

# Evaluating the Quality of Data: Case of Sarcasm Dataset

Girma Yohannis Bade  
Olga Kolesnikova  
koleolga@gmail.com

Jose Luis Oropeza



---

## Research Article

**Keywords:** Sarcasm detection, natural language processing, deep learning, machine learning, state-of-the-art

**Posted Date:** December 23rd, 2024

**DOI:** <https://doi.org/10.21203/rs.3.rs-5678459/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Abstract

The models of artificial intelligence (AI) rely on data as their primary fuel. Accurate and efficient AI models that generated by high-quality data may guarantee AI-safe use. Sentiment analysis (SA), one of the tasks in natural language processing (NLP), highly relies on sarcasm detection. Sarcasm's cryptic character, however, makes it difficult and degrades its quality. Even though the problem was thoroughly investigated, it has been limited by the restrictions resulting from improper labeling and data not specifically gathered to identify sarcasm. This paper evaluates the quality of the sarcasm data by the performance of similarly parameterized models. To make an analysis, we compiled four distinct datasets—SARC, SemEval2022, NewsHeadline, and Multimodal. Undersampling and over-sampling techniques were used to balance the data size as well as class-label variations among the corpora. We perform extensive and fair evaluations on various models ranging from machine learning to transfer learning algorithms and employ TF-IDF vectorization and word embedding text representation techniques. Based on the experimental result, the NewsHeadline corpus exhibited greater quality, achieving a notable F1 score of 0.93 in RoBERTa's model performance. We have created a new (Sarcasm-Quality) dataset combining the best-performing datasets based on the experimental analysis and made it available for public use.

## 1 Introduction

In today's digital age, people frequently share their opinions on topics like products, politics, and entertainment through online platforms. This open expression allows businesses and policymakers to analyze public sentiment and improve their offerings [1–3]. However, sarcasm poses these sentiments as it obscures the true sentiments due to its subtle and often contradictory natures [4].

Sarcasm is a communication style that often involves making humorous remarks, but it can also serve as a means to subtly criticize, mock, insult, or express hidden discontent [1, 2, 4]. The apparent meaning of these remarks generally conflicts with their intended meaning. Understanding sarcasm involves an awareness of the context in which the sentiment was stated [5]. For example, sarcastic expressions such as "I finally found the lover that annoys me" or "Oh, great, another rainy day!" convey sentiments that are contrary to their literal meanings. Similarly, phrases like "Isn't this just the best day ever?" may be used to express dissatisfaction following an unfavorable event. The inherent contradiction between the literal and intended meaning of sarcasm presents a significant challenge for tasks such as sentiment analysis, opinion mining, and the interpretation of online reviews [6].

A classification task that determines whether a statement is sarcastic or not is called "sarcasm detection" [7]. This task needs the model trained on quality data for correct SA analytics [8, 9], however, the obscure nature of sarcasm and the data quality issues make the detection challenging and affect the decisions. Despite significant effort, the earlier sarcasm detection studies used the datasets that were collected through hashtag-based supervision, resulting in language and labeling noise [10], and much of the data was not specifically curated for this task. However, sarcasm's inherent ambiguity, coupled with these data quality issues, presents challenges that often impact detection accuracy because the detection process can be hindered by unclear labeling [7, 10–14]. The reason for these are 1) people use highly casual language on social media, which results in a sparse vocabulary and the lack of trained embeddings for many words. 2) it lacks contextual information in replied messages to identify sarcasm. 3) textual data (plain text) lacks indications like body language, facial expressions, speech tone, etc [11, 15]. Thus, the model that was trained with these kinds of datasets may suffer from finding genuine ironic (sarcastic) elements because of the noise present in several areas.

To address this issue, a manual labeling approach has been proposed [10, 16, 17]. However, this method is costly and resource-intensive, as it requires the collection of high-quality labels with an adequate representation of ironic samples. This occurs because of people's varying interpretations, and in many cases, there may be little understanding of sarcasm [10]. Consequently, different datasets are yielding inconsistent results due to varying dataset features even using similar methods [18]. To cope with this dilemma, we evaluate the most utilizable datasets and create a benchmarking new dataset based on the experimental analysis.

To this end, we compile the four publicly available and most frequently utilized sarcastic datasets namely SARC, Semeval2022, NewsHeadLine, and Multimodal. Then we apply unique preprocessing techniques including data cleaning, text embedding, and imbalance handling through some sampling techniques for all datasets. After preprocessing, we select the models from three levels of algorithms (machine, deep, transfer) learning and employ two-phase experimentation. Finally, we validate the compiled four datasets on the performance metrics of these models. The result shows that newsheadlines sarcasm dataset gained the highest metrics, and that indicates a better quality level of data. The main contributions of this paper can be summarized as follows:

- a. Assuring the quality of sarcasm data: We compile and evaluate the state-of-the-art datasets on the different models to assure the quality. By doing so, we provide insights for researchers who are interested in this research career.

- b. Benchmarking best-performing models : While examining the data quality, we evaluated the model performance too. This can help to benchmark the state-of-the-art model in this particular domain.
- c. Dataset benchmarking: Finally, we created our new benchmark dataset by amalgamating the compiled corpus. This also will boost the researchers' motivation in the NLP area for further investigation.

## 2 Review of Literature

Research communities' interest in examining social media and social networks has significantly increased in recent years. Platforms such as Twitter, Reddit, and others have become rich sources of data for a wide range of natural language processing (NLP) tasks. For instance, tweets and Reddit posts are frequently analyzed for: sentiment analysis [19–22], sarcasm detection [23–27], political opinions mining [28, 29], fake news identification [30, 31], hate speech detection [32], hope speech detection [33, 34], and so on. These and other tasks are not only valuable for understanding social discourse but also for practical applications such as market analysis, public opinion tracking, and the detection of harmful content. The growing complexity of social media data has pushed the boundaries of NLP, driving innovation in both algorithm design and computational techniques. The ability to classify and analyze such data is becoming increasingly relevant across industries as businesses, governments, and organizations to harness the power of social media to gain deeper insights into user behavior and public opinion. This is the reason why sentiment analysis is becoming the heart of classification [15].

