Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model

Zainab Mahmood (Data Curation)[a], Iqra Safder[a], Rao Muhammad Adeel Nawab[b], Faisal Bukhari[c], Raheel Nawaz[d], Ahmed S. Alfakeeh[a], Naif Radi Aljohani[e], Saeed-Ul Hassan[a,*]

[a] *Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan*
[b] *Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan*
[c] *Punjab University College of Information Technology (PUCIT), University of the Punjab (PU), Lahore, Pakistan*
[d] *Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Manchester, United Kingdom*
[e] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*

ARTICLE INFO

Keywords:
Roman Urdu corpus
Urdu sentiment analysis
Under resource language
Deep learning
Text Classification
Recurrent Convolutional Neural Networks (RCNN)

ABSTRACT

Although over 64 million people worldwide speak Urdu language and are well aware of its Roman script, limited research and efforts have been made to carry out sentiment analysis and build language resources for the Roman Urdu language. This article proposes a deep learning model to mine the emotions and attitudes of people expressed in Roman Urdu - consisting of 10,021 sentences from 566 online threads belonging to the following genres: Sports; Software; Food & Recipes; Drama; and Politics. The objectives of this research are twofold: (1) to develop a human-annotated benchmark corpus for the under-resourced Roman Urdu language for the sentiment analysis; and (2) to evaluate sentiment analysis techniques using the Rule-based, N-gram, and Recurrent Convolutional Neural Network (RCNN) models. Using Corpus, annotated by three experts to be positive, negative, and neutral with 0.557 Cohen's Kappa score, we run two sets of tests, i.e., binary classification (positive and negative) and tertiary classification (positive, negative and neutral). Finally, the results of the RCNN model are analyzed by comparing it with the outcome of the Rule-based and N-gram models. We show that the RCNN model outperforms baseline models in terms of accuracy of 0.652 for binary classification and 0.572 for tertiary classification.

1. Introduction

In recent years, the internet has become the most important platform for social interactions between different people (Hassan, Akram & Haddawy, 2017a; Liu et al., 2019; Lytos, Lagkas, Sarigiannidis & Bontcheva, 2019). As technology expands rapidly, more individuals use the Internet for online shopping, distance learning, telemedicine, and correspondence on various aspects of life (Vuong, Saastamoinen, Jacucci & Ruotsalo, 2019). The growing use of social networking sites such as Blogs, Forums, Twitter, or Facebook involves people to express their opinions and engage in discussion groups (Al-Smadi, Al-Ayyoub, Jararweh & Qawasmeh,

* Corresponding author.
E-mail addresses: mscs1605@itu.edu.pk (Z. Mahmood), iqra.safder@itu.edu.pk (I. Safder), adeelnawab@ciitlahore.edu.pk (R.M.A. Nawab), faisal.bukhari@pucit.edu.pk (F. Bukhari), r.nawaz@mmu.ac.uk (R. Nawaz), asalfakeeh@kau.edu.sa (A.S. Alfakeeh), nraljohani@kau.edu.sa (N.R. Aljohani), saeed-ul-hassan@itu.edu.pk (S.-U. Hassan).

https://doi.org/10.1016/j.ipm.2020.102233
Received 10 August 2019; Received in revised form 20 February 2020; Accepted 27 February 2020

2019; Fatima, Hasan, Anwar & Nawab, 2017; Thompson et al., 2013). The anonymity of the web has allowed an individual to engage in hostile social media speech content that has made text discourse (Hassan et al., 2017b; Hassan, Safder, Akram & Kamiran, 2018) or, more specifically, sentiment analysis an important task of understanding the behavior of individuals (Batista-Navarro et al., 2013; Nawaz, Thompson & Ananiadou, 2013, 2010; Shardlow et al., 2018; Thompson, Nawaz, McNaught & Ananiadou, 2017; Hassan et al., 2017).

The great significance of sentiment analysis can be seen from our need to understand both the attitude of others towards the problem and what they think (Sailunaz & Alhajj 2019). Companies and organizations are seeking some useful data from these views, such as the sentiments behind customer opinions (Ananiadou, Thompson & Nawaz, 2013; Anwaar, Iltaf, Afzal & Nawaz, 2018; Imran et al., 2018; Nawaz, Thompson & Ananiadou, 2012; Qadir, Khalid, Khan, Khan & Nawaz, 2018). Sentiment analysis relates to the analysis of emotions behind user-written texts using natural language processing, text analytics, computational linguistics, and machine learning techniques (Ayyaz et al., 2018; Xing, Pallucchini & Cambria, 2019; Zhang et al., 2019). Due to growing interest in sentiment analysis, companies are interested to run campaigns, winning marketing strategies, have more customers, overcome weaknesses and lead scoring. Companies are keen to understand what individuals tell about their products or services and their emotions behind their words (Luo, Huang & Zhu, 2019). In addition, politicians are interested in knowing their reputations and what media sources say about them. In recent years, the focus of sentiment analysis has shifted to the analysis of the emotions of social media texts. Many topics, such as politics, sports, medicine, disasters, and harassment, have increased the use of sentiment analysis (Araque, Zhu & Iglesias, 2019). Sentiment analysis includes advanced methods for natural language processing, data mining for predictive analysis and contextual understanding of documents become an exciting area of research (Jahangir, Afzal, Ahmed, Khurshid & Nawaz, 2017; Safder & Hassan 2019; Wang et al., 2011; Zhang, Wang & Liu, 2018).

English and other European languages are regarded as affluent languages when it comes to the accessibility of the tools needed to carry out sentiment analysis, but there are many other languages that are deemed resource-poor (Al-Ayyoub, Khamaiseh, Jararweh & Al-Kabi, 2019), and Roman Urdu is one of them. Unfortunately, Roman Urdu requires standard corpora to perform natural language processing tasks. Urdu is Pakistan's national language spoken in five other areas of South Asia. Due to its morphological difficulty, the Urdu script is not very prevalent. Roman Urdu is a term used for the Urdu language written in Roman script i.e., with English alphabets. In Pakistan, India and other South Asian nations, users mostly interact on social media platforms using the Roman Urdu script or *"Romanagari"* which is a term used for Hindi language written in Roman script. Urdu and Hindi are the same languages with a distinction in their writing script. However, both languages are very comparable and can be interpreted by individuals from both nations if they are presented in the Roman Urdu script. For instance, the Roman Urdu script for "*Your work is awesome*" goes as "ap ka kaam lajawab hey".

Sentiment analysis for the Roman Urdu language is equally important as for other languages as it helps non-Urdu speaking users to understand the opinion of others and the emotion behind a text. Urdu is the national language of Pakistan and is widely used across Asia. When it comes to social media, People of Urdu as native language mostly use its Roman script to express their opinions and thoughts. Its use can be seen on Twitter, Facebook and other websites nowadays. So sentiment analysis has become important even for the Roman Urdu language so that people can understand the opinions of people expressed in this language. The challenge of sentiment analysis for Roman Urdu has not yet been explored fully even after its immense use, therefore, in this paper, the primary focus is Roman Urdu.

In this research, we aim to contribute a significant benchmark manually annotated Roman Urdu corpus for the task of sentiment analysis, hereafter referred to as the RUSA-19 Corpus. To build the suggested Corpus, we gathered 10,021 sentences from five different genres: sport, food and recipes, politics, software, and drama. First, human annotators manually annotated each evaluation and fell into one of three classifications: (1) positive; (2) negative; and (3) neutral. Second, we constructed a state-of-the-art model that draws on the advancements of deep learning to show the efficiency of our RUSA Corpus in sentiment analysis tasks.[1] Finally, we provide a Python code for our Roman Urdu sentiment analysis framework. We believe that the RUSA-19 Corpus and model library will be useful in 1 fostering research in Roman Urdu, an under-resourced language; (2) allowing an immediate comparison of current state-of-the-art sentiment analysis techniques in Roman Urdu language; and (3) developing and evaluating new techniques in Roman Urdu sentiment analysis.

