

3) Bidirectional Gated Recurrent Units: RNN is a type of ANN (Artificial Neural Network) that is designed with internal memory to deal with sequential data. The RNN is widely studied in language translation, speech recognition, and NLP tasks. However, it fails to capture long-range term dependencies. Therefore, LSTM (long short-term memory network) and GRU are introduced to capture the long-range term dependencies. Both LSTM and GRU work like a standard RNN, but they work differently in each recurrent unit. The LSTM is designed with three gating mechanisms as input, output, and forget gates. Similarly, the GRU is designed with two gating mechanisms such as reset and update gates [16], [17], [42]. These gates help to observe new information and preserve the previous information. The GRU works faster than LSTM networks. Therefore, the authors use the BiGRU for the tasks of sarcasm detection. The BiGRU concatenates the forward context information and backward context information of the given input sequence.

4) BERT-Based Classifier: BERT is a neural network-based language representation model [39] that is designed based on the encoder structure of the transformer model [43]. It uses the self-attention mechanism and FNN (feed-forward neural network) concepts to learn the bidirectional context information of an input token based on the previous and next input tokens. In particular, BERT is implemented in two steps, namely, pre-training and fine-tuning. First, the pre-training step is designed on unlabeled data for masked language prediction and next sentence prediction (NSP) tasks. Second, the fine-tuning model initializes pre-trained parameters for downstream tasks. The BERT model has two sizes, namely, BERT-Base and BERT-Large. The BERT-Base contains 12 encoder layers, 768 hidden state representations, and 12 self-attention layers with 110M trainable parameters. Similarly, the BERT-Large consists of 24 encoder layers, 1024 hidden state representations, and 16 self-attention layers with 340M trainable parameters. In this paper, the authors use the BERT-base fine-tuning task for sarcasm detection.

TABLE II
CONFUSION MATRIX

Model	Class	Version1						Version2					
		Training		Validation		Testing		Training		Validation		Testing	
		NS	S	NS	S	NS	S	NS	S	NS	S	NS	S
LogReg	NS	11224	913	1202	147	1330	169	10891	1246	1140	209	1272	227
	S	1404	8093	228	827	235	937	1113	9931	201	1026	184	1179
NBSVM	NS	11203	934	1176	173	1321	178	10806	1331	1143	206	1271	228
	S	1256	8241	209	846	223	949	1190	9854	193	1034	176	1187
BiGRU	NS	12038	99	1189	160	1323	176	12014	123	1181	168	1304	195
	S	118	9379	162	893	182	990	135	10909	173	1054	197	1166
BERT-Base	NS	12102	35	1256	93	1411	88	12085	52	1266	83	1411	88
	S	50	9447	85	970	108	1064	75	10969	97	1130	102	1261
DistilBERT-Base	NS	12034	103	1271	78	1401	98	12029	108	1252	97	1405	94
	S	125	9372	97	958	100	1072	170	10874	115	1112	132	1231
RoBERTa-Base	NS	12091	46	1286	63	1422	77	12098	39	1295	54	1443	56
	S	184	9313	99	956	107	1065	277	10767	117	1110	121	1242
XLNet-Base	NS	11582	555	1243	106	1388	111	11300	737	1227	122	1371	128
	S	915	8582	100	955	118	1054	787	10357	92	1135	91	1272

* NS-Not Sarcastic, S-Sarcastic

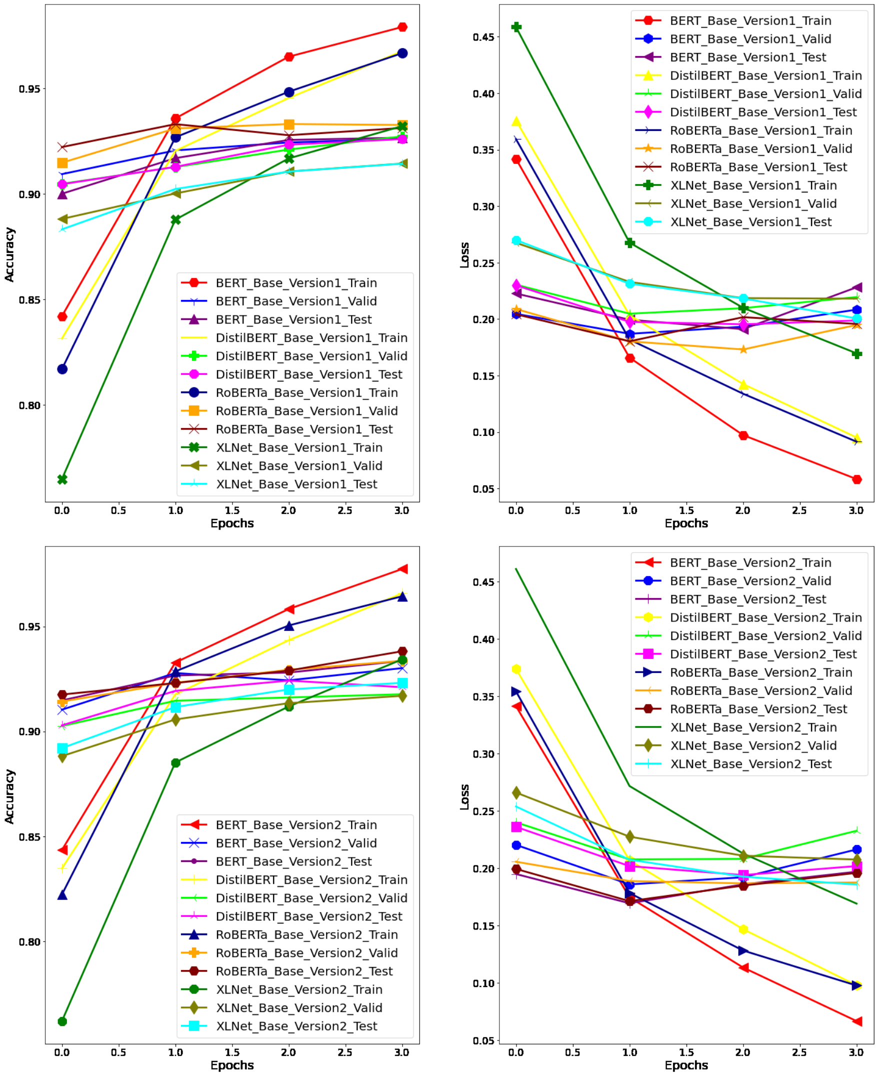

Fig. 2. Loss and accuracy learning curves for context-dependent feature-based classifiers

TABLE III

CLASSIFIERS PERFORMANCE FOR SARCASM DATASET VERSION1

Model	Class	Valid			Test		
		P	R	F1	P	R	F1
LogReg	Not Sarcastic	0.8406	0.8910	0.8651	0.8498	0.8873	0.8681
	Sarcastic	0.8491	0.7839	0.8152	0.8472	0.7995	0.8227
	Macro-average	0.8448	0.8375	0.8401	0.8485	0.8434	0.8454
	Micro-average	0.8440	0.8440	0.8440	0.8487	0.8487	0.8487
NBSVM	Not Sarcastic	0.8491	0.8718	0.8603	0.8556	0.8813	0.8682
	Sarcastic	0.8302	0.8019	0.8158	0.8421	0.8097	0.8256
	Macro-average	0.8397	0.8368	0.8380	0.8488	0.8455	0.8469
	Micro-average	0.8411	0.8411	0.8411	0.8499	0.8499	0.8499
BiGRU	Not Sarcastic	0.8801	0.8814	0.8807	0.8791	0.8826	0.8808
	Sarcastic	0.8481	0.8464	0.8472	0.8491	0.8447	0.8469
	Macro-average	0.8641	0.8639	0.8640	0.8641	0.8636	0.8639
	Micro-average	0.8661	0.8661	0.8661	0.8660	0.8660	0.8660
BERT-Base	Not Sarcastic	0.9366	0.9311	0.9338	0.9289	0.9413	0.9351
	Sarcastic	0.9125	0.9194	0.9160	0.9236	0.9078	0.9157
	Macro-average	0.9246	0.9252	0.9249	0.9263	0.9246	0.9254
	Micro-average	0.9260	0.9260	0.9260	0.9266	0.9266	0.9266
DistilBERT-Base	Not Sarcastic	0.9291	0.9422	0.9356	0.9334	0.9346	0.9340
	Sarcastic	0.9247	0.9081	0.9163	0.9162	0.9147	0.9155
	Macro-average	0.9269	0.9251	0.9259	0.9248	0.9246	0.9247
	Micro-average	0.9272	0.9272	0.9272	0.9259	0.9259	0.9259
RoBERTa-Base	Not Sarcastic	0.9285	0.9533	0.9407	0.9300	0.9486	0.9392
	Sarcastic	0.9382	0.9062	0.9219	0.9326	0.9087	0.9205
	Macro-average	0.9333	0.9297	0.9313	0.9313	0.9287	0.9299
	Micro-average	0.9326	0.9326	0.9326	0.9311	0.9311	0.9311
XLNet-Base	Not Sarcastic	0.9255	0.9214	0.9235	0.9216	0.9260	0.9238
	Sarcastic	0.9001	0.9052	0.9026	0.9047	0.8993	0.9020
	Macro-average	0.9128	0.9133	0.9131	0.9132	0.9126	0.9129
	Micro-average	0.9143	0.9143	0.9143	0.9143	0.9143	0.9143