However, this kind of tasks highly relies on the data that has no complication [35]. But in contradiction, many people intentionally express their ideas in a hidden manner when they want to amuse, insult, or criticize someone or something [1, 2]. This hidden nature vanishes the performance of sentiment analysis. To alleviate this issue, much research has been conducted using different algorithms and datasets. For example, Kumar et al [18] suggested deep learning techniques using convolution neural networks (convNet) and soft attention-based bidirectional long short-term memory (sAtt-BLSTM). They have used two different datasets from two sources: the SemEva2015 task and the 40,000 tweets with sarcasm. Their model achieved an accuracy of 93.71% and 97.87% for SemEva2015 and tweet datasets, respectively.

Similarly, Srivastava et al [36] have proposed a Hierarchical BERT-based model. It utilized datasets from two online forums, Twitter and Reddit, with 5,000 and 4,400 training samples and 1,800 test instances each. Gupta et al [12] conducted with 16,000 English tweets tagged with "irony," "sarcasm," and "not". The dataset has balanced class-labels i.e. an equal number of sarcastic and non-sarcastic tweets. Several machine learning techniques including KNN, Decision Tree, Random Forest, and SVM, were employed in their task. Among these, the voting classifier performed the best, achieving an accuracy of 83.53%, surpassing the other models tested.

Using news headline datasets from HuffPost and TheOnion, Misra et al [10] attempted sarcasm detection once more. Professional writers create news headlines, which guarantees good data quality and makes them a valuable asset for opinion-mining research.

On the other hand, forward embedding methods such as dependency weight-based features, Latent Semantic Analysis (LSA), GloVe, and Word2Vec were employed in combination with Support Vector Machines (SVM) for sarcasm detection, as demonstrated by [37]. Of these, Word2Vec and dependence weight-based features were more effective than GloVe and LSA. The primary datasets used for this study were sourced from 'GoodReads' and 'philosophy' domains. The resulting model achieved an F1-score of 72.53%.

In addition to English data, there has been an increasing research interest in this area for other languages too though progress remains limited (e.g., Dutch, Italian, Brazilian, Portuguese, Czech) [38]. Furthermore, the WANLP2021 shared task focused on sarcasm detection and sentiment analysis in Arabic [39]. Researchers developed a model called "SarcasmDet" using an ensemble approach with the ArBERT pre-trained model, achieving an F1-score of 59.89% on the test set. However, its performance was notably lower compared to models developed for English.

## 3 Methodology

### 3.1 Datasets

We compiled four (4) different datasets from past works and public repositories. Our study only focuses on the English corpus because sometimes model training algorithms do a favor on the perspective of language [40].

#### 3.1.1 SARC Dataset

This sarcasm corpus is collected from Reddit social media and deposited on Kaggle public repository. It has 1,010,827 labeled comments aimed at the sole dataset collection task. The studies like [41, 42] have utilized this corpus. Compared to other dataset collections in this study, SARC has a greater data size. However, it also needs serious pre-processing as some instances did not make sense.

### 3.1.2 SemEval2022 Dataset

This dataset is prepared for SemEva2022 shared task 6 [14, 18] and made available in GitHub repository for public use. It consists of 6,934 tweets with 'sarcastic' and 'non-sarcastic' two-class labels. Compared to SARC, this set is well-pre-processed but class labels are not balanced yet.

### 3.1.3 News Headlines Dataset

This resource is sourced from two news websites, theonion.com and huffingtonpost.com and deposited on Kaggle public repository. It consists of 26,709 tweets with two sarcastic and non-sarcastic classes. Several studies [10, 13, 43, 44] have utilized this corpus for their study. If the the headline is sarcastic, the class becomes 1, and otherwise 0. In contrast to Twitter and Reddit datasets, which are collected with hashtag-based collection methods and often contain informal or inconsistent language, the News Headlines dataset is crafted by professional writers, ensuring a higher level of linguistic quality and clarity [13].

### 3.1.4 Multimodal Dataset

This corpus includes several modalities, including text, audio, and image concepts. However, only text modality is available in the GitHub repository. Therefore, we utilize only this text modality data for our study. The dataset contains 31,449 records, each categorized into two class labels, sarcastic and non-sarcastic. In addition, another Mustard\_TTS multimodal dataset (audio and text) modalities [45] and MultilingualMultimodal('SarcNet') dataset were also introduced. However, since these dataset are bilingual (English and Chinese), they were not included in our analysis. Table 1 presents a summary of the datasets considered for this study.

Table 1  
Dataset Statistics. The column headings nonsarcastic and sarcastic represent the number of instances in class labels that were encoded into 0 and 1 representation respectively, for machine understandability.

Datasets' Name	#Data Size	#non-sarcastic	#sarcastic
SARC	1,010,827	505,413	505,413
SemEval2022	6,934	5,200	1,734
News Headlines	26,709	14,985	11,724
Multimodal	31,449	16,234	15,215

## 3.2 Preprocessing

Preprocessing is an essential step for ensuring data consistency and quality, next to accurate annotation or labeling. In the context of social media reviews, the unstructured nature of text introduces a variety of challenges, including the presence of colloquial expressions (e.g., "gonna," "wanna"), spelling errors, special characters, hyperlinks, and symbols [46]. These inconsistencies can create noise in the dataset and confuse the model during tokenization, potentially leading to poor classification performance [47]. To mitigate these issues, preprocessing becomes a vital task. Various preprocessing techniques such as stemming, stopword removal, and normalization can be employed for different NLP tasks depending on the specific problem at hand. However, in this study our preprocessing efforts centered on three key areas: cleaning the data, addressing class imbalance, and applying text encoding techniques to represent the text in a more machine-understandable manner. We mention their details in the following subsections.

### 3.2.1 Data Cleaning

Data cleaning includes identifying and correcting errors or inconsistencies in a dataset [48]. Common tasks during this step include the removal of punctuation marks (e.g., "?:-"), numbers (0–9), hyperlinks (e.g., "http:", "www", "HTML"), and special characters (e.g., "@#&"). Typically, these inaccuracies are achieved using the regex module. Additionally, columns containing missing values or entries that don't match the expected class labels are reviewed and corrected at this stage. The panda's library often flags such inconsistencies as NaN (Not a Number) when algorithms encounter undefined class labels. These issues can be managed either by manually inspecting and correcting the data or by automatically handling them with the panda's dropna method. Here, we chose the latter approach to manage missing data and maintain dataset integrity efficiently.

## 3.2.2 Handling Data Imbalances

The data and class-label imbalances are the two key issues that need to be addressed in the classification task. Imbalances significantly affect model performance, particularly class imbalance. In class imbalance, one class appears more frequently than the other [49]. In our case, non-sarcastic tweets are more common than sarcastic ones, an issue known as class-label imbalance [50]. Secondly, when working with multiple datasets, size imbalance can arise. For example, the SARC dataset is much larger than the others. To fairly evaluate the quality of dataset resources, we employ two sampling techniques: oversampling and undersampling [51]. In case of class label imbalance, we oversampled the minority sarcastic class ('1') in the size of the majority non-sarcastic class ('0') to ensure balanced class distributions in all datasets except SARC.

To address data imbalance, we applied both oversampling and undersampling techniques across the datasets. In the first oversampling approach, we increased the size of the SemEval2022 dataset and decreased the size of the SARC dataset to match the size of the Multimodal dataset. We left the NewsHeadline dataset unchanged, as its size was already comparable to Multimodal.

In the second undersampling approach, we reduced the sizes of all three larger datasets—SARC, NewsHeadline, and Multimodal to match the smaller size of SemEval2022. Although some data loss might occur, this approach offers a fair basis for evaluating data quality. To support the analysis, we conducted experiments under both conditions.

## 3.2.3 Data Encoding

Any model training algorithms require numerical data, so text inputs must be encoded into numerical representations [48]. For traditional machine learning algorithms, we used the TF-IDF vectorization technique [32]. In deep learning models, tokenization was done using "BERT AutoTokenizer" followed by an embedding layer to convert the tokens into numeric form.

Beside converting text data to numeric form, this technique extracts and puts the most identifying features in similarity vectors. The most sarcasm-indicating features are:

1) Prosodies: the words like "O, great, wow" following unpleasant context. 2) Subjectivity: flipping from positive to negative. For example, "I found the lover that annoys me." This sentence flipped from the positive word "lover" to the negative word "annoy".

3) Polarity indicators: Conjunction words like "but" or "in the same way" can indicate polarity. Generally, sarcastic sentences or dialogues possess huge contrasts of sentiments. That's the characteristic that we exploit and train our model. Figure 1 shows how text representation techniques encode raw texts into a numeric form where models start training.

## 3.3 Models

We employ the models from three levels of algorithms. These are machine learning (ML), deep learning (DL), and transfer learning (TL). We mention their details below.

ML algorithms are mainly divided into five categories: supervised, unsupervised, semi-supervised, self-supervised, and reinforcement learning. In the case of supervised learning, input-output pairs are given in the labeled data; in unsupervised learning, patterns are found in unlabeled data. Both labeled and unlabeled data are used in semi-supervised learning. Self-supervised learning generates labels from the data itself, and reinforcement learning adapts based on feedback [52]. To evaluate datasets as a baseline, we utilized four supervised ML algorithm models: Logistic Regression, Multinomial Naive Bayes, Random Forest, and Support Vector Classifier (SVC). With these models, we used a TF-IDF vectorizer with tri-grams (`ngram_range = 3`) to extract features in numeric.

DL is another state-of-the-art approach, particularly through artificial neural networks (ANNs), which are designed to solve complex problems by using multiple layers of interconnected neurons, inspired by the structure of the human brain. The use of layered architecture combined with non-linear activation functions allows DL models to capture intricate patterns within data effectively [53]. Thus to assess data quality in depth, we employ three models: Convolutional Neural Networks (CNN), Bi-directional Long Short-Term Memory (Bi-LSTM), and Recurrent Neural Networks (RNN) models from the DL algorithm [54–56].

Thirdly, we examine the same datasets further using transfer learning models (TL). Today, TL is representing a significant advancement in AI architecture. It utilizes pre-trained models that have already learned essential features from large datasets, reducing the need for extensive training data and computational resources while enhancing model generalization capabilities [15, 57, 58]. From TL algorithm models, we used two prominent models: Bidirectional Encoder Representations from Transformers (BERT) and the Robustly Optimized BERT Pretraining Approach (RoBERTa).

BERT was first presented by Devlin et al [59]. The main technological advancement of BERT is the use of Transformers trained in both directions for language modeling. It has two objectives: next-sentence prediction (NSP) and masked language modeling (MLM). In the transformer layer, we employ the pre-trained models "bertbase-uncased" which includes 110M parameters, 12 layers, 12 attention heads, and 768 hidden state sizes. Each hidden state has an embedding output of 768 lengths and a size of (max seq len, 768).

RoBERTa is an enhanced version of BERT, designed to boost performance by utilizing more extensive training data, larger batch sizes, and eliminating certain pretraining constraints found in the original BERT model. It was developed by the Facebook AI Research (FAIR) team for natural language processing (NLP) tasks. RoBERTa introduces several refinements to BERT’s architecture, resulting in improved performance across a range of NLP applications [60]. The task pipeline of this work is illustrated in Fig. 2.

## 4 Experiments and Results

### 4.1 Experiments on Oversampled Version

In Section 3.2.2, we introduced two methods: oversampling and undersampling. This section is the implementation of oversampling. The system oversamples small datasets and minority class labels first. To this end, the SemEval2022 dataset is expanded to match the size of the Multimodal dataset, plus all minority class labels further balanced to equal the majority classes.

Firstly, we conducted initial experiments on ML classifiers namely Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Multinomial Naive Bayes (MNB). We employ the TF-IDF word representation method with tri-grams (n-gram = 3) [61] for these models. 20% of data from each corpus was reserved to test the performances of selected models. Then we train each 90% corpus on selected models and evaluate the test data to assess dataset quality. Metrics such as accuracy, precision, recall, and F1 scores were used, with decisions primarily based on F1 scores due to their balance of precision and recall.

Next, we evaluated the same datasets using deep learning models (CNN, BiLSTM, and RNN). Consistent parameters were applied across all three models: 5 epochs, a batch size of 32, a dense layer activation of ‘relu’ (at 64 units), an output layer activation of ‘sigmoid,’ and a maximum token length of 128.

Lastly, we applied transfer learning models (BERT and RoBERTa) with hyperparameters outlined in Table 2 to enhance our assessment. These hyperparameters define the structure and guide the behavior of the training process [62].

Table 2  
The hyperparameters of BERT  
and Roberta on oversampled  
version.

Hyperparameters	Values
learning rate	1e-5
batch_size	32
epochs	5
weight_decay	0.01

Table 3  
The performance results of selected models on oversampled data.

Models	SARC				SemEva2022				NewsHeadlines				Multimodal			
ML	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
LR	0.73	0.73	0.75	0.74	0.99	0.99	0.99	0.99	0.93	0.92	0.95	0.93	0.92	0.93	0.90	0.91
SVM	0.80	0.82	0.76	0.79	0.99	0.99	0.99	0.99	0.96	0.96	0.95	0.96	0.94	0.95	0.93	0.94
RF	0.78	0.79	0.75	0.77	0.99	0.99	0.99	0.99	0.94	0.94	0.94	0.94	0.94	0.96	0.90	0.93
MNB	0.75	0.72	0.84	0.77	0.99	0.99	0.99	0.99	0.95	0.95	0.96	0.95	0.93	0.92	0.95	0.93
DL																
CNN	0.64	0.66	0.65	0.65	0.99	1.0	0.99	0.99	0.85	0.81	0.86	0.84	0.81	0.82	0.76	0.79
BiLSTM	0.64	0.67	0.64	0.65	0.99	0.99	0.99	0.99	0.84	0.83	0.80	0.81	0.80	0.79	0.80	0.79
RNN	0.62	0.64	0.64	0.64	0.99	0.99	0.98	0.98	0.82	0.83	0.73	0.78	0.78	0.76	0.79	0.78
TL																
BERT	0.70	0.71	0.70	0.70	0.99	0.99	0.99	0.99	0.94	0.94	0.94	0.94	0.89	0.88	0.91	0.90
RoBERTa	0.70	0.71	0.68	0.70	0.99	0.99	0.99	0.99	0.93	0.93	0.93	0.93	0.92	0.92	0.92	0.92

Table 3 shows the results of each model and in that SemEva2022 was expanded to match the size of Multimodal (31,449 tweets), SARC was undersampled to align with Multimodal's size, and all class labels except for SARC (already balanced) were oversampled.

The results synthesized from the oversampled datasets are the following. Except for the last one, RNN managed to achieve an F1 score of 0.98, Semeval2022 achieved an impressive F1 score of 0.99 for eight out of nine models. Because it was the sole dataset that was subjected to oversampling, it has a lot of repetitive values increasing the probability of overfitting. In the case of NewsHeadlines, the maximum F1 score was recorded at 0.96 using the SVM model in the ML category. Only the BERT model in TL outperformed its peers with an F1 score of 0.94 while in DL, only BiLSTM had an F1 score of 0.84. Both datasets of NewsHeadlines and Multimodal were benchmarked as they are without any oversampling or undersampling. But their class labels are adjusted so that they correspond to the sizes of the majority class to avoid bias

For ML, Multimodal scored F1 of 0.94 with SVM: the highest score. For DL, it showed comparable results as F1 score 0.79 among all the models except RNN and F1 score of 0.90 for TL. Multimodal class labels were oversampled, so that they had the same size as majority class. The lowest F1 score using SVM in the ML category performed for SARC is 0.79. In DL, it recorded an average F1 score of 0.65, with RNN slightly lower at 0.64. In TL, both models achieved an F1 score of 0.74. Unlike other datasets, the SARC dataset was undersampled to match the size of the Multimodal dataset. The SARC dataset was undersampled to match the size of the Multimodal dataset.

Figure 3 demonstrates the dataset quality validation in an oversampled version.

While all four datasets performed well, their metrics were not used to confirm data quality due to the presence of duplicates from oversampling. This introduces a risk of overfitting, indicating that additional validation is necessary to ensure a reasonable assessment.

## 4.2 Experiment on Under-sampled Version

For this version of the experiment, we followed the same method as we did in Section 4.1, including model selection, encoding methods, and evaluation metrics. However, all datasets here were downsampled to align with the size of the SemEval2022 dataset. In addition, the class labels are balanced by downsampling majority classes to match minority classes. For the DL and TL models, we used the same hyperparameters as in the oversampled experiment, except for setting the number of epochs to 10, as undersampled data requires less processing time per run. Table 4 presents the results from each model in the undersampling experiment. We use the results of this undersampled version to validate data quality.

In Table 4 presents the results on the undersampled datasets in which other datasets are resized to match the size of the SemEval2022 dataset. Additionally, class labels are also downsampled and balanced without duplication. The F1 score served as the key metric to assess dataset quality and model performance.

The key performances from this version are as follows:

1. The NewsHeadlines dataset achieved an F1 score of 0.93 with the RoBERTa model. In the ML approach, it reached an F1 score of 0.87 with the SVM model, and 0.78 with BiLSTM of DL.
2. Multimodal dataset has gained the second highest F1-score 0.88 with BERT. It also achieved an F1 score of 0.82 on SVM, and 0.73 on CNN.
3. SemEva2022 dataset has achieved the third highest F1-score 0.81 in SVM. It also achieved a 0.79 F1-score with BiLSTM and 0.76 on RoBERTa.
4. In SARC, the highest F1 score is 0.80 with SVM. It also gained an F1 score of

Table 4  
The performance results of selected models on undersampled data.