The rest of the paper is organized into six sections. In the second section, we focus on the previous related research on sentiment analysis. Section 3 discusses the entire corpus generation process, including the steps involved, annotation process, and corpus characteristics. Section 4 focuses on the methodology. Section 5 discusses the experimental results obtained from our sentiment analysis of Roman Urdu text classification models. Finally, Section 6 concludes the paper.

2. Literature review

In this section, we first discuss a brief overview of SemEval corpora and notable methods for sentiment analysis. We will then introduce a survey of the Roman Urdu corpora and related methods of sentiment analysis.

2.1. Brief review on SemEval corpora and sentiment analysis

The most prominent attempts taken in the literature to develop benchmark corpora for sentiment analysis are the series of

[1] The data and code used in this paper can be accessed from the following URL to reproduce the results: https://github.com/slab-itu/rusa_19

SemEval competitions. In each contest, scientists perform distinct tasks using various corpora to assess semantic analysis systems. The result of such contests is a collection of benchmark corpora and distinct methods for sentiment analysis. These corpora have been created in the English and Arabic dialects (Kiritchenko, Mohammad & Salameh, 2016). Generally, reviews/tweets fall under a number of classifications, such as laptops, restaurants, hotel reviews or tweets/SMS messages.

Each year, the sizes of the SemEval corpora are different. The 2013 version used Twitter and SMS datasets, and the Twitter dataset was split into 9728 training data, 1654 development data, and 3813 testing data, while the SMS data set of 2093 sentences was used only for testing. The 2014 edition of the Twitter test included 1853 tweets containing 86 sarcasm tweets and 1142 Live Journal data (Rosenthal, Nakov, Ritter & Stoyanov, 2016). There were five separate subtasks in 2016 and 2017 and, for each, the dataset was divided into training, development, development testing and testing. For subtask A, a dataset of 30,632 sentences was used, for subtask B and D a dataset of 17,639 sentences was used and for subtask C and E a dataset of 30,632 sentences was used (Nakov, Ritter, Rosenthal, Sebastiani & Stoyanov, 2016). Subtask A training and testing data for SemEval 2016 were used in 2017 (Ayata, Saraclar & Ozgur, 2017). Apart from SemEval Competitions, research on sentiment analysis has also been conducted for different languages such as Korean, German, Indonesian and Italian.

KOSAC, A Korean Corpus, contains 332 news articles chosen from the Sejong Corpus to evaluate feelings in Korean using an annotation system recognized as the Korean Subjectivity Markup Language Jang, Kim & Shin, 2013). A corpus for German consumer feedback was developed by extracting Amazon item reports using the Amazon review parser1. Each sentence in the corpus has been annotated according to its specific meaning. A sum of 63,067 sentences has been obtained from various product domains (Boland, Wira-Alam & Messerschmidt, 2013). The Indonesian tweets corpus of 5.3 million tweets was generated using the Twitter Streaming API. Tweets geolocation was used to filter tweets in the Indonesian dialect (Wicaksono, Vania, Distiawan & Adriani, 2014). The Italian Corpus, composed of 2648 sentences relating to the genre of film, was produced for aspect-based sentiment analysis. Each sentence in the corpus was manually annotated on the basis of various aspects. Five sentiment categories were described to depict sentiments towards each aspect: ((1) strongly negative; (2) negative; (3) neutral; (4) positive; and (5) strongly positive (Sorgente, Flegrei, Vettigli & Mele, 2014).

Various techniques have been suggested in the literature for the purpose of sentiment analysis. For example, in SemEval 2014, both unsupervised (Rule-based) and supervised (Support Vector Machine) machine learning techniques are used. The constructed Rule-based method was used to determine the sentiment polarities of sentences using a lexicon to generate a feature vector for supervised machine learning techniques. The sentence polarity is measured by adding the polarity scores of all words in the sentence and dividing by their distance from the aspect term. If the polarity score of the aspect term is higher than 0, it is classified as positive; if the score is less than 0, it is classified as negative; otherwise, it is classified as neutral. These features and N-gram features obtained from this Rule-based strategy have been used to train a four-way SVM (Wagner et al., 2014). In the SemEval 2016 version, Random Forest, Gaussian Regression and Linear Regression are used. To show the power of the term associated with the positive sentiment in a sentence, a score between 0 and 1 is used using these supervised methods. These results were assessed using Kendall's rank correlation coefficient and Spearman's rank regression (Kiritchenko et al., 2016).

In the SemEval 2017 publication, two classification methods are described: machine learning-based using word embeddings for feature representation and Long Short Term Memory Recurrent Neural Networks (LSTM) centered on a method that uses word indexes as input chains for feature representation. Note that term embeddings are an advanced NLP system that links words or phrases to a real number vector depiction. Machine learning models such as Support Vector Machines will discover a hyperplane that separates tweets/sentences according to their classes. In the same way, Random Forest produces various decision trees and ultimate choices are taken by evaluating the individual trees. Naive Bayes is also a probabilistic classifier relying on Bayes Theorem, and the LSTM is capable of studying long-term dependencies (Ayata et al., 2017).

Turney collaborated on a semantic analysis of the genre of a film using an unsupervised method called Thumbs up or Thumbs down. This technique focuses on the adjectives/adverbs in the sentences and then the polarity of the words that refer to the sentence classification (Turney, 2002). Another approach was to use artificial neural networks to split the text into positive, negative and fuzzy tones (Jian, Chen & Wang, 2010). In addition, In another research, a neural network approach is used for accumulating the benefits and applications of machine learning and information retrieval techniques. It utilizes semantic direction indices as inputs for neural networks to determine sentiments (Chen, Liu & Chiu, 2011). Dos Santos and Gatti suggested a deep Convolutional Neural Network that utilizes character-to sentence-level information to conduct sentiment analysis in SemEval assignments (Attardi & Sartiano 2016). In the 2017 task, the Ensemble technique of three classifiers has been used, Convolutional Neural Network, the Multilayer Perceptron and Logistic regression Model for Topic-Based message polarity and tweet quantification (El-Beltagy, Kalamawy & Soliman, 2017). Recently, ArWordVec a customized word embeddings model is designed for tweets classification that claims to be the largest study in the Arabic language (Fouad, Mahany, Aljohani, Abbasi & Hassan, 2019).

Studies have been conducted to perform sentiment analysis of different under resource languages such as Hindi, Thai, Khmer and Arabic. Research on Hindi language reviews for sentiment analysis based on negation and discourse relation has been carried out. An annotated Hindi reviews dataset was created which was classified using the polarity-based method. An accuracy of 80.21% was achieved using the algorithm (Mittal, Agarwal, Chouhan, Bania & Pareek, 2013). Some research has also been done in the Thai language also considered as an under-resourced language (Tuarob & Mitrpanont 2017). A machine learning algorithm is proposed to detect the abusive language in the Thai language. The algorithm yields 86% f-measure. Another research is done in the Bengali language (Al-Amin, Islam & Uzzal, 2017). In this work, the sentiment analysis of Bengali comments are carried out using Word2Vec and sentiment detection of words. They have presented an important insight into Bengali sentiment analysis with the Word2Vec method. An accuracy of 75.5% is obtained from the methodology.

According to the literature review conducted on all the under-resourced languages, we have observed that none of the researchers have implemented a neural networks approach in the classification task of these languages. Therefore, It is important to study this approach test on under resource language as we have done in our research work.

2.2. Brief review on Roman Urdu corpora and sentiment analysis

Multiple efforts have been made by different researchers to develop corpora, techniques, and other language resources to tackle the issue of Roman Urdu sentiment analysis. The difficulty of sentiment analysis in the Roman Urdu dialect, however, has not been fully studied.

The most significant condition for performing natural language tasks in Roman Urdu is the lexical normalization of the words employed by Sharf & Rehman. The method began by gathering reviews/tweets of Roman Urdu from various locations, including Twitter, Urdu biography, IT Dunya, Reddit, Names4muslims, Pakish news, and shashca.com. They have created an algorithm centered on phonetic algorithms such as Soundex and NYSIIS to standardize Roman Urdu text (Sharf & Rehman 2017). More lately, a significant commitment is produced in performing natural language processing on Roman Urdu datasets. The dataset comprises 15,000 statements gathered from separate sources, namely Twitter, Reddit, Urdu Poetry and Social Workers Biographies. The primary job is to conduct discourse-based sentiment analysis on Roman Urdu, for which the data were initially gathered, then lexically normalized into standard expression, and then lastly recognized the data for absence or presence of a discourse element. On average, the accuracy was 80%. This assignment has been carried out in the path of implementing a neural network method to sentiment analysis in the future (Sharf & Rehman 2018).

