symmetry

MDPI

Article

Innovations in Urdu Sentiment Analysis Using Machine and Deep Learning Techniques for Two-Class Classification of Symmetric Datasets

Khalid Bin Muhammad * and S. M. Aqil Burney

Department of Computer Science, College of Computer Science and Information Systems, Institute of Business Management Karachi, