Sentiment analysis for Urdu online reviews using deep learning models

Iqra Safder[1] | Zainab Mehmood[1] | Raheem Sarwar[2] | Saeed-Ul Hassan[1] | Farooq Zaman[1] | Rao Muhammad Adeel Nawab[3] | Faisal Bukhari[4] | Rabeeh Ayaz Abbasi[5] | Salem Alelyani[6,7] | Naif Radi Aljohani[8] | Raheel Nawaz[9]

[1]Department of Computer Science, Information Technology University, Lahore, Pakistan

[2]Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK

[3]Department of Computer Science, COMSATS University Lahore, Lahore, Pakistan

[4]Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan

[5]Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan

[6]Center for Artificial Intelligence (CAI), King Khalid University, Abha, Saudi Arabia

[7]College of Computer Science, King Khalid University, Abha, Saudi Arabia

[8]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[9]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

Correspondence
Saeed-Ul Hassan, Department of Computer Science, Information Technology University, Pakistan.
Email: saeed-ul-hassan@itu.edu.pk

Funding information
King Khalid University

Abstract

Most existing studies are focused on popular languages like English, Spanish, Chinese, Japanese, and others, however, limited attention has been paid to Urdu despite having more than 60 million native speakers. In this paper, we develop a deep learning model for the sentiments expressed in this under-resourced language. We develop an open-source corpus of 10,008 reviews from 566 online threads on the topics of sports, food, software, politics, and entertainment. The objectives of this work are bifold (a) the creation of a human-annotated corpus for the research of sentiment analysis in Urdu; and (b) measurement of up-to-date model performance using a corpus. For their assessment, we performed binary and ternary classification studies utilizing another model, namely long short-term memory (LSTM), recurrent convolutional neural network (RCNN) Rule-Based, *N*-gram, support vector machine , convolutional neural network, and LSTM. The RCNN model surpasses standard models with 84.98% accuracy for binary classification and 68.56% accuracy for ternary classification. To facilitate other researchers working in the same domain, we have open-sourced the corpus and code developed for this research.

KEYWORDS
artificial intelligence, deep learning models, sentiment analysis, Urdu online reviews

1 | INTRODUCTION

With the worldwide web's increasing adaptation, social media networks have become a critical means of sharing knowledge and contact across the globe (Borgman et al., 2019; Yu et al., 2020; Sarwar et al., 2021). The world wide web has transformed into a dynamic collection of user and corporate generated contents, where anyone can contribute (Hassan et al., 2017; Muneer et al., 2019). Millions of users use blogs, fora and social networking websites to express their views about personalities, events, places, and products (Edara et al, 2019; Jahangir et al., 2017). Sentiment analysis (SA) is of great importance in providing an understanding of people's attitudes and giving insights into the behavioural analysis (Martınez-Camara et al., 2014; Hathlian & Hafez, 2017; Liu et al., 2019). It is beneficial both in discouraging the spread of misinformation from extremist elements and in promoting commercial interests (Alotaibi et al., 2020). It is used to understand customers' behaviours for devising marketing strategies. Moreover, customer service, campaign success and product dissemination can be improved by analysing the sentiments (Asghar et al., 2019; Asghar et al., 2018; Hassan & Haddawy, 2015; Jarwar et al., 2017; Lee et al., 2019).

SA is referring to the subjective interpretation behind a user's phrase. It uses the techniques of natural language processing computational linguistics, text analysis, and machine learning (ML) to determine whether the term is positive, negative, or neutral (Imran et al., 2018; Melo et al., 2019; Śmieja et al., 2019). SA is also used in opinion mining for determining the attitudes of people towards products, places, and other entities. The significance of SA can be appreciated by our need to know the attitudes of people towards various issues (Alotaibi et al., 2020). More recently, the methods around SA have attracted the interest of practitioners with advancements in technology and the widespread use of social networking websites and online marketing (Arshad et al., 2019; Safder et al., 2018; Safder et al., 2017; Sarwar & Hassan, 2015). For instance, in the US presidential elections 2012, the Obama administration used SA to gauge public opinion to campaign messages and policy announcements. Moreover, companies can improve their reputation by determining customers' satisfaction or dissatisfaction towards their products and services using SA. Furthermore, SA is important in forecasting marketing trends through the sentiments extracted from news, blogs, and fora. Many organizations are affected by comments on social media and blogs. They depend on client reviews and try to incorporate the application of SA into their systems to get its benefits (Nagarajan & Gandhi, 2019; Safder & Hassan, 2018; Sarwar & Nutanong, 2016; Shardlow et al., 2018; Zheng et al., 2019). The primary requirement in carrying out SA efficiently is the availability of a corpus. It is the key in understanding how sentiments are conveyed on various fora, blogs or websites. The purpose of a corpus in SA is to train the machine learning models with high accuracy. Unfortunately, most of the corpora available as a resource for SA are in English or other popular languages (Ananiadou et al., 2013; Batista-Navarro et al., 2013; Hassan et al., 2020; Hassan et al., 2020; Hassan et al., 2017; Limkonchotiwat et al., 2020; Sarwar et al., 2020).

Urdu is a member of the Indo-Aryan language family. Urdu is the national language of Pakistan and is widely spoken in the Indian subcontinent. It uses Arabic script in cursive format (Nastaliq style) with the segmental writing system. Specifically, the Urdu language is based on an 'abjad' system where the long vowels and consonants are necessarily written while the short vowels (diacritics) are optional. It is a bidirectional language where the numerals are written from left-to-right, while the characters are written from right-to-left. When characters are joined to make the words, they develop different shapes based on the context. Specifically, a character can have a maximum four shape variants known as initial, medial, final and isolated. The characters that can develop all four shapes are known as joiners, while the characters that can only have two shapes (final and isolated) are known as non-joiners. A large number of online resources, such as blogs and various websites, enable users to express their views or opinions in Urdu. In addition, a significant number of people worldwide, particularly in Southwest Asia, use Urdu to interact in real life and on social media sites such as Twitter and Facebook (Asghar et al., 2019; Sarwar et al., 2018; Sabah et al., 2019; Sarwar et al., 2020). The challenge of performing SA in Urdu has not been thoroughly explored due to its grammatical and morphological characteristics. Nonetheless, implementations for SA in Urdu are still limited, primarily because of the following challenges.

- ***Lack of consideration.*** The present internet resources are preponderated by popular languages like English, Chinese, Spanish and others. Therefore, these prevalent languages have been the primary academic subject in recent decades. Moreover, the inimitable obstacles raised by the inherent features of Urdu have impeded the delayed study interests for the Urdu script.
- ***Differences from other languages.*** There are numerous intrinsic differences between Urdu and other popular languages, making the existing Sentiment Analysis techniques inapt to Urdu. For instance, lack of capitalization, grammatical and morphological characteristics, and free word order.
- ***Lack of a large Sentiment Analysis corpus.*** While few Urdu SA corpora have been created yet either these corpora are not openly accessible or not as huge as other popular languages. Hence, it is impeding the developments and assessment of Urdu SA techniques.

Contributions of this investigation:

1. Thus, keeping all the above narrated challenges in the view, this study aims to contribute a large benchmark corpus for SA of Urdu, hereafter called the SAU-18 Corpus. This proposed corpus was constructed by collecting 10,008 reviews from various domains, including sports, food, software, politics, and entertainment. Human annotators manually tagged the reviews into positive ($n = 3662$), negative ($n = 2619$), and neutral ($n = 3727$) categories.
2. We demonstrate how the SAU-18 corpus can be used for Urdu SA growth and evaluation. It is expected that the SAU-18 corpus will help (a) promote research in Urdu, a language which is under-resourced; (b) to make a clear comparison of contemporary SA methods in Urdu; and (c) create and test new methods in Urdu SA.
3. We present a solution that relies on a state-of-the-art deep learning model. We performed comprehensive experimental studies to compare our solution against the competitive methods.

The remainder of the paper is arranged as follows; Section 2 introduces a formal literature review followed by a thorough debate on the generation of corpus (see Section 3). The Section 4 focuses on the characteristics of the corpus. Section 5 discusses the employed approaches for the task of SA. Section 6 discusses results and the evaluation of employed deep learning model with classic machine learning and deep learning models. Eventually, the paper is completed and guidance for the future is given in the Section 7.

2 | LITERATURE REVIEW

Recent years have seen an overwhelming research on sentiment analysis and opinion mining. Additionally, enormous reports and competitions have been carried out to develop benchmark corpora and techniques for SA. Therefore, we categorize the related work into three sections; the first section presents the details of different corpora developed for open competitions. The second one discussed machine learning techniques developed for SA. The last section covers the studies and techniques designed for Urdu SA.

2.1 | Benchmark corpora

Efforts have been made in literature to establish SA benchmark corpora, the most conspicuous being the series of SemEval competitions.[1] These competitions have helped us in understanding the semantics of various natural languages. Different tasks are set at each competition, using multiple corpora to evaluate semantic analysis systems. Which have resulted in a range of standard corpora along with the contemporary SA techniques. Such corpora were developed specifically for English and Arabic (Kiritchenko et al., 2016).

Each year SemEval generates corpora of different sizes and characteristics from multiple data sources. Twitter and SMS datasets were included in the 2014 version. The Twitter dataset made up about 15,000 tweets, and the SMS dataset consisted of about 2000 messages. The tasks 4 and 9 were related to SA. The interest in SA grew over time, as in 2015 four tasks (9–12) were related to SA. In 2016 tasks 4–7, in 2017 tasks 4–8, in 2018 tasks 1–3, in 2019 tasks 3–6 and finally in 2020 tasks 7–10 were related to SA or related topics (Ayata et al., 2017; Nakov et al., 2016).

In addition to SemEval competitions, SA was also run for other languages such as Indonesian, Korean, Italian and German. The Korean corpus, KOSAC, comprises of approximately 8000 sentences chosen from the annotated Sejong corpus newspaper papers using Korean subjectivity markup language (Jang et al., 2013). A corpus for German product reviews was assembled by securing Amazon product reviews using Amazon's review parser.[2]

Growing sentences in the corpus are annotated based on their specific frame of reference. Sixty-three thousand sixty-seven sentences were derived from different commodity domains (Boland et al., 2013). Using the Twitter Streaming API, an Indonesian tweet corpus consisting of 5.3 million tweets was developed. The deployed model used the Tweets geo-location to process tweets in Indonesian (Wicaksono et al., 2014). In addition, an Italian corpus composed of 2648 movie-related sentences has been established for aspect-based SA (Sorgente et al., 2014).

2.2 | Sentiment analysis techniques

The literature has suggested different approaches for SA (Khan et al., 2016; Aung et al., 2019; Masood et al., 2020, Osmani et al., 2020, Basiri et al., 2020). Turney (2002) developed an unsupervised methodology for the semantic interpretation of the film genre entitled Thumbs up or Thumbs down. The proposed approach emphasizes on determining which particular polarities of the phrases have adjectives or adverbs. Another approach used artificial neural networks with recursive least squares back-propagation training algorithm for SA (Safder et al., 2020; Safder & Hassan, 2019). In the SemEval 2014 edition, Wagner et al. (2014) used both unsupervised (a rule-based approach) and supervised (support vector machine [SVM]) machine learning for SA. They developed a rule-based method to identify polarity in feelings using a lexicon and then turned them into features used by supervised machine learning algorithms.

Likewise, in the SemEval 2016, Kiritchenko et al. (2016) used three supervised machine learning algorithms random forest (RF), Linear Regression and Gaussian Regression to provide a score between 0 and 1 indicating the term's strength of association with the positive sentiment. These scores were evaluated through usage of Kendall's rank correlation coefficient and Spearman's rank correlation. Moreover, in SemEval 2017, two systems were deployed for classification (Ayata et al., 2017). The first one was based on word embeddings for the feature representation and classification of tweets, using SVM, RF and Naive Bayes (NB). The second module focused on long short-term memory (LSTM), which uses word indexes to describe features as input sequences. In addition, Dos Santos and Gatti (2014) and Attardi and Sartiano (2016) proposed a deep convolutional neural network that uses phrase-level information to perform SA. El-Beltagy et al. (2017) used a series of Convolutional Neural Network, Multilayer Perceptron and Logistic Regression models for the classification and tweet quantification of the subject-based message polarity. Ali et al. (2019) proposed a Word2Vec model with a fuzzy ontology-based semantic knowledge in order to improve the transportation features extraction task and text classification using Bidirectional LSTM approach. Fuzzy ontology help describe the semantic knowledge about entities, features and their relation in transportation domain. Moreover, in order to store and analyse healthcare data and improve the classification accuracy Ali et al. (2020) developed a healthcare monitoring system which relies on the cloud environment and a big data analytics engine. This engine relies on ontologies, data mining methods and Bi-LSTM. Li et al. (2020) proposed conversational sentiment analysis system which is faster, compact and parameter-efficient. This system relies on a generalized neural tensor block which is followed by a two-channel classifier and is designed to perform sentiment classification and context compositionality, respectively.

2.3 | Sentiment analysis for Urdu

Researchers have also attempted to establish corpora and methods for Urdu sentiment analysis (Mahmood et al., 2020; Mukhtar et al., 2018). Nevertheless, SA's task in the Urdu language was not discussed in detail. Syed et al. (2010) performed SA in Urdu, focused on lexicons. The proposed technique worked in two phases: 1st for creating a sentiment annotated lexicon, then making a classification model which would process alongside classify text. A dataset of 753 reviews (361 positives and 392 negatives) was used for experimentation, comprising 435 movie reviews and 318 product reviews. The intended approach was based on the identification and extraction of senti-units, using shallow parsing, to identify words that convey the sentiment of the whole sentence.

Almas and Ahmad (2007) applied SA on financial trading texts for English, Arabic, and Urdu. For Urdu, a 1.03 million token corpus was developed; comprising financial news items published between 2006 and 2007 in a major daily newspaper in Pakistan. To identify sentiments for text, they used a local grammar, constructed using a bottom-up approach. Moreover, a classification mechanism was proposed to distinguish subjective sentences from objective sentences, using linear SVM and the vector space model. The corpus was obtained from BBC Urdu and parsed using an in-house HTML parser to produce cleaned data. The sentences were annotated according to the Multi Perspective Question Answering standards set for English[3] (Mukund & Srihari, 2010).

Mukund et al. (2011) used sequence kernels to identify opinion entities in an Urdu corpus consisting of news articles from BBC Urdu.[4] Firstly, they constructed opinion-entity candidates and combined them with opinion expressions to generate candidate sequences. Secondly, they used SVM with a combination of linear and sequence kernels. Structural correspondence learning was also used in another work to move SA learning from Urdu newswire data to Urdu blog data. Furthermore, they validated their approach by using machine learning algorithms (Mukund & Srihari, 2012).

Rajput (2014) constructed an annotation framework to annotate Urdu texts. The framework used a domain-specific ontology created manually using the domain knowledge and context keywords. Their corpus constituted 350 online car advertisements obtained from a popular Urdu newspaper Jang.[5] Additionally, Syed et al. (2014) proposed a lexicon-based classification approach for Urdu data SA by extracting senti-units. Each sentence was divided into the source, senti-unit (appraisal) and the target of the appraisal. Each sentence was associated with its target to avoid misclassification. Zafar et al. (2016) performed SA on tweets corpus of controversial topics in Pakistan. Using Twitter Streaming, a data collection of around 6000 random Twitter users across Pakistan was obtained. API. Hashtags from tweets on four controversial topics covering media, foreign policy, politics, and religion were used to collect tweets. Furthermore, a retweet graph was constructed from the data, which was further divided into two communities.

Rehman and Bajwa (2016) attempted to build a lexicon-based SA using a publicly available lexicon consisting of 2607 positive and 4728 negative sentiment words. The polarity of a sentence was calculated using the tokens of the comment and the sentiment lexicon. The corpus was generated by extracting data from Urdu news websites[6] and user opinions from a blog.[7] Khan et al. (2017) conducted SA in Urdu using an English lexicon. For this purpose, four sentiment lexicons were built from four English lexicons: Affective Norms for English Words; English lexicon developed by Finn Årup Nielsen; SenticNet; and National Research Council Canada (NRC) Word-Emotion Association. Each lexeme in each lexicon was translated into Urdu using Google Translator. The results showed that the NRC Word Emotion Association Lexicon gave better results, with 60.24% accuracy.

Mukhtar and Khan (2018) carried out an Urdu SA study using three supervised machine-learning techniques: decision trees (DT), K-nearest neighbours (KNN) and SVM. The results were compared and improved using feature extraction. It was observed that KNN performed better than SVM and DT. The dataset used for this purpose was constructed from 14 topics and was annotated by two annotators. Hassan and Shoaib (2018) used a method of analysing sub-opinions in Urdu text to determine the overall polarity of a sentence. For this purpose, two datasets were collected: the first consisted of 443 reviews related to cars and cosmetics and the second consisted of 401 reviews related to electronic appliances. In comparison to the baseline bag-of-words technique, their proposed method increased precision by 8.46%, recall by 37.25% and accuracy by 24.75%.