Models	SARC				SemEva2022				NewsHeadlines				Multimodal			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
ML																
LR	0.70	0.70	0.69	0.70	0.80	0.79	0.84	0.81	0.84	0.83	0.86	0.84	0.79	0.86	0.70	0.77
SVM	0.78	0.74	0.87	0.80	0.79	0.75	0.89	0.81	0.86	0.82	0.92	0.87	0.82	0.84	0.81	0.82
RF	0.69	0.76	0.56	0.64	0.75	0.73	0.80	0.76	0.81	0.83	0.78	0.81	0.77	0.93	0.58	0.71
MNB	0.75	0.79	0.68	0.73	0.80	0.78	0.86	0.81	0.84	0.90	0.78	0.83	0.81	0.89	0.71	0.79
DL																
CNN	0.62	0.62	0.66	0.64	0.89	0.88	0.68	0.77	0.80	0.81	0.74	0.77	0.76	0.78	0.69	0.73
BiLSTM	0.60	0.62	0.56	0.59	0.89	0.86	0.74	0.79	0.79	0.77	0.80	0.78	0.74	0.77	0.65	0.71
RNN	0.56	0.58	0.48	0.53	0.74	0.52	0.51	0.51	0.80	0.81	0.74	0.77	0.72	0.74	0.63	0.68
TL																
BERT	0.64	0.67	0.64	0.65	0.69	0.70	0.65	0.68	0.92	0.94	0.90	0.92	0.89	0.91	0.85	0.88
RoBERTa	0.78	0.80	0.75	0.75	0.76	0.79	0.73	0.76	0.93	0.95	0.91	0.93	0.88	0.92	0.84	0.87

0.75 on RoBERTa and 0.64 on CNN.

Figure 4 illustrates the dataset quality in bar chart.

Finally, the overall rank is NewsHeadline, Multimodal, SemEva2022, and SARC in this descending order.

## 5 Benchmark Dataset and Its Performance

We created the Sarcasm-Quality dataset, which is suited for tasks from this viewpoint, by combining related but separately accessible datasets into a single resource. Although earlier studies have mostly concentrated on small, individual datasets, there is still a lack of comprehensiveness regarding robustness, which hinders the advancement of future study [63]. To close this gap and support upcoming studies in the sarcasm detection challenge, the Sarcasm-Quality benchmark dataset was created.

This new benchmark dataset is made from the combination of SemEva2022, NewsHeadline, and Multimodal because they showed relatively high quality during evaluation. Additionally, they were originally collected from different sources and manually labeled for sarcasm detection purposes [10].

After creating this dataset, we also evaluate it to see how it performs. For its evaluation, we select a RoBERTa model from transfer learning algorithms because it has outperformed the others during evaluation. We have applied the default hyperparameters of RoBERTa on the new dataset and it achieved an F1-score of 0.92. This shows that the optimization has been working still combining the better performing features.

## 6 Conclusion and Recommendation



This research study presents an efficient method for analyzing sarcasm datasets. It compares four distinct datasets—SARC, Semeva2022, NewsHeadline, and Multimodal created for similar purposes. To evaluate these datasets, three hierarchical layers of algorithm models were tested. To mitigate bias, two essential preprocessing techniques; cleaning and sampling, were implemented alongside model selection.

The experimental results revealed that the NewsHeadline dataset outperformed the others, showcasing its potential to retain distinctive features of sarcasm in text and informing the development of more accurate and effective models. Among the models evaluated from the three algorithms, the RoBERTa model, utilizing transfer learning, achieved the highest F1 score of 0.93.

Based on these findings, a new benchmark dataset was created by integrating the best-performing datasets as a foundational standard. To further enhance progress in this field, it is recommended that multilingual datasets should be evaluated against this benchmark to ensure a consistent quality level, ultimately promoting the development of a more robust and adaptable model.

## Declarations

## Acknowledgements

The work was carried out with the partial support of the Mexican Government through grant A1-S-47854 from CONACYT, Mexico, grants 20241816, 20241819, and 20240951 from the Research and Postgraduate Secretariat of the Polytechnic Institute National, Mexico. The authors thank CONACYT for the computing resources. presented through the Deep Learning Platform for Technologies of the Language of the Supercomputing Laboratory of the INAOE, Mexico and ac- Learn about Microsoft's support through the Microsoft Latin America Doctoral Award.

### Conflict of Interests

We authors declare that there is no conflict of interest among us. We aim to widen our knowledge and skill level in the area of artificial intelligence and share it with the global community through publication.

### Data source availability

The datasets are already available in <https://www.kaggle.com/datasets/danofer/sarcasm>, <https://github.com/AmirAbaskohi/SemEval2022-Task6-Sarcasm>, <https://www.kaggle.com/datasets/saurabhbagchi/sarcasm-detection>, <https://github.com/headacheboy/data-of-multimodal-sarcasm-detection> and made available upon request from corresponding author.

## References

1. Mayur, Wankhade (2022) Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* 55(7):5731–5780
2. Walaa Medhat A, Hassan, Korashy H (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng J* 5(4):1093–1113
3. Bing, Liu et al (2010) Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010
4. Yacoub AD, Slim S, Aboutabl A (2024) A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends. *Int J Electr Comput Eng Syst* 15(1):69–78
5. Singh GV, Firdaus M, Chauhan DS (2024) Asif Ekbal, and Pushpak Bhattacharyya. Well, now we know! unveiling sarcasm: Initiating and exploring multimodal conversations with reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18981–18989
6. Ellen Riloff A, Qadir P, Surve LD, Silva (2013) Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714
7. David Bamman and Noah Smith (2015) Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577
8. Tonja AL (2021) Michael Melese Woldeyohannis, and Mesay Gameda Yigezu. A parallel corpora for bi-directional neural machine translation for low resourced ethiopian languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 71–76. IEEE