Another research has been done on the sentiment analysis of the bilingual twitter dataset (English and Roman Urdu) on the subject of general elections. A dataset of 89,000 tweets has been gathered and the suggested technique consists of three phases. (i) Classification of topic and language (ii) semi-automatic building of lexicons using current Senti-strength and Wordnet. (iii) Tweets are evaluated and the degree of sentiments is calculated for each tweet using a bilingual sentiment lexicon (Javed, Afzal, Majeed & Khan, 2014). In another study, a linguistic reservoir was suggested, consisting of sentiment scores of Roman Urdu text by conducting a spatial analysis of bilingual (Urdu and English) tweets under the theme of General Elections 2013. The dataset of the tweets was gathered from four major political parties in Pakistan. Tweets are categorized using two methods. The first discriminates against tweets relating to political and non-political groups. The second strategy discriminates against tweets in the English and Roman Urdu languages. For Roman Urdu, a lexicon was built that gives a sentimental score to words comparable to the SentiStrength framework. A total of 91,804 tweets, along with 21,821 non-political tweets, were collected. Additionally, 7186 tweets belong to the Roman Urdu language in the language classification. Furthermore, in order to improve the coverage of Roman Urdu lexicon, a bigram based cosine similarity measure. was used (Javed & Afzal 2013).

Later, An approach towards the construction of bilingual (English, Roman Urdu) small text datasets relating to the political sphere was presented (Javed & Afzal 2014). Javed & Afzal performed tweet classification on the basis of the topic of interest and language. A total of 21,762 tweets were classified as Roman Urdu after 7.6% noise removal. In another research work related to sentiment classification of Roman Urdu and English opinions, three classification models were used: Naive Bayes, Decision Tree, and KNN. The opinions were extracted from blogs on a dataset of 150 positive and 150 negative opinions. According to the outcomes, Naive Bayes outperformed KNN and the decision tree in aspects of accuracy, precision, recall, and F-measurement.

An unsupervised feature-based strategy was proposed in research to find lexical differences in Roman Urdu text using two datasets (Rafae et al., 2015). They scrapped their first dataset from the websites on media, poetry, SMS and blog sites, while the second SMS dataset was obtained from Chopaal, an Internet-based SMS service. A Roman Urdu opinion mining scheme was suggested in another work to perform multiple experiments on the collected data, from a mobile website (whatmobile.com.pk). The scheme comprises five primary steps: data collection, translation of Roman Urdu reviews into English reviews, identification of opinion polarities and giving a rating in graphical format. A maximum of 1620 assessments was evaluated for experimental purposes. The findings obtained showed only a 21.1% deviation from the initial outcomes (Daud, Khan & Daud, 2015). In another work related to the field, A database comprising 40,000 words for classification of dialogue acts was suggested (Shaikh, Strzalkowski & Webb, 2011). The corpus was gathered by inviting subjects to online chat sessions. Experiments have been intended around topics that can readily be seen in certain types of behavior.

Another effort has been made on sentiment analysis of Roman Urdu text using supervised methods depending on feature selection (Arif et al., 2016). First, to retrieve the information, an openly available English Hotel review dataset was downloaded and transformed to the Urdu script using the Google Translate API. Later on, the Urdu script data was transformed into Roman Urdu using an online tool. Finally, the textual data is transformed into a numerical format for feature selection. The dataset used comprises 800 documents each containing both positive and negative assessments. Multiple classifiers were employed, such as SVM, KNN, Decision Tree Classifier, Nearest Centroid, Perceptron, Passive Aggressive Classifier, Stochastic Gradient Classifier, Ridge Classifier, and Naive Bayes. Results have shown that the Ridge classifier, the multinomial Naive Bayes and the SVC with linear kernel and linear SVC have the best performance at the accuracy of 96% with TF-IDF vectorizer.

Generally, there is no standard lexicon in Roman Urdu and there are many possible spelling variations exist for a given word. For instance, the word “hakumat” (Government) is also written as “Khakumat”, “hokumat”, and “hakamat”. More specifically, following standardization problems occur such as: (i) words with different spelling variations (see above example), (ii) words with similar spellings that are lexically different (e.g. word “Bahar” means “outside” in Urdu and “spring season” in English) and, (iii) words with similar spellings in English and Urdu (e.g. the word ‘had’ in English has same spellings for word ‘had’ (limit) in Urdu language) (Rafae et al., 2015). Such irregularities pose a problem of data sparseness in the processing of basic NLP tasks such as Urdu word segmentation (Khan, Khan & Khan, 2018), POS tagging (Sajjad & Schmid, 2009), spell checking and, machine translation

((Durrani, Sajjad, Fraser & Schmid, 2010; Naseem & Hussain, 2007), etc.

Note that Roman Urdu is an under-resourced language, technically and linguistically. According to the literature review, many of the techniques relevant to sentiment analysis of other languages are inapplicable to the Roman Urdu language due to the spelling variations of this language. The same word of Roman Urdu can be written with different spellings even by the same person (Masroor, Saeed, Feroz, Ahsan & Islam, 2019; Rafae et al., 2015). Moreover, the lack of language and linguistic resources such as datasets and lexicons and one to one mapping between Urdu letters and its corresponding Roman letter also makes it harder to implement the existing sentiment analysis techniques as mentioned in the literature review, such as the accessibility of corpora and lexicons.

In addition, available annotated corpora are not big enough to carry out efficient sentiment analysis. Moreover, most of the corpora and sentences of limited genres belonging to positive and negative classes only. To overcome this limitation, this paper focuses on constructing a Roman Urdu corpus comprising sentences belonging to five different genres. We focus on the classification of sentences into positive, negative and neutral classes. To perform sentiment analysis, we have implemented a deep learning model RCNN on our constructed corpus RUSA-19 which has not yet explored fully for the sentiment analysis of Roman Urdu data.

3. Corpus generation

This section introduces the measures adopted to build an annotated Roman Urdu corpus to perform sentiment analysis. The steps taken in the construction of our RUSA-19 Corpus include the collection of raw data from different websites, the preparation of annotation guidelines, the execution of manual annotation, the storage of corpus, standardization and finally, the presentation of the characteristics of the corpus.

3.1. Data collection from online resources

To construct a benchmark corpus for the evaluation of the Roman Urdu sentiment, we accumulate data from the websites that provide unlimited access and enable consumers to post in Roman Urdu. Table 1 highlights all the websites that we have searched for the data collection. Roman Urdu is an under-resourced language; therefore, internet repositories render information relating to different genres available and easily accessible to the generation of a challenging benchmark for the Roman Urdu corpus. We gathered information from five different genres, including drama, movies and, talk shows; food and recipes; politics; sport; and software, blogs, forums and, gadgets. The data was collected manually by two individuals who were well aware of the objective within the 8 to 9 months' time frame. Initially, each review was stored in an excel spreadsheet along with the following information: (i) review ID; (ii) topic to which the review belongs; (iii) URL from where the review was collected; (iv) date of collection; and (v) annotation tag. The corpus was finally standardized in XML format after some data cleaning such as removal of illegal characters and URLs from the comments.

3.2. Annotation process and guidelines

This section defines the full annotation method that we attended to manually annotate the full corpus. This stage involves the preparing of rules for annotation, the manual annotation of the entire corpus by human annotators, and the computation of an inter-annotator agreement. In order to evaluate the performance of our corpus and add value to our corpus, three annotators which were native speakers of Urdu language and well familiar with the Roman Urdu script and the purpose annotated the entire corpus. We prepared guidelines for sentiment analysis from distinct existing corpora prior to corpus annotation. Table 2 demonstrates some of the examples of reviews belonging to the positive, negative, and neutral class.