Note that Urdu is an under-resourced language lacking publicly available corpora and lexicons. It has morphological complexity that makes SA for Urdu more challenging. Riaz (2007) stated that very few researches in the Information Retrieval (IR) community have focused on the challenges in Urdu stemming and many of the techniques developed for SA in other languages are not applicable to Urdu. Moreover, existing annotated corpora are not large enough, cover only a few topics and have binary classes (positive and negative). The lack of large-scale resources is a significant obstacle in carrying out research on SA in Urdu. Therefore, in this paper, we present a novel contribution by generating an Urdu language corpus of data covering five topics extracted from multiple social media platforms. We classify data into three classes: positive, negative, and neutral. Besides, we have applied a state-of-the-art deep learning method (RCNN) on this corpus. Recurrent convolutional neural network (RCNN) has not been used in Urdu for SA, to the best of our understanding.

3 | CORPUS GENERATION

This segment discusses the measures involved in developing the SAU-18 Corpus including raw data processing, annotation method (annotation guidelines, annotations and inter-annotator agreement), corpus standardization and corpus characteristics.

3.1 | Data collection

We collected data from online websites in order to create a gold standard SAU-18 corpus for Urdu SA that provides free access to its contents. The main reasons for selecting online data repositories are: (a) they are free and readily available; (b) data scrapping is not prohibited from their platforms (c) since Urdu is an under-resourced language, therefore it is difficult to collect a large amount of data from online resources; (d) text on online websites is available in a digital format that is easy to use for corpus generation; and (e) reviews are available for various genres, which helps to create a more realistic and challenging benchmark corpus. In our case, the data was gathered from genres which are popularly discussed and have wide coverage such as dramas, films, along with chat shows; meals in addition to ingredients; politics; sports; plus apps, websites, seminars, together with tools.

The research team manually extracted the reviews from websites mentioned in Table 1. This data is publically available; therefore, it does not come under any violation in terms of services of these websites. A total of 10,008 reviews were collected from 566 online threads. Initially, each review (a collection of sentences) was stored in an Excel file, along with the following information: (a) review ID; (b) review theme; (c) the review's URL; (d) the date of compilation; and (e) annotation tag. The corpus was later also translated into XML format.

3.2 | Annotation process

This section explains the annotation process, including the preparation of the annotation instructions, the manual annotation of human annotator's feedback and the calculation of the inter-annotator agreement (IAA).

3.2.1 | Annotation guidelines

Firstly, we prepared the annotation guidelines and separated sentences from each review. Next, we tagged each sentence of the review. The polarity of each sentence was calculated by applying the criteria for the annotation, adapted from those of the different current corpora for SA as mentioned below. Table 2 provides several instances of reviews of a positive, negative and neutral type.

TABLE 1 Web sources of data collection

Genre	Sources	No of reviews
Films, chat shows, plays	www.reviewit.pk, www.tweettunnel.com, www.urduweb.org, www.dramasonline.com, www.dailydose.pk, www.fashionuniverse.net, www.hamariweb.com, www.zemtv.com, www.siasat.pk	2000 Reviews
Meal and Ingredients	www.urduweb.org, www.paksitan.web.pk, www.friendskorner.com, www.facebook.com, www.kfoods.com	1507 Reviews
Civics	www.siasat.pk, www.twitter.com	2000 Reviews
Apps, forums, gadgets	www.baazauq.blogspot.com, www.dufferistan.com, www.mbilalm.com, www.urduweb.org, www.urdupoint.com, www.itdarasgah.com, www.urdudaan.blogspot.com, www.itforumpk.com, www.itdunya.com, www.achidosti.com, www.mobilesmspk.net, www.tafrehmella.com, www.sachiidosti.com	2501 Reviews
Games	www.urduweb.org, www.tafrehmella.com	2000 Reviews

TABLE 2 Examples of the negative, positive, and neutral review instances

Examples of the positive review (translation)	Examples of the negative review (translation)	Examples of the neutral review (translation)
(it's a good attempt)	(there are errors in the software)	(I think it will rain during the match)
(Very interesting app)	(This plugin is not functional with Wordpress)	(I have watched Shawshank redemption)
(Congratulations)	(This is not working)	(National game of Pakistan is hockey)

- Positive rules:
  - A sentence is classified as positive if it communicates a good feeling about all the terms of the aspect or the context on which the statement is made (Pontiki et al., 2016).
  - If a sentence expresses both neutral and positive sentiment, then positive feeling overcomes over the neutral one. This phrase is declared as a good sentence.
  - Agreements and approvals are labelled positive (Abdul-Mageed & Diab, 2012).
  - Illocutionary Speech Acts-actions such as apology, gratitude, appreciation and a constructive statement (Abdul-Mageed & Diab, 2012).
- Negative rules:
  - A statement is defined as negative because it communicates a bad feeling about the word element (Maynard & Bontcheva, 2016).
  - If the sentence contains more negative terms and fewer positive ones, it is known to be negative.
  - Direct un-softened disagreements are considered as negative (Abdul-Mageed & Diab, 2012).
  - Banning, bidding, penalizing and evaluating makes derogatory sentences (Rehman & Bajwa, 2016).
  - When a negative word comes with a positive adjective, then the sentence is considered as negative (Ganapathibhotla & Liu, 2008).
  - If a simple negation happens without a positivity or negativity in a sentence, then it is called a negative sentence.
- Neutral rules:
  - When a sentence includes some truthful details than it is marked as neutral (Boland et al., 2013).
  - When in a sentence, thought is exchanged, then it is marked as neutral (Boland et al., 2013).
  - Terms such as possibly (shayad) minimize the degree of certainty and liability; thus, these words are called neutral (Abdul-Mageed & Diab, 2012).
  - A phrase with many attributes and sources that convey both positive and negative attitudes or emotions is defined as neutral (Pontiki et al., 2016).

3.2.2 | Annotation and inter-annotator agreement

We conducted manual annotation on sentence-level for each review with the help of three independent annotators (A, B and C). The annotators were all graduates, Urdu native speakers and acquainted with SA's mission. The annotation process was carried out in two steps. In the first step, an initial set of 100 reviews was manually annotated by Annotators A and B on a sentence level, using the annotation guidelines. Conflicting pairs were discussed, and the annotation guidelines were duly revised. The updated annotation guidelines were used by Annotators A and B to annotate the sentences from the remaining 9908 reviews. Following annotation of the entire corpus, Annotator C annotated the contrasting sentences. In order to measure the IAA, we computed a standard metric such as Cohen's Kappa (Artstein & Poesio, 2008; Cohen, 1960) to evaluate the quality of the annotated data. Note that IAA simply computes the agreement between annotators; while Cohen's Kappa also added a chance adjustment to determine how much better the annotators did than chance (see Equation (1)).

$$\text{Cohen's Kappa} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

Although Pr(a) is the probability of consensus between the two annotators, Pr(e) is the possibility that the two annotators decide by chance. In our case, we found 0.66 Cohen's Kappa score. Furthermore, we also computed the IAA percentage by simply calculating the agreement between annotators and achieved a 78.01% IAA score. These scores are good, considering the differentiation between the three classes. This also highlights the fact that the annotation guidelines were well defined and adhered to by the annotators during the annotation process. Our conflict review found that most conflicts existed when differentiating between the positive and neutral classes (11.01%) and the negative and neutral classes (7.79%). Also, note that we have a verbal consent of annotators to use this dataset without any restrictions. Annotations are the co-authors of this study as well.

4 | CORPUS CHARACTERISTICS AND STANDARDIZATION

Our proposed SAU-18 Corpus comprises 10,008 Urdu reviews: 36% positive instances; 26% negative instances; and 37% neutral instances, as seen in Table 3. From these statistics, it can be noted that our proposed corpus comprises a good class balance. Also, SAU-18 Corpus is free and publicly available for research purposes.[8]

5 | DATA AND METHODS

This segment demonstrates the implementation of a deep learning model named RCNN (Lai et al., 2015; Safder & Hassan, 2019) using SAU-18 corpus. Two methods were used to test the model: binary classification, and ternary classification. The following sections describe the dataset details for the experiments, the applied techniques, the evaluation methodology and finally the evaluation measures are discussed.

5.1 | Datasets

The datasets used to conduct experiments of SA belong to five genres as mentioned in the above section. For the binary classification, we only withdraw and segmented the positive and negative reviews from our dataset. For ternary classification, we identified reviews in the positive, negative, and neutral classes of the dataset. The explanations of the datasets used for the two experiments are given in Table 4.