9. Tonja AL, Azime IA, Belay TD, Yigezu MG, Mehamed MA, Ayele AA, Jibril EC, Woldeyohannis MM, Kolesnikova O, Slusallek P et al Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. arXiv preprint arXiv:2403.13737, 2024.
10. Rishabh Misra and Prahal Arora (2023) Sarcasm detection using news headlines dataset. AI Open 4:13–18
11. Diana G, Maynard, Mark A, Greenwood (2014) Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In Lrec 2014 proceedings. ELRA
12. Gupta R, Kumar J, Agrawal H et al (2020) A statistical approach for sarcasm detection using twitter data. In. *4th international conference on intelligent computing and control systems (ICICCS)*, pages 633–638. IEEE, 2020
13. Rishabh Misra and Prahal Arora (2019) Sarcasm detection using hybrid neural network. arXiv preprint arXiv:1908.07414
14. Amirhossein Abaskohi A, Rasouli TZ, Bahrak B (2022) UTNLP at SemEval-2022 task 6: A comparative analysis of sarcasm detection using generative-based and mutation-based data augmentation. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval- pages 962–969*, Seattle, United States, July 2022. Association for Computational Linguistics. 10.18653/v1/2022.semeval-1.135. URL <https://aclanthology.org/2022.semeval-1.135>
15. Kalaivani A, Thenmozhi D Sarcasm identification and detection in conversation context using BERT. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 72–76, Online, July 2020. Association for Computational Linguistics. 10.18653/v1/2020.figlang-1.10. URL <https://aclanthology.org/2020.figlang-1.10>
16. Gavin Abercrombie and Dirk Hovy (2016) Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In Proceedings of the ACL 2016 student research workshop, pages 107–113
17. Joshi A, Tripathi V, Bhattacharyya P, Carman M (2016) Harnessing sequence labeling for sarcasm detection in dialogue from tv series ‘friends’. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 146–155
18. Kumar A, Sangwan SR, Arora A, Nayyar A, Abdel-Basset M et al (2019) Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE access 7:23319–23328
19. Md Arif M, Hasan SAA, Shiam MP, Ahmed et al (2024) Mazharul Islam Tusher, Md Zikar Hossan, Aftab Uddin, Suniti Devi, Md Habibur Rahman, Md Zinnat Ali Biswas,. Predicting customer sentiment in social media interactions: Analyzing amazon help twitter conversations using machine learning. International Journal of Advanced Science Computing and Engineering, 6(2):52–56
20. Ahmad Amjad Mir (2024) Sentiment analysis of social media during coronavirus and its correlation with indian stock market movements. Integr J Sci Technol, 1(8)
21. Yiming Pan B, Wu H, Zheng Y, Zong, Wang C (2024) The application of social media sentiment analysis based on natural language processing to charity. In The 11th International scientific and practical conference Advanced technologies for the implementation of educational initiatives(March 19–22, 2024) Boston, USA. International Science Group. 254 p., page 216, 2024
22. Paulraj D, Ezhumalai P, Prakash M et al (2024) A deep learning modified neural network (dlmnn) based proficient sentiment analysis technique on twitter data. J Exp Theor Artif Intell 36(3):415–434
23. Iraisha Fadilah and Agus Wijayanto (2024) Sarcasm in social media: A study of comments on sam smith’s instagram posts. Jurnal Onoma: Pendidikan Bahasa dan Sastra 10(1):92–104
24. Vidyullatha Sukhavasi V, Dondeti et al (2024) Sarcasm detection using optimized bi-directional long short-term memory. Knowl Inf Syst, pages 1–29
25. Chetana Thaokar JK, Rout (2024) Minakhi Rout, and Niranjana Kumar Ray. N-gram based sarcasm detection for news and social media text using hybrid deep learning models. SN Comput Sci 5(1):163
26. Sharma DK, Singh B, Agarwal S, Pachauri N, Alhussan AA, Hanaa A, Abdallah (2023) Sarcasm detection over social media platforms using hybrid ensemble model with fuzzy logic. Electronics 12(4):937
27. Rajnish Pandey and Jyoti Prakash Singh (2023) Bert-lstm model for sarcasm detection in code-mixed social media post. J Intell Inform Syst 60(1):235–254
28. Aniket K, Shahade KH, Walse VM, Thakare, Atique M (2023) Multi-lingual opinion mining for social media discourses: An approach using deep learning based hybrid fine-tuned smith algorithm with adam optimizer. Int J Inform Manage Data Insights 3(2):100182
29. Ratnapuri CI, Karmagatri M, Kurnianingrum D, Utama ID, Darisman A (2023) Users opinion mining of tiktok shop social media commerce to find business opportunities for small businesses. J Theoretical Appl Inform Technol 101(1):214–222

30. Femi Olan U, Jayawickrama EO, Arakpogun J, Suklan, Liu S (2024) Fake news on social media: the impact on society. *Inform Syst Front* 26(2):443–458
31. Girma Bade O, Kolesnikova G, Sidorov, José, Oropeza (2024) Social media fake news classification using machine learning algorithm. In Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors, *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29, St. Julian's, Malta, mar 2024. Association for Computational Linguistics. URL <https://aclanthology.org/dravidianlangtech-1.4>
32. Girma Bade O, Kolesnikova G, Sidorov, José, Oropeza (2024) Social media hate and offensive speech detection using machine learning method. In Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors, *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/dravidianlangtech-1.40>
33. Girma Yohannis Bade O, Koleniskova (2024) José Luis Oropeza, Grigori Sidorov, and Kidist Feleke Bergene. Hope speech in social media texts using transformer. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, colocated with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org
34. Yigezu MG, Bade GY, Kolesnikova O (2023) Grigori Sidorov, and Alexander F Gelbukh. Multilingual hope speech detection using machine learning. *IberLEF@ SEPLN*
35. Girma Yohannis Bade and Akalu Assefa Afaro (2018) Object oriented software development for artificial intelligence. *Am J Softw Eng Appl* 9(3):22–24
36. Himani Srivastava V, Varshney S, Kumari, Srivastava S (2020) A novel hierarchical bert architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97
37. Joshi A, Tripathi V, Patel K, Bhattacharyya P, Carman M Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883*, 2016.
38. Tomáš Ptáček I, Habernal, Hong J (2014) Sarcasm detection on czech and english twitter. In *COLING 2014, the 25th International Conference on Computational Linguistics*, pages 213–223
39. Dalya Faraj and Malak Abdullah (2021) Sarcasmdet at sarcasm detection task 2021 in arabic using arabert pretrained model. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 345–350
40. Girma Yohannis Bade (2021) Natural language processing and its challenges on omotic language group of ethiopia. *J Comput Sci Res* 3(4):26–30
41. Jens Lemmens B, Burtenshaw E, Lotfi I, Markov, Daelemans W (2020) Sarcasm detection using an ensemble approach. In *proceedings of the second workshop on figurative language processing*, pages 264–269
42. Y Alex Kolchinski and Christopher Potts (2018) Representing social media users for sarcasm detection. *arXiv preprint arXiv:1808.08470*
43. Rasikh Ali T, Farhat S, Abdullah S, Akram M, Alhajlah (2023) Awais Mahmood, and Muhammad Amjad Iqbal. Deep learning for sarcasm identification in news headlines. *Applied Sciences*, 13(9), ISSN 2076–3417. 10.3390/app13095586. URL <https://www.mdpi.com/2076-3417/13/9/5586>
44. Parnavi Shrikhande V, Setty, Sahani A (2020) Sarcasm detection in newspaper headlines. In *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, pages 483–487. IEEE
45. Tan Yue X, Shi R, Mao ZH, Cambria E (2024) Sarcnet: A multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14325–14335
46. Mersha MA, Bade GY, Kalita J, Kolesnikova O, Gelbukh A et al (2024) Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. *Procedia Comput Sci* 244:133–142
47. Kokab ST, Asghar S, Naz S (2022) Transformer-based deep learning models for the sentiment analysis of social media data. *Array* 14:100157
48. Girma Yohannis Bade and Hussien Seid (2018) Development of longest-match based stemmer for texts of wolaita language. 4:79–83
49. Kushankur Ghosh C, Bellinger R, Corizzo P, Branco (2024) Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Mach Learn* 113(7):4845–4901