3.2.1. Positive class guidelines

i A sentence is marked as positive if the given sentence expresses a positive sentiment for all the aspect term (Pontiki et al., 2016).

ii In the case of a sentence expressing both positive and neutral sentiment, the positive sentiment dominates and the sentence is

Table. 1
Roman Urdu data collection sources.

Genre	Sources	No of reviews
Food and recipes	www.kfoods.com, www.urduweb.org, www.facebook.com, www.friendskorner.com, www.paksitan.web.pk,	1504 reviews
Sports	www.urduweb.org, www.tafrehmella.com	2001 reviews
Politics	www.siasat.pk, www.twitter.com	2014 reviews
Dramas, movies and talk shows	www.hamariweb.com, www.zemtv.com, www.urduweb.org, www.dailydose.pk,www.siasat.pk, www.reviewit.pk, www.tweettunnel.com, www.dramasonline.com, www.fashionuniverse.net	2001 reviews
Software, blogs, fora, and gadgets	www.mobilesmspk.net, www.itforumpk.com,www.baazauq.blogspot.com, www.dufferistan.com, www.mbilalm.com, www.urduweb.org, www.urdudaan.blogspot.com, www.itdunya.com, www.achidosti.com, www.itdarasgah.com, www.tafrehmella.com, www.sachiidosti.com, www.urdupoint.com	2501 reviews

Table. 2
Examples of reviews belonging to positive, negative and neutral class.

Positive review examples (with English translation)	Negative review examples (with English translation)	Neutral review examples (with English translation)
Ap ka kaam lajawab.*(Your work is awesome)*	Mujhe software pasand nai aya.*(I don't like the software)*	Beshak.*(Indeed)*
Us ka camera ki resolution aala ha.*(His-camera resolution is awesome)*	Bohat bura kaam hai apka *(Your work is very bad)*	Software acha ha per mujhe samaj nai araha.*(The software is good, but I can't comprehend it.)*
Bohat achi recipe ha.*(It is a very good recipe)*	Fazool.*(Bad)*	Har jeet to khel ka hisa hoti hai.*(Winning and losing is part of the game.)*

classified as positive.

iii The presence of an agreement of approval makes the sentence positive. (Abdul-Mageed & Diab 2012).
iv Sentences with illocutionary speech act like praise, congratulations are classified as positive (Abdul-Mageed & Diab 2012).

3.2.2. Negative class guidelines

i If the aspect term in a sentence expresses a negative sentiment then the sentence is classified as negative (Maynard & Bontcheva 2016).
ii A sentence is classified as negative if it has more negative terms then other sentiments.
iii Direct un-softened disagreements in a sentence make it a negative sentence (Abdul-Mageed & Diab 2012).
iv Banning, bidding, penalizing, and assessing terms in a sentence makes sentences negative (Abdul-Mageed & Diab 2012).
v If a sentence has a negative word with a positive adjective then it is considered as a negative sentence as it negates the positive adjective in the sentence. (Ganapathibhotla & Liu 2008).
vi If negation is present in a sentence then it is considered as a negative sentence.

3.2.3. Neutral class guidelines

i Factual information in a sentence makes it a neutral sentence (Boland et al., 2013).
ii If thought is shared in a sentence then it is classified as neutral (Sorgente et al., 2014).
iii Sentences with a reduced degree of surety and liability such as words like “maybe (shayad)” are considered neutral sentences (Abdul-Mageed & Diab 2012).
iv A sentence with both positive and negative sentiment in terms of the aspects and entities are classified as a neutral sentence (Pontiki et al., 2016).

3.3. Corpus characteristics

The full RUSA-19 Corpus was manually annotated by three annotators (A,B, and C) in order to prepare the benchmark corpora. All the sentences were annotated by graduates and native speakers of Urdu also familiar with the Roman Urdu language and the task of sentiment analysis. In order to test the accuracy of our annotation guidelines, we sampled 100 reviews from all domains and sent them to two annotators (A and B) to provide a manual label for the reviews. For each evaluation, they introduced one of the following three categories: neutral, positive, and negative. The conflicting pairs between the two annotators have been discussed and the Annotation Guidelines have been updated as outlined above. The updated guidelines were used by two annotators to label the corpus as a whole and, in the event of a disagreement, the third annotator gave the label to the review. We have accomplished an Inter-Annotator Agreement (IAA) of 70.72% and a Cohens Kappa score of 0.557 (Moderate) for the full corpus. Moderate and acceptable scores show the reality that the Annotation Guidelines have been well prepared, well understood, and followed by annotators throughout the annotation phase. Our conflict analysis reveals that the bulk of disputes happened when the courses differed between positive and neutral (12.07%) and negative and neutral (11.57%) classes. RUSA-19 Corpus consists of 10,021 Roman Urdu reviews,

Table. 3
Statistics of the Corpus.

Positive reviews	3778
Negative reviews	2941
Neutral reviews	3302
Num. of tokens	146,558
Num. of types	21,750
Review min. length	1 word
Review maximum length	154 words
reviews avg. length	15 words
Num. of reviews	10,021

of which 38% are positive reviews; 29% negative reviews; and 33% neutral reviews, as shown in Table 3. It can be obviously seen from these statistics that there is a class balance in our corpus. Researchers have made efforts to develop corpora for carrying out experiments but unfortunately, most of the currently tagged corpora are not big enough and covers the content from just a few genres rather than various genres such as the UCL Roman Urdu dataset.[2] The UCL dataset is vast enough to carry out sentiment analysis task but the sentences in this dataset are not categorized into different genres as our corpus which comprises data belonging to five different genres. Most of the corpora are restricted to a small number of domains and concentrated on just two classes, positive and negative.

4. Data and methods

In this section, we addressed the experimental details of our Deep learning model, Rule-based and N-gram-based methods. These models have been implemented on our suggested RUSA-19 Corpus and for comparison, we run same experiments on the UCL Roman Urdu dataset. In addition, we performed two types of experiments for the evaluation, namely: binary classification and tertiary classification.

4.1. Experimental datasets

The RUSA-19 Corpus comprises 10,021 Roman Urdu sentences which belong to five genres: (i) drama, movies, and talk shows; (ii) food and recipes; (iii) politics; (iv) software, blogs, forums, and gadgets; and (v) sport collected from various social media websites. Each sentence in our corpus falls into one of three classes: positive (denoted by 1); negative (denoted by 2); and neutral (denoted by 0). We have conducted two experiments: binary classification and tertiary classification. To carry out the binary classification experiment, at first positive and negative sentences are extracted from the whole corpus and then the sentences are segmented. For the tertiary classification experiment, all the sentences belonging to positive, negative and neutral classes are used. For the comparison purpose, we have used the UCL Roman Urdu dataset which comprises 20,228 Roman Urdu sentences belonging to three classes: positive; negative; and neutral. The entire dataset consists of 6013 positive reviews; 5286 negative reviews; and 8929 neutral reviews. Table 4 provides the details of the datasets used in both experiments.

4.2. Classification approaches

This section presents the details of our designed baseline and deep learning method. The baseline method comprises a Rule-based and an N-gram technique. Furthermore, we constructed a deep learning RCNN network to build a customized framework for the Roman Urdu sentiment analysis. Fig. 1 shows high-level architecture from data inputs to classification outputs.

4.2.1. Rule-based approach

Lexicon comprising 500 positive words and 500 negative words. The lexicons were generated by selecting 300 random sentences from the RUSA-19 Corpus and the Roman Urdu UCL dataset. Fig. 2 illustrates the pseudo-code of this technique in detail. At first, we tokenized a sentence and classified each token as positive, negative or neutral by matching it to the polarity in the lexicon. The sentiment of the sentence is determined by the tokens and the lexicons.

The polarity of a sentence is determined by weighing the polarity of each individual token in a sentence. The following rules are considered to classify the sentence into one of three classes: positive, negative or neutral.

i A sentence is considered as positive with polarity score 1 if it has more positive words.
ii A sentence is considered as negative with a polarity score of 2 if it has more negative words.
iii A sentence is considered as neutral with a polarity score of 0 if it has equal positive and negative words.