5.2 | Employed approaches

This section contains the detail of the RCNN model that we deployed on our SAU-18 corpus along with details of the rule-based and *N*-gram based techniques used for the comparative analysis.

TABLE 3 Corpus statistics of Urdu dataset

Type	Statistics
No of positive reviews	3662 reviews
No of negative reviews	2619 reviews
No of neutral reviews	3727 reviews
Total no of tokens	175,399 tokens
Total no of types	16,487 types
Minimum length of reviews	1 word
Maximum length of reviews	208 words
Average length of reviews	18 words
Total no of reviews	10,008 reviews

TABLE 4 Dataset split for sentiment analysis experiments

Dataset	Class	Train set	Test set
Binary classifier	2 (positive and negative)	5024 instances	1257 instances
Ternary classifier	3 (positive, negative and neutral)	8000 instances	2008 instances

5.2.1 | Main method: Deep sentiments by recurrent convolutional neural network

Firstly, we performed pre-processing on review sentences to remove any junk characters that may come during data parsing. Afterward, the pre-processed sentences are fed to the RCNN model that classify them either positive or negative for binary classification and positive or negative or neutral for ternary classification (see Figure 1). Lai et al. (2015) have published detailed information on RCNN, its development, learning, and evaluation. Briefly, the RCNN model is a combination of the recurrent neutral network (RNN) and convolutional neural network (CNN) model. RNN has the ability to analyse word-by-word text and to store text contextual information in a hidden layer. It is called a biased model, though, because it prefers recent words, and this may influence a text's semantics. CNN was intended to address the drawback. CNN extracts relevant and valuable words in a text through a max-pooling layer utilizing a fixed convolutional kernel-like fixed window size, which makes learning more difficult. The RCNN model was then implemented to address the shortcomings of both RNN and CNN versions.

This model's main idea is to construct a representation of a word by adding a bi-directional RNN followed by the max-pooling layer. The actual word description comprises of the left context, resulting from the forwarding of RNN, the word embedding and the correct context, derived from the backward RNN. This unique property helps to perform much better than conventional neural network models that use a lesser part of the information about a text. The following are the architectural details of the deployed RCNN model.

$$c_l(w_i) = f\left(\left(W^l\right)c_l(w_{i-1}) + \left(W^{sl}\right)e(w_{i-1})\right) \tag{2}$$

$$c_r(w_i) = f((W^r)c_r(w_{i-1}) + (W^{sr})e(w_{i-1})) \tag{3}$$

$c_l(w_i)$ is defined here as the left context of the word w_i and $c_r(w_i)$ as the right context of the word w_i. The left and right meaning of a term w_i is determined using Equations (2) and (3) where $e(w_{i-1})$ is the word embedding of the word w_{i-1}, which is a vector of real value. $c_l(w_{i-1})$ is the left context of the preceding word w_{i-1}.

W^l is a matrix that converts the hidden layer into the next layer hidden up. W^{sl} is a matrix that blends the current word semantics with the left meaning of the next word. f is a non-linear activate function. The meaning on the right side $c_r(w_i)$ is also measured in a similar way as shown in Equation (3). The meaning matrix incorporates the semantics of both left-and right-side contexts. For instance, at Figure 1, $c_l(w_6)$ encodes the left-hand meaning semi context from word w_1 to w_6.

Equation (4) defined the word w_i, which is the concatenation of the left- context vector $c_l(w_i)$, the word embedding $e(w_i)$, and the right-context vector $c_r(w_i)$. It's Equation (4) introduces the final input vector x_i for term w_i which is then passed into a regular layer where a linear transformation is used to it along with the tan$h$ function.

$$x_i = [c_l(W_i) : e(w_i); c_r(W_i)] \tag{4}$$

The resultant vector y represents a semantic vector having the most useful textual features used for the text representation. When all word representations are determined, a max-pooling layer is used, as seen in the Equation (5).

$$y_i^{(1)} = \max^n\left[\tanh\left(W^{(1)}x_i + b^{(1)}\right)\right] \tag{5}$$

FIGURE 1 Overview of our deep sentiment recurrent convolutional neural network architecture: The pre-processed Urdu sentences are fed to the recurrent convolutional neural network (RCNN) model that classify them either positive or negative for binary classification and positive or negative or neutral for ternary classification

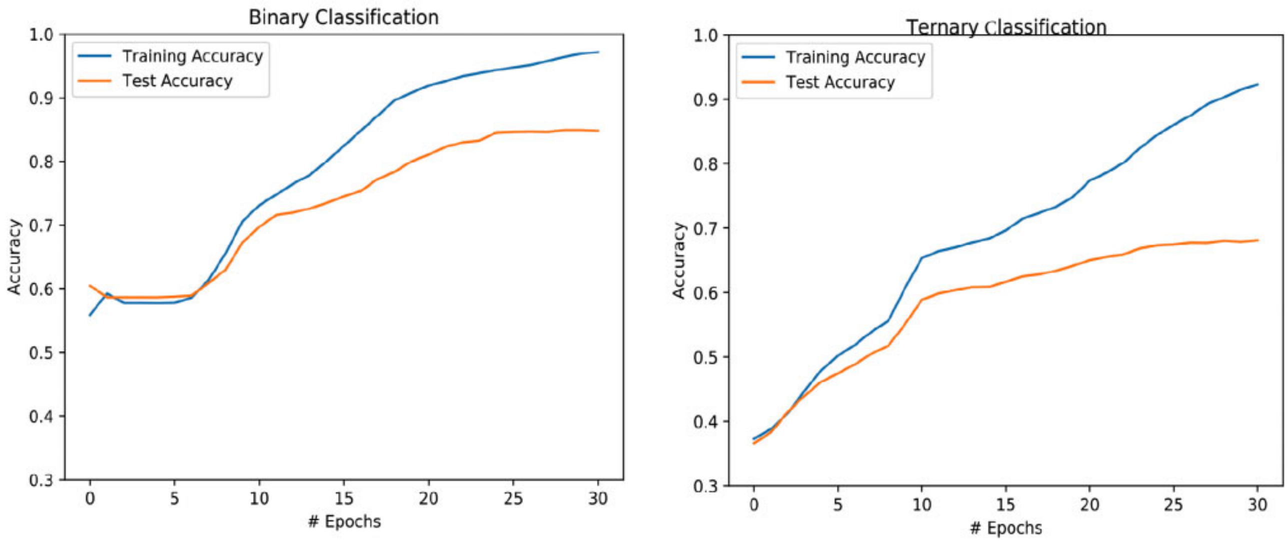

FIGURE 2 Recurrent convolutional neural network (RCNN) model training accuracy for binary and ternary classification: The *x*-axis shows the epochs and *Y*-axis represents the training accuracy

Note that the max-pooling layer takes the most important features of each word representation. The max function is an element-wise function which takes as much as possible from all the elements of a word representation *i*. The last part of the model is output layer $y^{(2)}$, a fully connected layer of the neuron, is then finally passed through softmax activation function which converts the output numbers into probabilities as shown in Equation (6).

$$p_{(i)} = \frac{\exp\left(y_k^{(2)}\right)}{\sum_{k=1}^{n} \exp^{\left(y_k^{(2)}\right)}} \tag{6}$$

One of the most appealing features of the aforementioned model is that it will retain longer contextual details and produce less noise. Hence it is considered useful for languages with low resources. Furthermore, we have fine-tuned the model parameters such as learning rate = 0.001 for Adam optimizer, word embedding vector size = 50, neurons in hidden layers = 1000, size of context vector =1000.

Figure 2 displays our RCNN model's training accuracy for the Binary and Ternary classification for 30 epochs to reflect the model's behaviour during training and validation testing. The *x*-axis shows the epochs and *Y*-axis represents the training accuracy. We observed that the training accuracy gradually increased to become stable for the rest of the epochs.

5.2.2 | Baseline methods

Rule-based approach

The rule-based approach that we aim to explore in our study utilizes a manually constructed Urdu lexicon that lists 1000 words, half positive words, and half negative words. These lexicons were generated by randomly choosing 300 sentences from the entire corpus while extracting only the positive and negative words. The sentiment of a sentence is determined by the tokens of the review and the lexicon generated. Firstly, the sentence is tokenized, and the polarity of each token is analysed by matching it with the polarity of the word in the lexicon. Secondly, the individual polarities of the tokens are established as positive or negative. Lastly, the overall sentiment of the sentence is determined by weighting negative or positive indications. Furthermore, we consider the following three rules when determining the sentiment of a sentence:

- If a sentence has a number of positive words than negative words, it is considered as a positive sentence with polarity equals to 1.
- If a sentence has more negative words than positive words, it is considered as a negative sentence with polarity 2.
- If numbers of positive and negative words are equal in a sentence, we consider it as a neutral sentence with polarity equals to 0.