50. Soujanya Poria E, Cambria D, Hazarika, Vij P A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint arXiv:1610.08815, 2016.
51. Eunnuri Cho T-W Chang, and, Hwang G (2022) Data preprocessing combination to improve the performance of quality classification in the manufacturing process. *Electronics*, 11(3), ISSN 2079–9292. 10.3390/electronics11030477. URL <https://www.mdpi.com/2079-9292/11/3/477>
52. Batta Mahesh (2020) Machine learning algorithms-a review. *Int J Sci Res (IJSR)* [Internet] 9(1):381–386
53. CM Suneera and Jay Prakash (2020) Performance analysis of machine learning and deep learning models for text classification. In 2020 IEEE 17th India council international conference (INDICON), pages 1–6. IEEE
54. Manjunath Jogin MS, Madhulika GD, Divya RK, Meghana S, Apoorva et al (2018) Feature extraction using convolution neural networks (cnn) and deep learning. In. 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT), pages 2319–2323. IEEE, 2018
55. Prasnurzaki Anki and Alhadi Bustamam (2021) Measuring the accuracy of lstm and bilstm models in the application of artificial intelligence by applying chatbot programme. *Indonesian J Electr Eng Comput Sci* 23(1):197–205
56. Ruishuang Wang Z, Li J, Cao T, Chen, Wang L (2019) Convolutional recurrent neural networks for text classification. In 2019 international joint conference on neural networks (IJCNN), pages 1–6. IEEE
57. Yigezu MG, Mersha MA, Bade GY, Kalita J, Kolesnikova O, Gelbukh A (2024) Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. arXiv preprint arXiv:2410.02609
58. Amardeep Kumar and Vivek Anand (2020) Transformers on sarcasm detection with context. In *Proceedings of the second workshop on figurative language processing*, pages 88–92
59. Amardeep Kumar and Vivek Anand Transformers on sarcasm detection with context. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 88–92, Online, July 2020. Association for Computational Linguistics. 10.18653/v1/2020.figlang-1.13. URL <https://aclanthology.org/2020.figlang-1.13>
60. BV Kumar and Manchala Sadanandam (2024) A fusion architecture of bert and roberta for enhanced performance of sentiment analysis of social media platforms. *Int J Comput Digit Syst* 15(1):51–66
61. Bade GY, Kolesnikova O (2024) José Luis Oropeza, and Grigori Sidorov. Lexicon-based language relatedness analysis. *Procedia Comput Sci* 244:268–277
62. Yigezu MG, Kolesnikova O, Sidorov G, Alexander F, Gelbukh (2023) Transformer-based hate speech detection for multi-class and multi-label classification. *IberLEF@ SEPLN*
63. Tonja AL, Kolesnikova O, Gelbukh A, Sidorov G (2023) Low-resource neural machine translation improvement using source-side monolingual data. *Appl Sci* 13(2):1201

## Figures

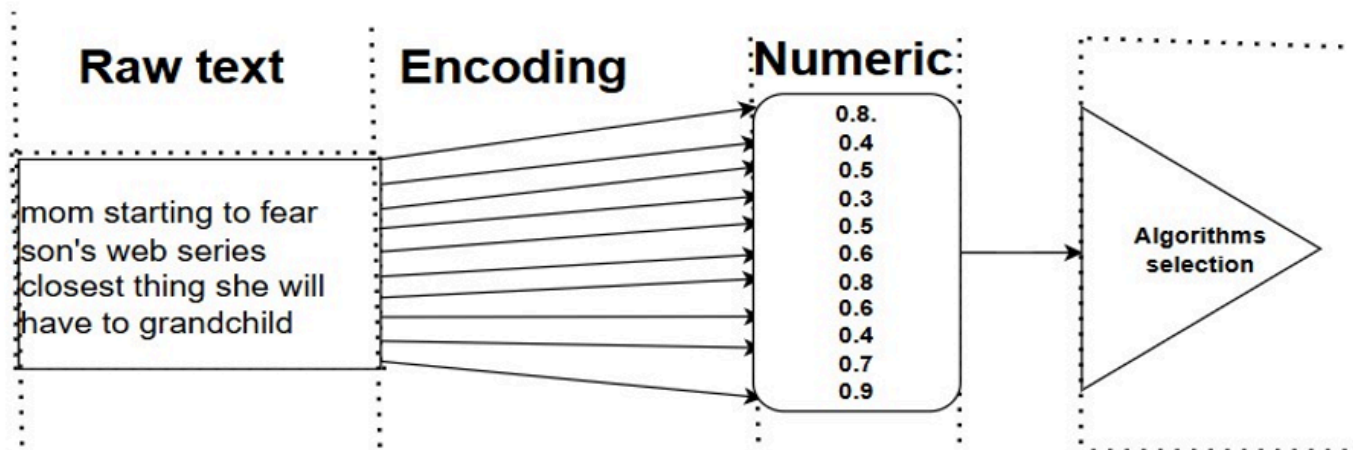


Figure 1

A block diagram that shows how vectorization and embedding techniques encode raw text into numeric form.