4.2.2. N-gram model

N-grams were first mentioned by Shannon in 1948 subjected to the theory of communication (Šilić, Chauchat, Bašić & Morin, 2007). N-grams are the sequence of words in a text with a fixed window size N which is used to find useful information in a corpus (Liu, 2007). In research, a word-based N-grams was used on a 37 million word corpus belonging to poetry and prose domain Adeeba, Akram, Khalid and Hussain, (2014). In this paper, we have implemented the character-based N-gram model where the length of N varies from 2 to 10.

4.2.3. Recurrent Convolutional Neural Network

The Recurrent Convolutional Neural Network model (Lai, Xu, Liu & Zhao, 2015; Safder, Hassan & Aljohani, 2018) was basically introduced to overcome some of the limitations of older traditional neural network models such as Recurrent Neural Network (RNN) (Safder & Hassan 2018) and Convolutional Neural Network (CNN) model. The RNN model analyses the text word by word and stores the contextual information of a sentence in a hidden layer. However, its disadvantage is that it only favors the recent words of a sentence which can affect the overall semantics of the text. To overcome this disadvantage, CNN was introduced with a max-pooling

[2] http://archive.ics.uci.edu/ml/datasets/Roman + Urdu + Data + Set

Table. 4
Dataset for experiments.

Corpus	Dataset	Classes	Train Set	Test Set
RUSA-19 Corpus	Binary classifier	2-positive and negative	5380 reviews	1339 reviews
	Tertiary classifier	3-positive, negative and neutral	8000 reviews	2021 reviews
UCL Roman Urdu dataset	Binary classifier	2-positive and negative	9039 reviews	2260 reviews
	Tertiary classifier	3-positive, negative and neutral	16,182 reviews	4046 reviews

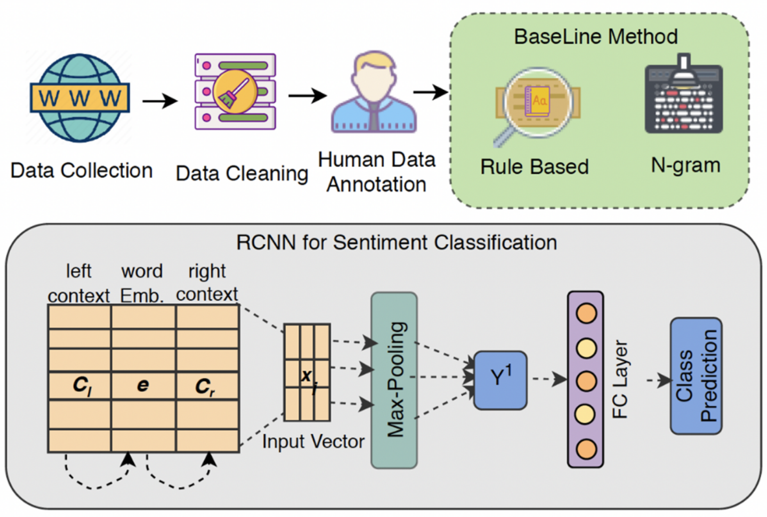
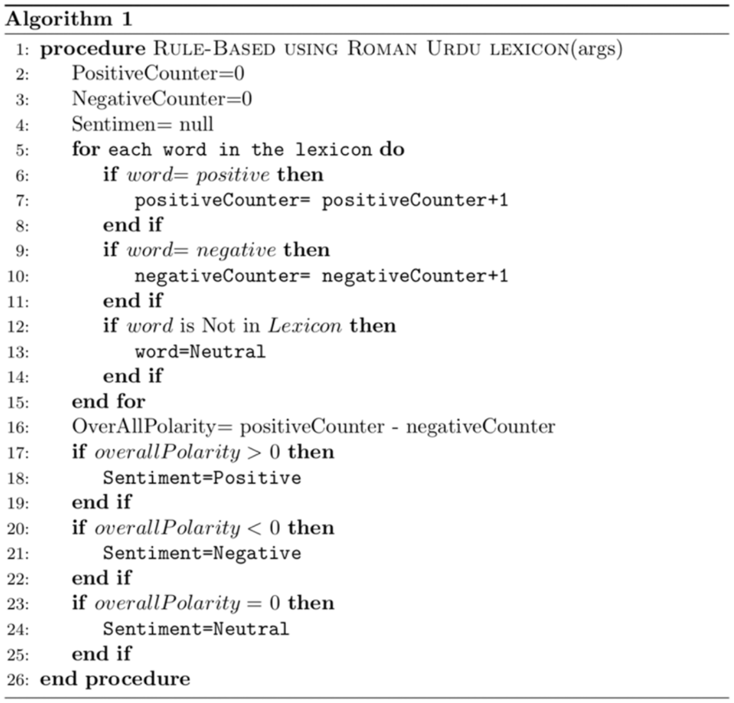


Fig. 1. High-level system architecture of the Roman Urdu sentiment classification framework.



Algorithm 1

```
procedure RULE-BASED USING ROMAN URDU LEXICON(args)
    PositiveCounter=0
    NegativeCounter=0
    Sentimen= null
    for each word in the lexicon do
        if word= positive 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
```

Fig. 2. Pseudo-code for a Rule-based technique using the Roman Urdu lexicon.

layer which helps in selecting the most important and useful words in a text. The major limitation of this CNN is that it uses a fixed kernel-like fixed window size, which makes the learning more challenging and time-consuming.

In an RCNN model, the main idea is to create a word representation that consists of (1) the left context, obtained from forward RNN; (2) word embedding; and (3) the right context, obtained from backward RNN. Each word representation is created by applying a bi-directional RNN and a max-pooling layer.

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

$$c_r\ (w_i) = f((W^{(r)})c_r(w_{i-1}) + (W^{(sr)})e(w_{i-1})) \tag{2}$$

Eqs. (1) and 2 are used to calculate the left and right context of the word W_i. In the above Equations, $c_l(w_i)$ and $c_r(w_i)$ are the left and right contexts of the word w_i. $e(w_{i-1})$ is the word embedding of word (w_{i-1}). $W^{(l)}$ matrix transforms the hidden layer to the next hidden layer. $W^{(sl)}$ matrix combines the semantics of the current word with the left context of the next word. f is a non-linear activation function.

Eq. (3) shows the word representation by concatenating $C_l(W_i)$, the left context vector, $e(W_i)$, the word embedding and $C_r(W_i)$, the right context vector.

$$x_i = [C_l(W_i):\ e(W_i);\ C_r(W_i)] \tag{3}$$

Each word representation x_i (see Eq. (3)) is passed through a standard layer where a linear transformation along with the *tanh* function is applied to it and results in y which contains a semantic vector that is used to find the most useful semantic in the text. As the next step, a max-pooling layer is applied as shown in Eq. (4). The max-pooling layer is used for the feature extraction of each word representation.

$$y_i^{(1)} = \overset{n}{\max}[\tanh(W^{(1)}x_i + \ b^{(1)})] \tag{4}$$

In Eq. (4), the max function takes the maximum from all the elements of a word representation x_i.

Finally, the output layer is computed using Eq. (5).

$$y_i^{(2)} = (W^{(2)}\ y_i^{(1)} + \ b^{(2)}) \tag{5}$$

$y_i^{(2)}$ is passed through a softmax function as shown in Eq. (6) which converts the output into probability by which we can classify the text into the class with the highest probability.

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

The training target of the network attempts to maximize the log-likelihood of a given class. The weights of the network were initialized from a uniform distribution.