The proposed pseudo-code for the Rule-based approach is shown in Algorithm 1.

Algorithm 1

```
procedure RULE-BASED USING URDU LEXICON(ARGS)
  PositiveCounter = 0
  NegativCounter = 0
  Sentiment = null
for each word in the lexicon do
  if word = negative then
    positiveCounter = positiveCounter+1
  end if
  if word = negative then
    negativeCounter = negativeCounter+1
  end if
  if word is not in Lexicon then
    word = Neutral
  end if
end for
overAllPolarity = positiveCounter − negativeCounter
if overallPolarity >0 then
  Sentiment = Positive
end if
if OverAllPolarity <0 then
  Sentiment = Negative
end if
if OverAllPolarity = 0 then
  Sentiment = Neutral
end if
end procedure
```

N-*gram model*

In 1948, Shannon first suggested *N*-grams which were subject to information theory (Silic et al., 2007). Liu (2007) defined *N*-grams as word sequence in a text with a permanent window size *N*. The *N*-grams give useful information of the corpus which can be used in different applications (Adeeba et al., 2014; Bonaccorsi et al., 2017; Hassan et al., 2018). We also implemented character-based *N*-grams in this research, where *N* ranged in length from 2 to 10.

5.3 | Evaluation measures

This section provides descriptions of evaluation methods used to evaluate the efficiency of our deployed techniques. Four performance metrics were used in this study: (a) accuracy; (b) precision; (c) recall; and (d) F1-score. The specifics for each calculation are below.

Accuracy is the ratio of correctly expected instances and the actual number of instances (see Equation (7)).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

TP is the count of true positive cases, the correctly categorized positive cases; TN is the count of true negatives, the accurately categorized negative cases; FP is the count of false positives, the wrongly graded cases that are negative; and FN is the count of false negatives, the incorrect instances that are actually good.

Precision is the fraction of correctly estimated positive instances (see Equation (8)), and recall is the fraction of correctly classified positive examples (see Equation (9)).

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

F1-score is the harmonic mean of Precision and Recall that takes False Positives (FP) and False Negatives (FN) into account. (see explanation in Equation (10))

$$F_1 - \text{score} = 2 * \frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}} \tag{10}$$

6 | RESULTS AND DISCUSSION

For the experiments, we applied a binary classification task and a ternary classification task. Furthermore, we have compared the performance of the RCNN model with two baseline approaches (a) Rule-Based Approach and (b) *N*-Gram Model. Furthermore, we also performed additional benchmark analysis against state-of-the-art machine learning models and prevailing deep learning models. The deployed RCNN based model outperforms existing state-of-the-art models on our designed data corpus.

6.1 | Results using RCNN model

This section presents the results obtained from the RCNN model with Binary and Ternary classification tasks. The primary objective is to gain insight into which classification paradigm becomes more suitable as uncertainty increase.

Table 5 shows the classification results of RCNN for both binary and ternary tasks. Overall, the highest results are achieved with a Binary classification task with approximately 85% accuracy along with 84% F1-score. These results exhibit that the number of classes has an impact on the performance of the RCNN model.

Since binary classification feeds only positive and negative instances as input. Therefore, the model can quickly learn to discriminate between binary instances. However, in the case of ternary classification as soon as the neutral instances are added, uncertainty increases that causes an abrupt drop in the accuracy. Moreover, the below-par performance of Ternary classification can be attributed to various factors. Firstly, the performance goes down by increasing the number of classes. Secondly, the number of reviews is not large—a set of 10,008 short text sentences is not considered a large dataset. Further, reviews differ in size (from 1 to 208 words), but the average number of words is 18—therefore, these reviews are short by nature. We need to look into other techniques to deal with this common problem of text shortness.

We have contrasted our best findings from the RCNN model with two simple models such as rule-based and character-*N*-grams model accompanied by a comparative study of up-to-the-minute machine learning and deep learning models.

6.2 | Results using rule-based models

Table 6 shows the results obtained from the rule-based approach. The results depict that the RCNN model performs better than the rule-based approach in terms of accuracy, precision, recall, and F1-score. The Precision is somewhat reasonable (64.30%) but Recall is very low (44.40%). The

TABLE 5 Recurrent convolutional neural network (RCNN) classification results

Model	Accuracy	Precision	Recall	F1-score
Binary classification	84.98%	84.56%	84.4%	84.48%
Ternary classification	68.56%	69.14%	67.78%	68.21%

rule-based method did not do well as only the words in the lexicon are used to define the classification. In contrast to the RCNN model, no semantic information was considered during the sentiment analysis.

6.3 | Results using character *N*-grams

Another popular approach that we have considered for the comparison is the *N*-gram model. For the task of SA, we have used character *N*-grams-where the length of *N* varies from 2 to 10. We have extracted character *N*-grams from our corpus. Table 7 shows the frequency (*N*) of some of the top *N*-grams with 2, 4, 6, 8 and 10 g. The space characters have been converted to "." for clarity. We evaluate the different values of *N* for character *N*-grams using naïve bayes classifier. The parameter settings are available in Table 8. The Table 9 displays the outcomes of the *N*-gram features evaluated with the NB algorithm. The NB classifier is a basic probabilistic classifier that determines the likelihood that a given sample belongs to a specific class. The Naïve Bayes classifier is based on the Bayes principle and operates on a conditional independence premise such that the attribute value of a given class is independent of the values of certain attributes. As can be seen that the character bigrams provide better performance. Based on the character bigrams we also provide the performance of other well-known machine learning models for sentiment analysis including random forests, DT and SVM in Table 10.

It is also observed that the rise in the size of the *N* produces inferior results relative to the RCNN model, hence the 2-Gram features work better in terms of accuracy, precision, recall and F1-score. Overall, we can assume that the RCNN's performance is superior to that of the character *N*-gram features tested in the NB algorithm. Figure 3. gives a description of the accuracy, precision, recall and f-measurement of the RCNN, the rule-based model and the *N*-gram model using the binary classification method. Typically, the RCNN outperforms both the rule-based model and the *N*-gram model with regard to both sizes.

TABLE 6 Rule-based binary classification results

Model	Accuracy	Precision	Recall	F1-score
Rule-Based Model	45.60%	64.30%	44.40%	52.50%

TABLE 7 High-frequency *N*-grams from the corpus

2-gram	Freq. (*N*)	4-gram	Freq. (*N*)	6-gram	Freq. (*N*)	8-gram	Freq. (*N*)	10-gram	Freq. (*N*)
	6145		723		457		16		13
	3398		152		143		3		3
	1522		111		46		1		2

TABLE 8 Implementations of the machine learning algorithms and their parameters (we used WEKA's implementation. Among these approaches, LibSVM is not available directly in WEKA, and we included it manually)

Approaches	Weka implementation	Parameters changed from default setting
SVM	*.functions.LibSVM	kernel: Linear
NB	.bayes.NaiveBayes	kernel: Radial Basis
DT	*.trees.J48	–
RF	*.trees.RandomForests	–

Abbreviations: DT, decision trees; NB, Naive Bayes; RF, random forest; SVM, support vector machine.

TABLE 9 *N*-gram model results for 2, 4, 6, 8, and 10 grams using Naïve Bayes

Features	Accuracy	Precision	Recall	F_1-score
2-Gram	0.622	0.598	0.513	0.552
4-Gram	0.332	0.456	0.325	0.379
6-Gram	0.293	0.346	0.292	0.316
8-Gram	0.187	0.221	0.289	0.250
10-Gram	0.163	0.152	0.236	0.184

TABLE 10 Experimental results using character bigrams

Method	Accuracy	Precision	Recall	F_1-score
NB	0.622	0.598	0.513	0.552
SVM	0.614	0.583	0.510	0.540
RF	0.589	0.584	0.496	0.533
DT	0.609	0.592	0.483	0.520

Abbreviations: DT, decision trees; NB, Naive Bayes; RF, random forest; SVM, support vector machine.