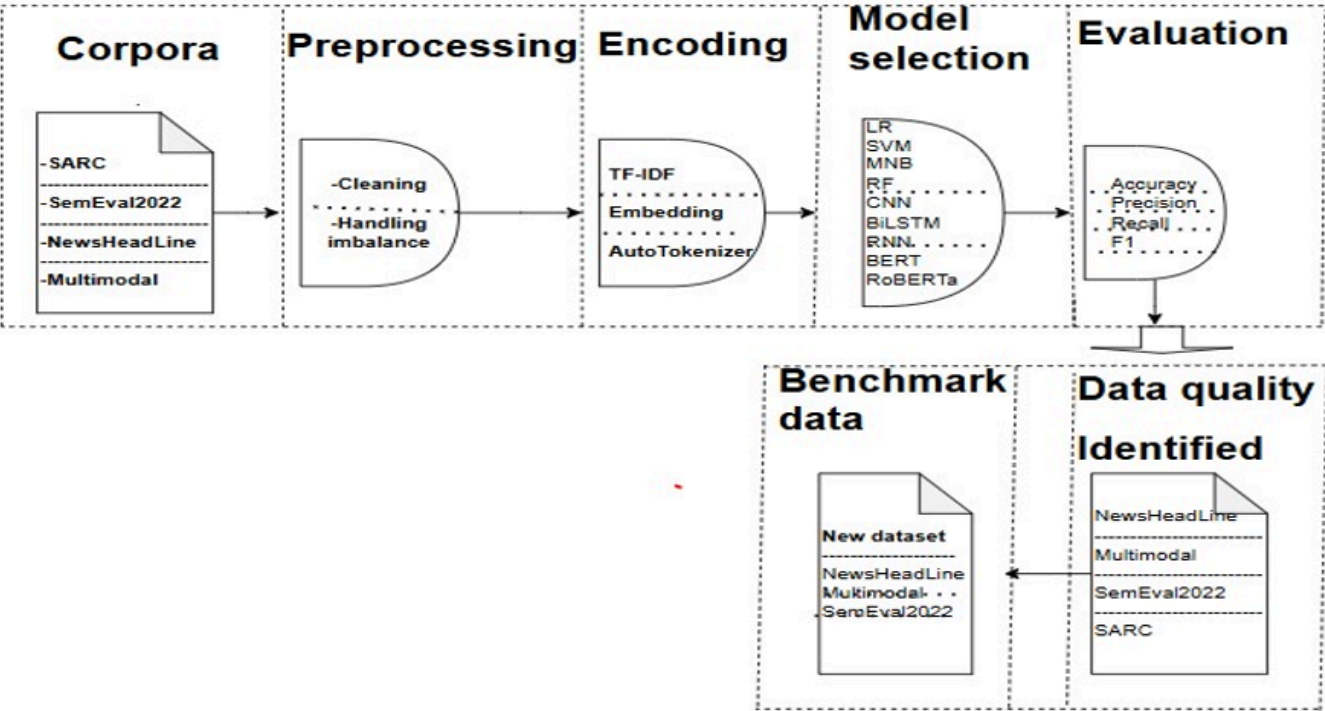


Figure 2

The task pipeline of this study

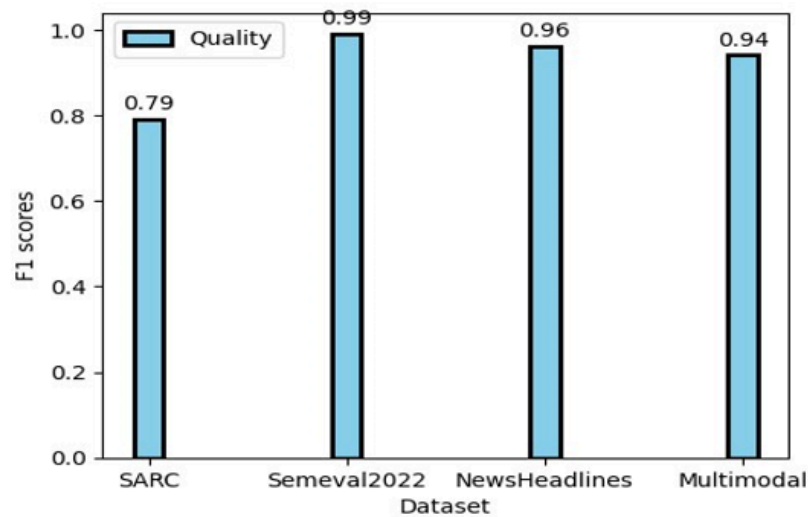
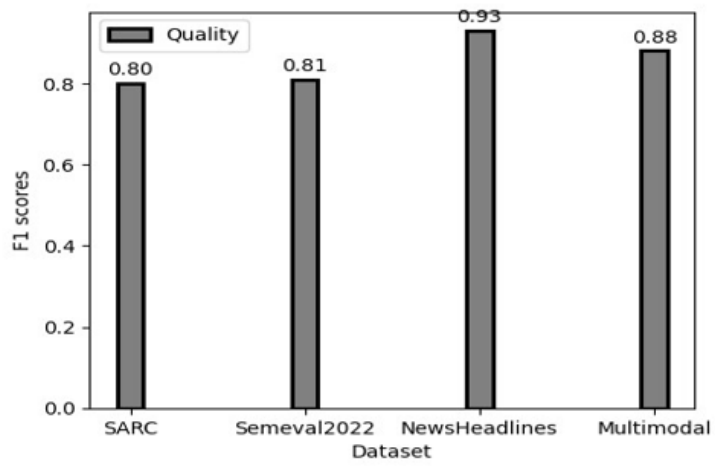


Figure 3

The result of over-sampled data quality



**Figure 4**

Data quality in undersampled data