We used standard parameters to set up the model for the task of sentiment analysis. We tuned the learning rate of the stochastic gradient descent as 0.05, word embedding as 50, the hidden layer size as 1000, the size of context vector is 1000 and no. of epochs as 100. For training the model to detect the sentiment behind a text, first, we initialized all network parameters as shown in Eq. (7). The parameters are real-valued word embedding E, real-valued bias vectors $b^{(2)}$ and $b^{(4)}$ and initial left and right context vectors $c_l(w_1)$ and $c_r(w_n)$.

$$\theta = \{E,\ b^{(2)},\ b^{(4)},\ c_l(w_1),\ c_r(w_n),\ W^{(2)}, W^{(4)},\ W^{(l)}, W^{(r)},\ W^{(sl)}, W^{(sr)}\} \tag{7}$$

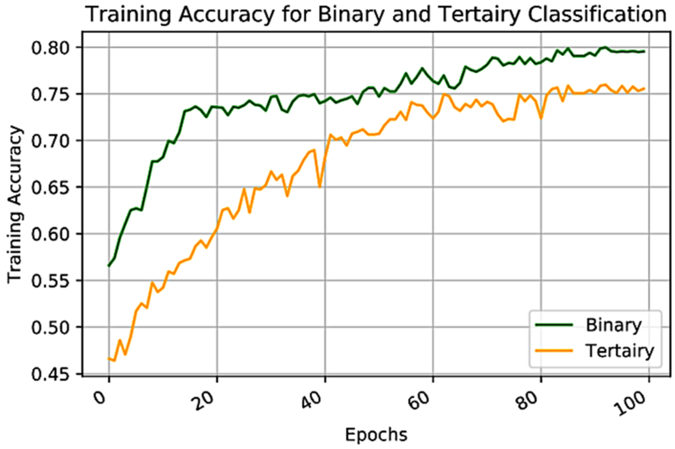


Fig. 3. Training accuracy of the RCNN model for binary and tertiary classification.

Fig. 3 shows the training accuracy curve of 100 epochs for binary and tertiary classification experiments to analyze the learning of the RCNN model. The epochs are represented on the x-axis and training accuracy is represented on the y-axis. It is observed from the training curves that at first the model training gradually increases and then becomes stable after some epochs. The RCNN model is considered to be an efficient model for low-resource languages such as Roman Urdu due to the fact that it can preserve longer contextual information and introduce less noise.

4.3. Evaluation measures

We performed two sets of experiments (binary classification & tertiary classification) to evaluate the suggested RCNN model. As a baseline model for comparison and better insights, we implemented a Rule-based approach with 6719 sentences for the binary classification and 10,021 sentences for the tertiary classification from our corpus. For the evaluation using a Rule-based approach, the sentences are tokenized and each token is then equated with the lexicon entry for polarity detection. We have also applied a character N-gram model which is evaluated using the Naive Bayes algorithm. Further, we depict four performance metrics that we have used for evaluation of our models namely: accuracy, precision, recall, and F-measure.

Accuracy is defined as the ratio of correctly classified subjects to the total number of subjects as shown in Equation 8.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN} + \text{FP+FN}} \tag{8}$$

where TP is the number of true positives TN is the number of true negatives, the correctly classified negative instances; FP is the number of false positives and FN is the number of false negatives. Precision is the ratio of correctly classified positive subjects to the total number of predicted positive subjects. Recall is the ratio of positive subjects that are correctly classified as shown in Eq. (9) and 10.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

F-Measure is defined as the harmonic mean of precision and recall. (see Eq. (11)).

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}} \tag{11}$$

5. Results and analysis

This section describes the results of experiments carried out in this research along with the demonstration of the effectiveness of our proposed model and the discussions of these outcomes in light of the aim of this research. In the comparison of the three models which we have implemented, it is observed that the RCNN model outperforms the Rule-based model and N-gram model.

5.1. Binary and tertiary classification by RCNN model

In this section, we have described the results obtained from binary and tertiary classification experiments using the RCNN model compared with the baseline models. Table 5 shows the experimental results obtained for binary classification and tertiary classification using the RCNN model on our RUSA-19 Corpus and Roman Urdu UCL dataset. It is observed that RCNN performance is better on RUSA-19 Corpus as compared to the UCL dataset on both experiments with an accuracy of 0.751 using binary classification and 0.713 using tertiary classification. The obtained result clearly indicated that binary classification outperformed Tertiary classification task with 0.652 accuracy, 0.653 precision, 0.503 recall, and 0.568 F1-score on our corpus. The UCL Roman Urdu dataset performs better in tertiary classification in terms of Precision, Recall and F-measure. It is observed that RCNN can perform better even for this dataset along with our corpus. We realized that the poor performance of Tertiary classification is due to multiple reasons, such as more number of classes in the deployed dataset and a larger number of reviews with longer lengths of sentences. Additionally, we can conclude that the number of classes affects the performance of the RCNN model. Furthermore, we have compared our best results obtained from the RCNN model for binary classification of both datasets with two baseline models: rule-based, and character n-grams model.

Table. 5
RCNN evaluation results.

Dataset	Model	Accuracy	Precision	Recall	F1 Score
RUSA-19 Corpus	Binary classification	0.751	0.738	0.744	0.741
	Tertiary classification	0.713	0.710	0.683	0.696
Roman Urdu UCL dataset	Binary classification	0.738	0.721	0.725	0.723
	Tertiary classification	0.693	0.732	0.699	0.715

Table. 6
Rule-based model achieved results.

Dataset	Model	Accuracy	Precision	Recall	F1 Score
RUSA-19 Corpus	Binary classification	0.544	0.721	0.541	0.618
	Tertiary classification	0.497	0.623	0.440	0.515
Roman Urdu UCL dataset	Binary classification	0.530	0.729	0.569	0.639
	Tertiary classification	0.486	0.513	0.554	0.532

5.2. Rule-based model

According to the results shown in Table 6 for the Rule-based approach of binary and tertiary classification using RUSA-19 Corpus and UCL dataset, it is observed that RUSA-19 Corpus performs better than the UCL dataset in terms of accuracy which is 0.544 for binary classification and 0.497 for tertiary classification. Moreover, we can say that the RCNN model outperforms the rule-based approach in terms of all evaluation measures we have used. The Rule-based approach didn't perform well in this experiment is merely due to the fact that no semantic information was considered while making the analysis; the classification is just based on the words in the lexicons.

5.3. N-gram model

We have applied character N-grams with a length of N varying from 2–10 characters. Table 7 and 8 show the results of N-gram models for binary classification and tertiary classification analyzed using a machine learning model called Naive Bayes algorithm. Naive Bayes classifier is based on Bayes theorem which predicts the probabilities of a given sample belonging to a particular.

The character-based N-gram model resulted in the poor outcome as compared to RCNN model for both datasets as we increase the size of N. For binary classification, the 2-gram features give the best results (accuracy = 0.743) for RUSA-19 Corpus and (accuracy = 0.755) for UCL dataset whereas, for tertiary classification, Roman Urdu UCL gives better result in terms of accuracy (0.531). Fig. 4 shows the comparison of accuracy, precision, recall, and F-measure obtained from RCNN, Rule-based model, and N-gram model using binary classification for RUSA-19 Corpus (see 5(a)) and UCL Roman Urdu dataset (see 5(a)). Fig. 5 shows the comparison of accuracy, precision, recall, and F-measure obtained from RCNN, Rule-based model, and N-gram model using tertiary classification for RUSA-19 Corpus (see 6(a)) and UCL Roman Urdu dataset(see 6(b)). It is found that the RCNN model outperforms the Rule-based model and the N-gram model with respect to all the evaluation measures we have used.

6. Concluding remarks

In recent times, quite a few studies have been reported in the direction of Roman Urdu Sentiment analysis. We find that high classification accuracy has been achieved on relatively small using an array of lexical approaches that require extensive data preprocessing such as TF-IDF, LDA, and linguistics-based models. However, these techniques involve comprehensive and laborious feature engineering to extract distinctive features for the task of Roman Urdu sentiment analysis. In contrast, our proposed deep learning-based model can automatically perform the feature engineering on relatively large data corpus effectively with an encouraging accuracy.