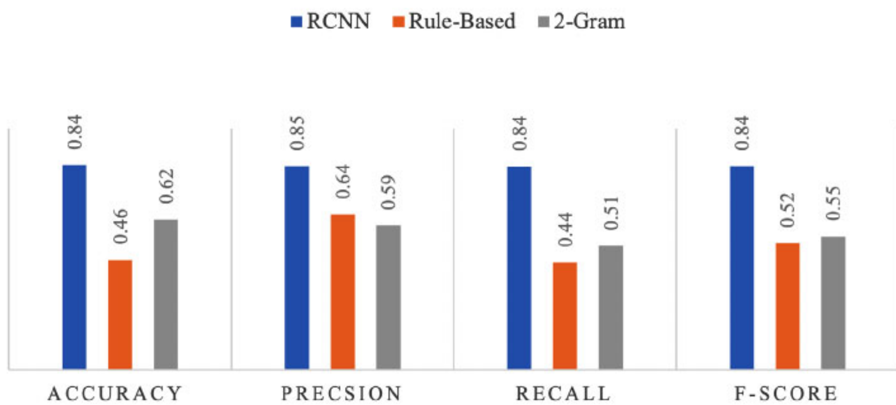

FIGURE 3 Comparison of accuracy, precision, recall and F1-score of the recurrent convolutional neural network (RCNN), the rule-based model and the 2-gram model using the binary classification method

TABLE 11 Comparative analysis of RCNN for other deep/ML models

Model	Accuracy	Precision	Recall	F1-score
RCNN	**84.98%**	**84.56%**	**84.40%**	**84.48%**
LSTM	82.35%	81.95%	82.37%	82.10%
CNN	81.75%	81.40%	81.48%	81.44%
SVM	81.64%	80.96%	80.94%	80.95%
DT	80.12%	81.29%	80.82%	81.09%
RF	80.92%	80.72%	80.55%	80.83%
NB	79.98%	81.03%	81.18%	80.91%

Note: The highest performance values for each evaluation measure are represented in bold.
Abbreviations: CNN, convolutional neural network; DT, decision trees; LSTM, long short-term memory; NB, Naive Bayes; RCNN, recurrent convolutional neural network; RF, random forest; SVM, support vector machine.

6.4 | Comparison of RCNN with ML and deep learning models

Furthermore, we have also compared RCNN results for binary class with other known machine learning and deep learning models including CNN, LSTM, and SVM. Table 11 shows the achieved results for all these classifiers. In case of CNN, we fine-tuned model parameters such as Conv1D (filters = 32, kernel size = 6), activation = ReLU with dropout = 0.5. Likewise, for LSTM, we used 128 memory units with a 0.001 learning rate. The parameter settings for SVM, DT, RF, and DT is provided in Table 8. We find that RCNN achieved the best accuracy results among all other Deep/ML models.

7 | CONCLUDING REMARKS

This work builds on the developments in Urdu-language sentiment analysis in the contemporary RCNN model. The results are promising, providing a path for more in-depth work to develop models for languages that lack enriched corpora. This study raises some opportunities in terms of both corpus development and the application of deep learning for detecting sentiments through social media platforms. Results suggest that the in-depth learning approach appears to be a good way to work with a morphologically rich language like Urdu. In addition, the lexicon entities (Nawaz et al., 2013; Wang et al., 2011) must be extended to include the entire Urdu language in order to enhance the accuracy of the sentiment analysis (Zhu et al., 2013).

In a future study, we would like to examine the role of character-level and word-level representations in the Urdu sentiment analysis. Additionally, we plan to measure the impact of using texts belonging to specific domains to conduct an unsupervised pre-training phase (Bonaccorsi et al., 2017; Hassan et al., 2016). We also plan to investigate the effect of the review sizes on the performance of the sentiment analysis task. To encourage and facilitate researchers who are interested in extending the research related to SA for Urdu, we have open-sourced the corpus and code developed for this research.[9]

ACKNOWLEDGEMENT

The authors (Salem Alelyani and Saeed-Ul Hassan) are grateful for the financial support received from King Khalid University for this research Under Grant No. R.G.P2/100/41.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in urdu_deep_sentiments at https://github.com/slab-itu/urdu_deep_sentiments, reference number 5120e1b.

ORCID

Raheem Sarwar https://orcid.org/0000-0002-0640-807X
Raheel Nawaz https://orcid.org/0000-0001-9588-0052

ENDNOTES

1 http://alt.qcri.org/semeval2020/, last accessed February 2, 2020
2 https://github.com/aesuli/Amazon-downloader, last accessed February 2, 2020
3 https://mpqa.cs.pitt.edu, last accessed on February 21, 2020
4 https://www.bbc.com/urdu, last accessed on February 15, 2020
5 https://jang.com.pk, last accessed on February 20, 2020
6 https://www.bbc.com/urdu and https://www.dawnnews.tv/, last accessed on February 2, 2020
7 https://web.archive.org/web/20161119072643/http://blog.jang.com.pk/, last accessed on February 2, 2020
8 https://github.com/slab-itu/urdu_deep_sentiments, last accessed March 19, 2020
9 https://github.com/slab-itu/urdu_deep_sentiments, last accessed March 19, 2020

REFERENCES

Abdul-Mageed, M., & Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC, vol. 515* (pp. 3907–3914). European Language Resources Association (ELRA).

Adeeba, F., Akram, Q., Khalid, H., & Hussain, S. (2014). Cle urdu books n-grams. In *Conference on language and technology*. Center for Language Engineering.

Ali, F., El-Sappagh, S., Islam, S. R., Ali, A., Attique, M., Imran, M., & Kwak, K. S. (2020). An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Generation Computer Systems*, *114*, 23–43.

Ali, F., El-Sappagh, S., & Kwak, D. (2019). Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel. *Sensors*, *19*(2), 234.

Almas, Y., & Ahmad, K. (2007). A note on extracting 'sentiments' in financial news in english, arabic & urdu. In *The second workshop on computational approaches to Arabic script-based languages* (pp. 1–12). Linguistics Society of America.

Alotaibi, B., Abbasi, R. A., Aslam, M. A., Saeedi, K., & Alahmadi, D. (2020). Startup initiative response analysis (Sira) framework for analyzing startup initiatives on twitter. *IEEE Access*, *8*, 10718–10730.

Ananiadou, S., Thompson, P., & Nawaz, R. (2013). Enhancing search: Events and their discourse context. In *International conference on intelligent text processing and computational linguistics* (pp. 318–334). Springer.

Arshad, N., Bakar, A., Soroya, S. H., Safder, I., Haider, S., Hassan, S.-U., Aljohani, N. R., Alelyani, S., & Nawaz, R. (2019). *Extracting scientific trends by mining topics from call for papers*. Library Hi Tech. https://doi.org/10.1108/LHT-02-2019-0048

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Asghar, M. Z., Kundi, F. M., Ahmad, S., Khan, A., & Khan, F. (2018). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. *Expert Systems*, *35*(1), e12233.

Asghar, M. Z., Sattar, A., Khan, A., Ali, A., Masud Kundi, F., & Ahmad, S. (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems*, *36*(3), e12397.

Attardi, G., & Sartiano, D. (2016). Unipi at semeval-2016 task 4: Convolutional neural networks for sentiment classification. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 220–224). Association for Computational Linguistics.

Aung, H. H., Zheng, H., Erdt, M., Aw, A. S., Sin, S. C. J., & Theng, Y. L. (2019). Investigating familiarity and usage of traditional metrics and altmetrics. *Journal of the Association for Information Science and Technology*, *70*(8), 872–887.

Ayata, D., Saraclar, M., & Ozg ¨ ur, A. (2017). Busem at semeval-2017 task 4a sentiment analysis with word embedding and long short term memory rnn approaches. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 777–783). Association for Computational Linguistics.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2020). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294.

Batista-Navarro, R. T., Kontonatsios, G., Mihail ˇ a, C., Thompson, P., Rak, R., Nawaz, R., Korkontzelos, I., & Ananiadou, S. (2013). Facilitating the analysis of discourse phenomena in an interoperable nlp platform. In *International conference on intelligent text processing and computational linguistics* (pp. 559–571). Springer.

Boland, K., Wira-Alam, A., and Messerschmidt, R. (2013). *Creating an annotated corpus for sentiment analysis of german product reviews*. Social Science Open Access Repository.

Bonaccorsi, A., Cicero, T., Haddawy, P., & Hassan, S.-U. (2017). Explaining the transatlantic gap in research excellence. *Scientometrics*, *110*(1), 217–241.

Bonaccorsi, A., Haddawy, P., Cicero, T., & Hassan, S.-U. (2017). The solitude of stars. An analysis of the distributed excellence model of european universities. *Journal of Informetrics*, *11*(2), 435–454.

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 69–78). Dublin City University and Association for Computational Linguistics.