5) DistilBERT-Based Classifier: DistilBERT is a neural network-based smaller general-purpose language model than BERT [19]. It represents the same architecture as BERT. However, the DistilBERT reduces the number of encoder-layers to half. The pre-trained DistilBERT model has 40% lesser trainable parameters and 60% faster performance than BERT. Moreover, it retains 97% of BERTs performance on various NLP tasks. For instance, the DistilBERT-Base contains 6 encoder layers, 768 hidden state representations, and 12 self-attention layers with 66M trainable parameters.

6) RoBERTa-Based Classifier: RoBERTa [20] is trained with simple modifications of BERT architecture that include longer training with more data and bigger batch sizes, removing the NSP loss, and dynamic mask pattern changes. Moreover, the RoBERTa is trained with 125K steps and 2K batch sizes to learn bidirectional context information of the input sequence. In this work, the authors use RoBERTa for the sarcasm detection task. This model has 12-encoder layers, 768-hidden units, 12-attention heads, and 125M parameters.

7) XLNet-Based Classifier: XLNet uses both autoregressive and auto encoding schemes. It captures bidirectional context using a permutation operation [21]. The permutation operation shares model parameters with all factorization orders. Moreover, the XLNet model adopts Transformer-XL schemes such as segment recurrent and relative encoding for pre-training tasks. It is designed with two architectures, namely, XLNet-base-cased, and XLNet-large-cased. In this work, the authors use the XLNet-base-cased model that has 12-encoder layers, 768-hidden units, 12-attention heads, and 117M parameters.

IV. RESULTS AND DISCUSSION

The authors performed supervised learning methods on two versions of the sarcasm headlines dataset. These methods were implemented in Google Colab Pro notebook using Python libraries with NVIDIA P100 GPU and 32GB RAM. Initially, the authors fix the broken Unicode in the texts using FTFY (fixes text for you) python library. The authors then converted upper case letters to lower case letters. Later, the authors applied a word contractions map to expand the shortened tokens. The authors divide these preprocessed data randomly into training (80%), validation (10%), and testing (10%) using stratified sampling. Table I shows the data distribution for all versions into training, validation, and testing. The first version of the sarcasm headlines dataset contains 21634 samples for the training set, 2404 samples for the validation set, and 2671 samples for the test set. Similarly, the second version of the sarcasm headlines dataset contains 23181 samples for the training set, 2576 samples for the validation set, and 2862 samples for the test set.

The authors employed four supervised learning methods on these datasets. Specifically, the authors used the NB-SVM, LR, BiGRU, BERT, DistilBERT, RoBERTa, and XLNet models. The BoW feature is used for both the NB-SVM and LR. In the BiGRU model, the unidirectional semantic context feature is used from the forward to backward and backward to forward directions. In the BERT-based fine-tuning models, the bidirectional context features are extracted from their respective pre-trained models for each input token from both directions. A unigram feature is considered for all models.