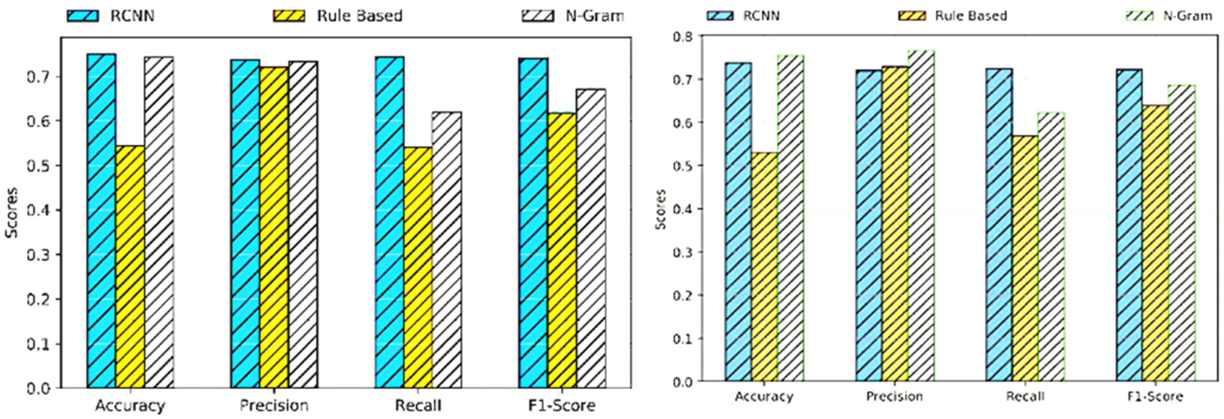
This study opens up a window for further research using deep learning in order to build language-independent models for low-resource languages. Also, our results highlight an essential insight that deep learning is a promising approach to handle complex languages like Roman Urdu. The future work will focus on the enhancements of the research for improved results. We believe that our publicly available data corpus would serve as a benchmark to carry out sentiment analysis for the Roman Urdu language. The data and code used in this study can be accessed from the following URL to replicate the results for future research: URL: https://github.com/slab-itu/rusa_19.

Table. 7
N-gram model results for binary classification.

Features	Dataset	Accuracy	Precision	Recall	F1 Score
2-gram	RUSA-19 Corpus	0.743	0.733	0.619	0.671
	Roman Urdu UCL dataset	0.755	0.765	0.623	0.686
4-gram	RUSA-19 Corpus	0.711	0.723	0.543	0.620
	Roman Urdu UCL dataset	0.698	0.734	0.571	0.642
6-gram	RUSA-19 Corpus	0.618	0.619	0.432	0.508
	Roman Urdu UCL dataset	0.603	0.598	0.411	0.487
8-gram	RUSA-19 Corpus	0.423	0.643	0.321	0.428
	Roman Urdu UCL dataset	0.459	0.547	0.309	0.394
10-gram	RUSA-19 Corpus	0.320	0.312	0.238	0.270
	Roman Urdu UCL dataset	0.279	0.288	0.256	0.271

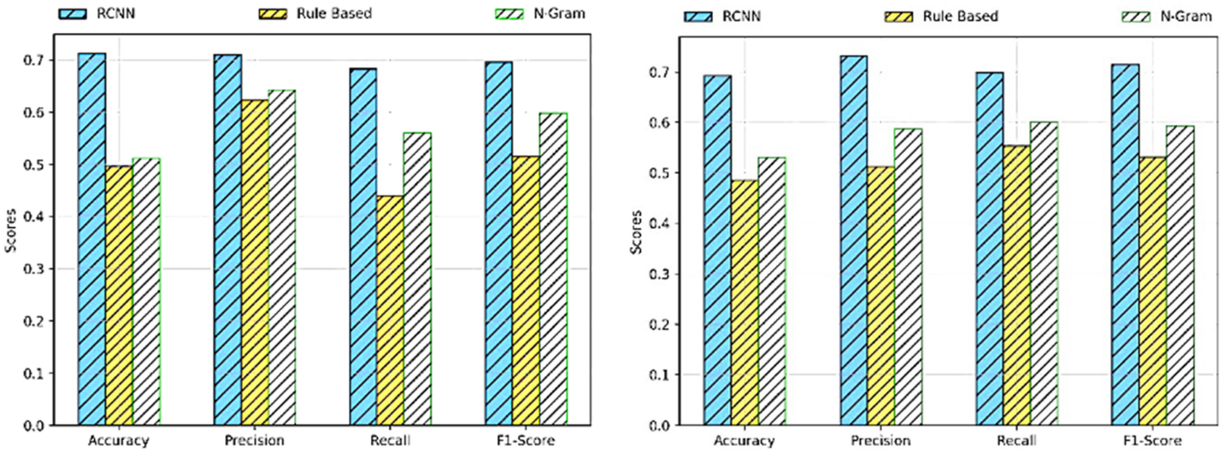
Table. 8
N-gram model results for tertiary classification.

Features	Dataset	Accuracy	Precision	Recall	F1 Score
2-gram	RUSA-19 Corpus	0.512	0.642	0.561	0.598
	Roman Urdu UCL dataset	0.531	0.587	0.601	0.593
4-gram	RUSA-19 Corpus	0.487	0.398	0.455	0.424
	Roman Urdu UCL dataset	0.500	0.499	0.510	0.504
6-gram	RUSA-19 Corpus	0.455	0.543	0.409	0.466
	Roman Urdu UCL dataset	0.401	0.451	0.499	0.473
8-gram	RUSA-19 Corpus	0.339	0.421	0.390	0.404
	Roman Urdu UCL dataset	0.320	0.352	0.305	0.326
10-gram	RUSA-19 Corpus	0.311	0.298	0.309	0.303
	Roman Urdu UCL	0.299	0.302	0.286	0.293



(a) (b)

Fig. 4. Comparison of Precision, Recall, and F1 score of binary classification for all techniques using RUSA-19 Corpus and UCL Roman Urdu dataset.



(a) (b)

Fig. 5. Comparison of Precision, Recall, and F1 score of tertiary classification for all techniques using RUSA-19 Corpus and Roman Urdu UCL dataset.

CRediT authorship contribution statement

Zainab Mahmood: Investigation, Writing - review & editing. **Iqra Safder:** Writing - review & editing. **Rao Muhammad Adeel Nawab:** Validation, Investigation, Writing - review & editing. **Faisal Bukhari:** Investigation, Writing - review & editing. **Raheel Nawaz:** Methodology, Writing - original draft. **Ahmed S. Alfakeeh:** Validation, Writing - review & editing. **Naif Radi Aljohani:**

Supervision, Writing - original draft, Writing - review & editing. **Saeed-Ul Hassan:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2020.102233.

References

Abdul-Mageed, M., & Diab, M. T. (2012). AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. *LREC. Vol. 515. LREC* (pp. 3907–3914).

Adeeba, F., Akram, Q., Khalid, H., & Hussain, S. (2014). *CLE Urdu books N-grams.* Conference on Language and Technology.

Al-Amin, M., Islam, M. S., & Uzzal, S. D. (2017). Sentiment analysis of bengali comments with word2vec and sentiment information of words. *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 186–190). IEEE.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of Arabic sentiment analysis. *Information Processing & Management, 56*(2), 320–342.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management, 56*(2), 308–319.

Ananiadou, S., Thompson, P., & Nawaz, R. (2013). Enhancing search: Events and their discourse context. *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 318–334). Springer.

Anwaar, F., Iltaf, N., Afzal, H., & Nawaz, R. (2018). HRS-CE: A hybrid framework to integrate content embeddings in recommender systems for cold start items. *Journal of computational science, 29*, 9–18.

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems, 165*, 346–359.

Attardi, G., & Sartiano, D. (2016). UniPI at SemEval-2016 task 4: Convolutional Neural Networks for sentiment classification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 220–224). .

Ayata, D., Saraclar, M., & Ozgur, A. (2017). BUSEM at SemEval-2017 task 4A sentiment analysis with word embedding and long short term memory RNN approaches. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 777–783). .

Ayyaz, S., Qamar, U., & Nawaz, R. (2018). HCF-CRS: A hybrid content based fuzzy conformal recommender system for providing recommendations with confidence. *PloS one, 13*(10), https://doi.org/10.1371/journal.pone.0204849.

Batista-Navarro, R. T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., et al. (2013). *Facilitating the analysis of discourse phenomena in an interoperable NLP platform. International Conference on Intelligent Text Processing and Computational Linguistics*. Springer559–571.

Boland, K., Wira-Alam, A., & Messerschmidt, R. (2013). Creating an annotated corpus for sentiment analysis of german product reviews. *GESIS-Technical Reports*.

Chen, L. S., Liu, C. H., & Chiu, H. J. (2011). A neural network based approach for sentiment classification in the blogosphere. *Journal of Informetrics, 5*(2), 313–322.