Edara, D. C., Vanukuri, L. P., Sistla, V., & Kolli, V. K. K. (2019). Sentiment analysis and text categorization of cancer medical records with LSTM. *Journal of Ambient Intelligence and Humanized Computing*, 1–17. https://doi.org/10.1007/s12652-019-01399-8.

El-Beltagy, S. R., El Kalamawy, M., & Soliman, A. B. (2017). NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 790–795). Publisher: Association for Computational Linguistics.

Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)* (pp. 241–248). Coling 2008 Organizing Committee.

Hassan, M., & Shoaib, M. (2018). Opinion within opinion: Segmentation approach for urdu sentiment analysis. *International Arab Journal of Information Technology*, *15*(1), 21–28.

Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 1–8). IEEE.

Hassan, S. U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., … Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in twitter. *Knowledge-Based Systems*, *192*, 105383.

Hassan, S. U., Aljohani, N. R., Shabbir, M., Ali, U., Iqbal, S., Sarwar, R., … Herrera, F. (2020). Tweet coupling: a social media methodology for clustering scientific publications. *Scientometrics*, *124*, 973–991.

Hassan, S.-U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy. *Scientometrics*, *103*(1), 33–46.

Hassan, S.-U., Imran, M., Iftikhar, T., Safder, I., & Shabbir, M. (2017). Deep stylometry and lexical & syntactic features based author attribution on plos digital repository. In *International conference on Asian digital libraries* (pp. 119–127). Springer.

Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, *116*(2), 973–996.

Hassan, S.-U., Sarwar, R., & Muazzam, A. (2016). Tapping into intra-and international collaborations of the organization of islamic cooperation states across science and technology disciplines. *Science and Public Policy*, *43*(5), 690–701.

Hathlian, N. F. B., & Hafez, A. M. (2017). Subjective text mining for arabic social media. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *13*(2), 1–13.

Imran, M., Akhtar, A., Said, A., Safder, I., Hassan, S.-U., & Aljohani, N. R. (2018). Exploiting social networks of twitter in altmetrics big data. In *23rd international conference on science and technology indicators (STI 2018), September 12–14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. In *2017 intelligent systems conference (IntelliSys)* (pp. 722–728). IEEE.

Jang, H., Kim, M., & Shin, H. (2013). Kosac: A full-fledged korean sentiment analysis corpus. In *Proceedings of the 27th Pacific Asia conference on language, information, and computation (PACLIC 27)* (pp. 366–373). Association for Computational Linguistics.

Jarwar, M. A., Abbasi, R. A., Mushtaq, M., Maqbool, O., Aljohani, N. R., Daud, A., Alowibdi, J. S., Cano, J. R., Garc'ıa, S., & Chong, I. (2017). Communiments: A framework for detecting community based sentiments for events. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *13*(2), 87–108.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Senti-CS: Building a lexical resource for sentiment analysis using subjective feature selection and normalized Chi-Square-based feature weight generation. *Expert Systems*, *33*(5), 489–500.

Khan, M. Y., Emaduddin, S. M., & Junejo, K. N. (2017). Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach. In *2017 IEEE 11th international conference on semantic computing (ICSC)* (pp. 242–249). IEEE.

Kiritchenko, S., Mohammad, S., & Salameh, M. (2016). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)* (pp. 42–51). Association for Computational Linguistics.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence* (pp. 2267–2273). Association for Computing Machinery.

Lee, Y., Song, S., Cho, S., & Choi, J. (2019). Document representation based on probabilistic word clustering in customer-voice classification. *Pattern Analysis and Applications*, *22*(1), 221–232.

Li, W., Shao, W., Ji, S., & Cambria, E. (2020). BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *arXiv Preprint arXiv: 2006.00492*.

Limkonchotiwat, P., Phatthiyaphaibun, W., Sarwar, R., Chuangsuwanich, E., & Nutanong, S. (2020). *Domain adaptation of Thai word segmentation models using stacked ensemble*. Association for Computational Linguistics.

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Liu, Y., Du, F., Sun, J., Silva, T., Jiang, Y., & Zhu, T. (2019). Identifying social roles using heterogeneous features in online social networks. *Journal of the Association for Information Science and Technology*, *70*(7), 660–674.

Mahmood, Z., Safder, I., Nawab, R. M. A., Bukhari, F., Nawaz, R., Alfakeeh, A. S., … Hassan, S. U. (2020). Deep sentiments in Roman Urdu text using recurrent convolutional neural network model. *Information Processing & Management*, *57*(4), 102233.

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), 1–28.

Maynard, D., & Bontcheva, K. (2016). Challenges of evaluating sentiment analysis tools on social media. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 1142–1148). LREC.

Melo, P. F., Dalip, D. H., Junior, M. M., Gonçalves, M. A., & Benevenuto, F. (2019). 10SENT: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *Journal of the Association for Information Science and Technology*, *70*(3), 242–255.

Mukhtar, N., & Khan, M. A. (2018). Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, *32*(02), 1851001.

Mukhtar, N., Khan, M. A., Chiragh, N., & Nazir, S. (2018). Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. *Expert Systems*, *35*(6), e12317.

Mukund, S., Ghosh, D., & Srihari, R. K. (2011). Using sequence kernels to identify opinion entities in urdu. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 58–67). Association for Computational Linguistics.

Mukund, S., & Srihari, R. K. (2010). A vector space model for subjectivity classification in urdu aided by cotraining. In *Proceedings of the 23rd international conference on computational linguistics: Posters, pages 860–868*. Association for Computational Linguistics.

Mukund, S., & Srihari, R. K. (2012). Analyzing urdu social media for sentiments using transfer learning with controlled translations. In *Proceedings of the second workshop on language in social media, pages 1–8*. Association for Computational Linguistics.

Muneer, I., Sharjeel, M., Iqbal, M., Nawab, R. M. A., & Rayson, P. (2019). CLEU-A cross-language english-urdu corpus and benchmark for text reuse experiments. *Journal of the Association for Information Science and Technology*, *70*(7), 729–741.

Nagarajan, S. M., & Gandhi, U. D. (2019). Classifying streaming of twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications*, *31*(5), 1425–1433.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1–18). Publisher: Association for Computational Linguistics.

Nawaz, R., Thompson, P., & Ananiadou, S. (2013). Negated bio-events: Analysis and identification. *BMC Bioinformatics*, *14*(1), 14.

Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched latent Dirichlet allocation for sentiment analysis. *Expert Systems*, *37*, e12527.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *10th international workshop on semantic evaluation (SemEval 2016)*. Association for Computational Linguistics.

Rajput, Q. (2014). Ontology based semantic annotation of urdu language web documents. *Procedia Computer Science*, *35*, 662–670.

Rehman, Z. U., & Bajwa, I. S. (2016). Lexicon-based sentiment analysis for urdu language. In *2016 sixth international conference on innovative computing technology (INTECH)* (pp. 497–501). IEEE.

Riaz, K. (2007). Challenges in urdu stemming (a progress report). In *BCS IRSG Symposium: Future Directions in Information Access* (pp. 1–6). Association for Computing Machinery.

Sabah, F., Hassan, S. U., Muazzam, A., Iqbal, S., Soroya, S. H., & Sarwar, R. (2019). *Scientific collaboration networks in Pakistan and their impact on institutional research performance: A case study based on Scopus publications*. Library Hi Tech.

Safder, I., & Hassan, S.-U. (2018). Ds4a: Deep search system for algorithms from full-text scholarly big data. In *2018 IEEE international conference on data mining workshops (ICDMW), pages 1308–1315*. IEEE.

Safder, I., & Hassan, S.-U. (2019). Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, *119*(1), 257–277.

Safder, I., Hassan, S.-U., & Aljohani, N. R. (2018). Ai cognition in searching for relevant knowledge from scholarly big data, using a multi-layer perceptron and recurrent convolutional neural network model. In *Companion proceedings of the the web conference 2018* (pp. 251–258). Association for Computing Machinery.

Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information Processing & Management*, *57*(6), 102269.

Safder, I., Sarfraz, J., Hassan, S.-U., Ali, M., & Tuarob, S. (2017). Detecting target text related to algorithmic efficiency in scholarly big data using recurrent convolutional neural network model. In *International conference on Asian digital libraries* (pp. 30–40). Springer.

Sarwar, R., & Hassan, S. U. (2015). A bibliometric assessment of scientific productivity and international collaboration of the Islamic world in science and technology (S&T) areas. *Scientometrics*, *105*(2), 1059–1077.

Sarwar, R., & Nutanong, S. (2016). The key factors and their influence in authorship attribution. *Research in Computing Science*, *110*, 139–150.