TABLE IV
CLASSIFIERS PERFORMANCE FOR SARCASM DATASET VERSION2

Model	Class	Valid			Test		
		P	R	F1	P	R	F1
LogReg	Not Sarcastic	0.8501	0.8451	0.8476	0.8736	0.8486	0.8609
	Sarcastic	0.8308	0.8362	0.8335	0.8385	0.8650	0.8516
	Macro-average	0.8404	0.8406	0.8405	0.8561	0.8568	0.8562
	Micro-average	0.8408	0.8408	0.8408	0.8564	0.8564	0.8564
NBSVM	Not Sarcastic	0.8555	0.8473	0.8514	0.8784	0.8479	0.8629
	Sarcastic	0.8339	0.8427	0.8383	0.8389	0.8709	0.8546
	Macro-average	0.8447	0.8450	0.8448	0.8586	0.8594	0.8587
	Micro-average	0.8451	0.8451	0.8451	0.8588	0.8588	0.8588
BiGRU	Not Sarcastic	0.8722	0.8755	0.8738	0.8688	0.8699	0.8693
	Sarcastic	0.8625	0.8590	0.8608	0.8567	0.8555	0.8561
	Macro-average	0.8674	0.8672	0.8673	0.8627	0.8627	0.8627
	Micro-average	0.8676	0.8676	0.8676	0.8630	0.8630	0.8630
BERT-Base	Not Sarcastic	0.9288	0.9385	0.9336	0.9326	0.9413	0.9369
	Sarcastic	0.9316	0.9209	0.9262	0.9348	0.9252	0.9299
	Macro-average	0.9302	0.9297	0.9299	0.9337	0.9332	0.9334
	Micro-average	0.9301	0.9301	0.9301	0.9336	0.9336	0.9336
DistilBERT-Base	Not Sarcastic	0.9159	0.9281	0.9219	0.9141	0.9373	0.9256
	Sarcastic	0.9198	0.9063	0.9130	0.9291	0.9032	0.9159
	Macro-average	0.9178	0.9172	0.9175	0.9216	0.9202	0.9207
	Micro-average	0.9177	0.9177	0.9177	0.9210	0.9210	0.9210
RoBERTa-Base	Not Sarcastic	0.9171	0.9600	0.9381	0.9226	0.9626	0.9422
	Sarcastic	0.9536	0.9046	0.9285	0.9569	0.9112	0.9335
	Macro-average	0.9354	0.9323	0.9333	0.9397	0.9369	0.9378
	Micro-average	0.9336	0.9336	0.9336	0.9382	0.9382	0.9382
XLNet-Base	Not Sarcastic	0.9303	0.9096	0.9198	0.9378	0.9146	0.9260
	Sarcastic	0.9029	0.9250	0.9138	0.9086	0.9332	0.9207
	Macro-average	0.9166	0.9173	0.9168	0.9232	0.9239	0.9234
	Micro-average	0.9169	0.9169	0.9169	0.9235	0.9235	0.9235

TABLE V
COMPARISON OF CLASSIFIERS WITH MICRO F1 SCORE

Model	Version1		Version2	
	Valid	Test	Valid	Test
LogReg	0.8440	0.8487	0.8408	0.8564
NBSVM	0.8411	0.8499	0.8451	0.8588
BiGRU	0.8661	0.8660	0.8676	0.8630
BERT-Base	0.9260	0.9266	0.9301	0.9336
DistilBERT-Base	0.9272	0.9259	0.9177	0.9210
RoBERTa-Base	**0.9326**	**0.9311**	**0.9336**	**0.9382**
XLNet-Base	0.9143	0.9143	0.9169	0.9235

In particular, the authors trained the proposed supervised learning methods using the following hyperparameters; 64 batch size, four epochs, 64 sequence length, 20000 maximum features, one-cycle learning rate policy (2e-5) [44] for the BERT-base, DistilBERT-base, RoBERTa-base, and XLNet models, and triangular learning rate policy (0.001) [45] for the NB-SVM, LR, and BiGRU. All models were evaluated using the standard classification metrics such as confusion matrix and Micro F1 and Macro F1-average scores [46]. The confusion matrix for all versions of datasets with four supervised learning models is shown in Table II. Table III and Table IV shows the performance of the proposed models for both datasets with validation, and testing. In these table, the RoBERTa-base fine-tuning model shows a better result than the NB-SVM, LR, BiGRU, BERT, DistilBERT, and XLNet for all versions of the sarcasm headlines dataset. Specifically, in Table V, the first version achieves 93.26% micro F1 for the validation data and 93.11% micro F1 for the test data.

Similarly, the second version of the dataset achieves 93.36% micro F1 for the validation data and 93.82% for the test data. Moreover, the context-dependent feature-based models achieve a better result than the BoW and context-independent feature models. Overall, the RoBERTa fine-tuning model achieves a higher micro F1 score in both versions of the sarcasm datasets. Moreover, the learning curve of the context-dependent feature-based models is shown in Figure 1 with loss and accuracy. This figure also indicates that the RoBERTa-base model achieves a better loss and accuracy for both datasets.