Daud, M., Khan, R., & Daud, A. (2015). Roman Urdu opinion mining system (RUOMiS). *Computer Science & Engineering: An International Journal (CSEIJ), 4*(5/6).

Durrani, N., Sajjad, H., Fraser, A., & Schmid, H. (2010). Hindi-to-Urdu machine translation through transliteration. *Proceedings of the 48th Annual meeting of the Association for Computational Linguistics* (pp. 465–474). Association for Computational Linguistics.

El-Beltagy, S. R., Kalamawy, M. E., & Soliman, A. B. (2017). Niletmrg at semeval-2017 task 4: Arabic sentiment analysis. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017* (pp. 790–795). .

Fatima, M., Hasan, K., Anwar, S., & Nawab, R. M. A. (2017). Multilingual author profiling on Facebook. *Information Processing & Management, 53*(4), 886–904.

Fouad, M. M., Mahany, A., Aljohani, N., Abbasi, A., & Hassan, S. U. (2019). ArWordVec: Efficient word embedding models for Arabic tweets. *Soft Computing* (pp. 1–8). .

Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. *Proceedings of the 22nd International Conference on Computational Linguistics. 1. Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 241–248). Association for Computational Linguistics.

Hassan, S. U., Akram, A., & Haddawy, P. (Akram and Haddawy, 2017a). Identifying important citations using contextual information from full text. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–8). IEEE.

Hassan, S. U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017b). Measuring social media activity of scientific literature: An exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics, 113*(2), 1037–1057.

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. *International conference on Asian digital libraries* (pp. 119–127).

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.

Imran, M., Akhtar, A., Said, A., Safder, I., Hassan, S. U., & Aljohani, N. R. (2018). Exploiting social networks of Twitter in altmetrics big data. *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018* 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. *2017 Intelligent Systems Conference (IntelliSys)* (pp. 722–728). IEEE.

Jang, H., Kim, M., & Shin, H. (2013). KOSAC: A full-fledged Korean sentiment analysis corpus. *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 366–373). .

Javed, I., & Afzal, H. (2013). *Opinion analysis of bi-lingual event data from social networks.* IA: ESSEM@ AI*164–172.

Javed, I., & Afzal, H. (2014). Creation of bi-lingual social network dataset using classifiers. *In International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 523–533). Springer.

Javed, I., Afzal, H., Majeed, A., & Khan, B. (2014). Towards the creation of linguistic resources for bilingual sentiment analysis of twitter data. *International Conference on Applications of Natural Language to Data Bases/Information Systems* (pp. 232–236). Springer.

Jian, Z. H. U., Chen, X. U., & Wang, H. S. (2010). Sentiment classification using the theory of ANNs. *The Journal of China Universities of Posts and Telecommunications, 17*, 58–62.

Khan, S. N., Khan, K., & Khan, W. (2018). Supervised Urdu word segmentation model based on POS information. *EAI Endorsed Trans. Scalable Information Systems, 5*(19), e2.

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

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Twenty-ninth AAAI conference on artificial intelligence*.

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data.* New York: Springer Berlin Heidelberg.

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*.

Luo, Z., Huang, S., & Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Information Processing & Management, 56*(3), 408–423.

Lytos, A., Lagkas, T., Sarigiannidis, P., & Bontcheva, K. (2019). The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management, 56*(6), 102055.

Masroor, H., Saeed, M., Feroz, M., Ahsan, K., & Islam, K. (2019). Transtech: Development of a novel translator for Roman Urdu to English. *Heliyon, 5*(5), e01780.

Maynard, D., & Bontcheva, K. (2016). Challenges of evaluating sentiment analysis tools on social media. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1142–1148). .

Mittal, N., Agarwal, B., Chouhan, G., Bania, N., & Pareek, P. (2013). Sentiment analysis of hindi reviews based on negation and discourse relation. *Proceedings of the 11th Workshop on Asian Language Resources* (pp. 45–50). .

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis on Twitter. *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 1–18). .

Naseem, T., & Hussain, S. (2007). A novel approach for ranking spelling error corrections for Urdu. *Language Resources and Evaluation, 41*(2), 117–128.

Nawaz, R., Thompson, P., & Ananiadou, S. (2012). Identification of manner in bio-events. *LREC* (pp. 3505–3510). .

Nawaz, R., Thompson, P., & Ananiadou, S. (2013). Negated bio-events: Analysis and identification. *BMC bioinformatics, 14*, 14. https://doi.org/10.1186/1471-2105-14-14.

Nawaz, R., Thompson, P., McNaught, J., & Ananiadou, S. (2010). Meta-Knowledge annotation of bio-events. *LREC* (pp. 2498–2507). .

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., & Mohammad, A. S. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 19–30). .

Qadir, H., Khalid, O., Khan, M. U., Khan, A. U. R., & Nawaz, R. (2018). An optimal ride sharing recommendation framework for carpooling services. *IEEE access : practical innovations, open solutions, 6*, 62296–62313.

Rafae, A., Qayyum, A., Moeenuddin, M., Karim, A., Sajjad, H., & Kamiran, F. (2015). An unsupervised method for discovering lexical variations in Roman Urdu informal text. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 823–828). .

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2016). SemEval-2014 task 9: Sentiment analysis on Twitter. *Proceedings of the international workshop on semantic evaluation (semeval-2014)* (pp. 502–518). .

Safder, I., & Hassan, S. U. (2018). DS4A: Deep search system for algorithms from full-text scholarly big data. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 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. *Companion Proceedings of The Web Conference 2018* (pp. 251–258). .

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*.

Sajjad, H., & Schmid, H. (2009). Tagging Urdu text with parts of speech: A tagger comparison. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 692–700). .

Shaikh, S., Strzalkowski, T., & Webb, N. (2011). Classification of dialogue acts in Urdu multi-party discourse. *KDIR,* 406–412.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from the scientific literature. *BMC medical informatics and decision making, 18*(1), 46.

Sharf, Z., & Rahman, S. U. (2017). Lexical normalization of roman Urdu text. *International Journal of Computer Science and Network Security, 17*(12), 213–221.

Sharf, Z., & Rahman, S. U. (2018). Performing natural language processing on Roman Urdu datasets. *International Journal Of Computer Science And Network Security, 18*(1), 141–148.

Šilić, A., Chauchat, J. H., Bašić, B. D., & Morin, A. (2007). N-grams and morphological normalization in text classification: A comparison on a croatian-english parallel corpus. *Portuguese Conference on Artificial Intelligence* (pp. 671–682). Berlin, Heidelberg: Springer.

Sorgente, A., Flegrei, V. C., Vettigli, G., & Mele, F. (2014). An Italian Corpus for aspect-based sentiment analysis of movie reviews. *CLICIT2014, 25*.

Thompson, P., Nawaz, R., Korkontzelos, I., Black, W., McNaught, J., & Ananiadou, S. (2013). News search using discourse analytics. *2013 Digital Heritage International Congress (DigitalHeritage). 1. 2013 Digital Heritage International Congress (DigitalHeritage)* (pp. 597–604). IEEE.

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2017). Enriching news events with meta-knowledge information. *Language Resources and Evaluation, 51*(2), 409–438.

Tuarob, S., & Mitrpanont, J. L. (2017). *Automatic discovery of abusive Thai language usages in social networks.* NovemberInternational Conference on Asian Digital Libraries267–278.

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

Vuong, T., Saastamoinen, M., Jacucci, G., & Ruotsalo, T. (2019). Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology*.

Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J. et al. (2014). Dcu: Aspect-based polarity classification for semeval task 4.

Wang, X., Rak, R., Restificar, A., Nobata, C., Rupp, C. J., Batista-Navarro, R. T. B., et al. (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. *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computing* (pp. 185–194). .

Xing, F. Z., Pallucchini, F., & Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management, 56*(3), 554–564.

Zhang, B., Xu, X., Li, X., Chen, X., Ye, Y., & Wang, Z. (2019). Sentiment analysis through critic learning for optimizing Convolutional Neural Networks with rules. *Neurocomputing, 356*, 21–30.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.