Sarwar, R., Rutherford, A. T., Hassan, S. U., Rakthanmanon, T., & Nutanong, S. (2020). Native language identification of fluent and advanced non-native writers. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *19*(4), 1–19.

Sarwar, R., Urailertprasert, N., Vannaboot, N., Yu, C., Rakthanmanon, T., Chuangsuwanich, E., & Nutanong, S. (2020). $ CAG $: Stylometric authorship attribution of multi-author documents using a co-authorship graph. *IEEE Access*, *8*, 18374–18393.

Sarwar, R., Yu, C., Tungare, N., Chitavisutthivong, K., Sriratanawilai, S., Xu, Y., … Nutanong, S. (2018). An effective and scalable framework for authorship attribution query processing. *IEEE Access*, *6*, 50030–50048.

Sarwar, R., Zia, A., Nawaz, R., Fayoumi, A., Aljohani, N. R., & Hassan, S. U. (2021). Webometrics: Evolution of social media presence of universities. *Scientometrics*, *126*(2), 951–967.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making*, *18*(1), 46.

Silic, A., Chauchat, J.-H., Basic, B. D., & Morin, A. (2007). N-grams and morphological normalization in text classification: A comparison on a croatian-english parallel corpus. In *Portuguese conference on artificial intelligence* (pp. 671–682). Springer.

Śmieja, M., Tabor, J., & Spurek, P. (2019). SVM with a neutral class. *Pattern Analysis and Applications*, *22*(2), 573–582.

Sorgente, A., Flegrei, V. C., Vettigli, G., & Mele, F. (2014). *An italian corpus for aspect based sentiment analysis of movie reviews* (p. 25). CLICIT.

Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2010). Lexicon based sentiment analysis of urdu text using sentiunits. In *Mexican international conference on artificial intelligence* (pp. 32–43). Springer.

Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2014). Associating targets with sentiunits: A step forward in sentiment analysis of urdu text. *Artificial Intelligence Review*, *41*(4), 535–561.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). Association for Computational Linguistics.

Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J., & Tounsi, L. (2014). *Dcu: Aspect-based polarity classification for semeval task* (p. 4). Association for Computational Linguistics.

Wang, X., Rak, R., Restificar, A., Nobata, C., Rupp, C., Batista-Navarro, R. T. B., Nawaz, R., & Ananiadou, S. (2011). Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature. *BMC Bioinformatics*, *12*(8), S11.

Wicaksono, A. F., Vania, C., Distiawan, B., & Adriani, M. (2014). Automatically building a corpus for sentiment analysis on indonesian tweets. In *Proceedings of the 28th Pacific Asia conference on language, information and computing* (pp. 185–194). Association for Computational Linguistics.

Yu, M., Li, Z., Yu, Z., He, J., & Zhou, J. (2020). Communication related health crisis on social media: a case of COVID-19 outbreak. In *Current issues in tourism*, (pp. 1–7). Taylor & Francis.

Zafar, S., Sarwar, U., Gilani, Z., & Qadir, J. (2016). Sentiment analysis of controversial topics on Pakistan's twitter user-base. In *Proceedings of the 7th annual symposium on computing for development*, article no. 4 (pp. 1–4). Association for Computing Machinery. https://doi.org/10.1145/3001913.3006644

Zheng, H., Aung, H. H., Erdt, M., Peng, T. Q., Sesagiri Raamkumar, A., & Theng, Y. L. (2019). Social media presence of scholarly journals. *Journal of the Association for Information Science and Technology*, *70*(3), 256–270.

Zhu, D., Wang, D., Hassan, S.-U., & Haddawy, P. (2013). Small-world phenomenon of keywords network based on complex network. *Scientometrics*, *97*(2), 435–442.

AUTHOR BIOGRAPHIES

Iqra Safder is a Ph.D. student at Information Technology University. She received her BS in software engineering and MS in computer science degree from PUCIT, Lahore. Iqra is the recipient of the prestigious MDPI and ICDM student travel awards. Her research interests include text mining, applied machine learning, artificial intelligence, digital libraries, and information retrieval. She has served as a reviewer for multiple journals including TALLIP and Scientometrics.

Zainab Mahmood received the MS degree in Computer Science from Information technology university, Pakistan in 2018. Her research interests include theory and application of different natural language processing tools and machine learning.

Raheem Sarwar received the Ph.D. degree in Computer Science from City University of Hong Kong. Currently, he is a Lecturer at RGCL, RIILP, University of Wolverhampton, United Kingdom. His research interests include Natural Language Processing, Altmetrics, Scientometrics, Digital Text Forensics, and Data Science.

Saeed-Ul Hassan is the Director of AI Lab and a faculty member at Information Technology University (ITU), a former Postdoctoral Fellow at the United Nations University—with more than 15 years of hands-on experience of advanced statistical techniques, artificial intelligence, and software development client work. He earned his Ph.D. in the field Information Management from Asian Institute of Technology—AIT. He has also served as a Research Fellow at National Institute of Informatics in Japan. Dr. Saeed is also the recipient of the James A. Linen III Memorial Award in recognition of his outstanding academic performance. More recently, he has been awarded Eugene Garfield Honorable Mention Award for Innovation in Citation Analysis by Clarivate Analytics, Thomson Reuters.

Farooq Zaman is PhD Fellow at the Department of Computer Science, Information Technology University, Lahore, Pakistan. His research interests include Text Summarization, Deep Learning and Machine Translation.

Rao Muhammad Adeel Nawab is working as Assistant Professor in the Computer Science Department of COMSATS Institute of Information Technology, Lahore, Pakistan. He is also leading the Natural Language Processing (NLP) research group at CIIT, Lahore. He holds a PhD in Computer Science from University of Sheffield, UK. His research interests include Plagiarism Detection, Information Retrieval and Natural Language Processing.

Faisal Bukhari received an M.Sc. in Statistics (2000) and an M.Sc. in Computer Science (2006) from the University of Punjab (PU), Lahore, Pakistan. He completed a Masters in Computer Science (2009) and a Ph.D. in Computer Science (2012) from the Asian Institute of Technology (AIT), Thailand. Currently, he is an Associate Professor and supervising the Imaging and Data Science Lab at PUCIT. His research interests lie in computer vision, image processing, data science, machine learning, virtual reality, and health informatics.

Rabeeh Abbasi is working as a tenured associate professor at the Department of Computer Science, Quaid-i-Azam University, Islamabad. He has a vast research experience in the fields of social media analytics and social network analysis. His research focuses on leveraging positive aspects of social media. He has published more than 50 articles in reputed journals and conferences.

Salem Alelyani has been an assistant professor in the Computer Science Department at King Khalid University, Saudi Arabia, since 2014. He serves as a consultant to the president of the university and the director of the Center for Artificial Intelligence. He obtained his PhD from Arizona State University 2013 in Machine Learning and Data Mining. He has several publications in the field. He serves as a reviewer in multiple international conferences and scientific journals including ICTAI, AAAI, ICMLA, ICML, IRI, IEEE Access, Artificial Intelligence Review, AJSE, Information Sciences and others. Also, he serves as a board member and as a PC member in other international journals and conferences.

Naif Aljohani is the Director of Web Observatory Lab and Associate Professor at the Faculty of Computing and Information Technology in King AbdulAziz University, Jeddah, Saudi Arabia. He holds a PhD in Computer Science from the University of Southampton, UK. He received the Bachelor's degree in Computer Education from King Abdul Aziz University, 2005. In 2009, he received the Master degree in Computer Networks from La Trobe University, Australia. His research interests are in the areas of Learning and Knowledge Analytic, Semantic Web, Web Science, Big Data Analytics, Bibliometrics, Scientometrics and Evidence-based Policy Formulation.

Raheel Nawaz has served in various senior leadership positions in the private Higher and Further Education sector; and was an Army Officer before that. He is currently the Director of Digital Technology Solutions and a Reader in Analytics and Digital Education with Manchester Metropolitan University (MMU). He has founded and/or headed several research units specializing in artificial intelligence, digital transformations, data science, digital education, and apprenticeships in higher education. He has led on numerous funded research projects in the United Kingdom, EU, South Asia, and Middle East. He holds adjunct or honorary positions with several research, higher education, and policy organizations, both in the United Kingdom and overseas.

How to cite this article: Safder, I., Mehmood, Z., Sarwar, R., Hassan, S.-U., Zaman, F., Nawab, R. M. A., Bukhari, F., Abbasi, R. A., Alelyani, S., Aljohani, N. R., & Nawaz, R. (2021). Sentiment analysis for Urdu online reviews using deep learning models. *Expert Systems*, e12751. https://doi.org/10.1111/exsy.12751