V. CONCLUSION

In this work, the authors presented the sarcasm detection task using supervised learning models. In particular, the authors performed sarcasm detection with the BoW features, context-independent features, and context-dependent features using six supervised learning models such as NB-SVM, LR, BiGRU, BERT, DistilBERT, RoBERTa, and XLNet. Our results show that the RoBERTa achieves a comparable result with the context-independent features. In particular, this model achieves 93.26% and 93.11% in version1 dataset for the validation and testing, respectively. Similarly, it achieves 93.36% and 93.82% in version2 dataset. In the future, the authors extend this task with gender information using dependency tree features.

ACKNOWLEDGMENT

This work was supported by the UGC (University Grants Commission), Government of India under UGC National Fellowship.

References

[1] S. S. Hossain, Y. Arafat, and M. E. Hossain, "Context-based news headlines analysis: A comparative study of machine learning and deep learning algorithms," *Vietnam Journal of Computer Science*, pp. 1–15, 2021.

[2] L. Oneto, F. Bisio, E. Cambria, and D. Anguita, "Statistical learning theory and ELM for big social data analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 45–55, 2016.

[3] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Statistical approaches to concept-level sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 6–9, 2013.

[4] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on twitter: A behavioral modeling approach," in *Proceedings of the eighth ACM international conference on web search and data mining*, 2015, pp. 97–106.

[5] I. Chaturvedi, Y.-S. Ong, I. Tsang, R. Welsch, and E. Cambria, "Learning word dependencies in text by means of a deep recurrent belief network," *Knowledge-Based Systems*, vol. 108, pp. 144–154, 2016.

[6] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Common sense computing: From the society of mind to digital intuition and beyond," in *Biometric ID Management and Multimodal Communication*, ser. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, 2009, vol. 5707, pp. 252–259.

[7] R. Misra and P. Arora, "Sarcasm detection using hybrid neural network," *arXiv preprint arXiv:1908.07414*, 2019.

[8] R. Misra and J. Grover, *Sculpting Data for ML: The first act of Machine Learning*, 01 2021.

[9] E. Cambria, S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi, "The CLSA model: A novel framework for concept-level sentiment analysis," in *LNCS*. Springer, 2015, vol. 9042, pp. 3–22.

[10] L. Xu and V. Xu, "Project report: Sarcasm detection."

[11] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–22, 2017.

[12] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.

[13] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *arXiv preprint arXiv:1804.09170*, 2018.

[14] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.

[15] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[21] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[22] S. Amir, B. C. Wallace, H. Lyu, P. Carvalho, and M. J. Silva, "Modelling context with user embeddings for sarcasm detection in social media," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 167–177.

[23] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "CASCADE: Contextual sarcasm detection in online discussion forums," in *COLING*, 2018, pp. 1837–1848.

[24] Y. A. Kolchinski and C. Potts, "Representing social media users for sarcasm detection," *arXiv preprint arXiv:1808.08470*, 2018.

[25] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," *arXiv preprint arXiv:1906.01815*, 2019.

[26] A. K. Jena, A. Sinha, and R. Agarwal, "C-net: Contextual network for sarcasm detection," in *Proceedings of the Second Workshop on Figurative Language Processing*, 2020, pp. 61–66.

[27] H. Nayel, E. Amer, A. Allam, and H. Abdallah, "Machine learning-based model for sentiment and sarcasm detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 386–389.

[28] M. Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.

[29] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter," in *Ninth international AAAI conference on web and social media*, 2015.

[30] M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, 2016, pp. 2449–2460.

[31] L. Liu, J. L. Priestley, Y. Zhou, H. E. Ray, and M. Han, "A2text-net: A novel deep neural network for sarcasm detection," in *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2019, pp. 118–126.

[32] S. Mukherjee and P. K. Bala, "Sarcasm detection in microblogs using naïve bayes and fuzzy clustering," *Technology in Society*, vol. 48, pp. 19–27, 2017.

[33] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 757–762.

[34] R. Speer, "ftfy," Zenodo, 2019, version 5.5. [Online]. Available: https://doi.org/10.5281/zenodo.2591652

[35] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Exploring impact of age and gender on sentiment analysis using machine learning," *Electronics*, vol. 9, no. 2, p. 374, 2020.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[37] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[38] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[40] J. S. Cramer, "The origins of logistic regression," 2002.

[41] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[42] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1539–1548, 2017.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[44] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.

[45] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[46] R. Alejo, J. Antonio, R. M. Valdovinos, and J. H. Pacheco-Sánchez, "Assessments metrics for multi-class imbalance learning: A preliminary study," in *Mexican Conference on Pattern Recognition*. Springer, 2013, pp. 335–343.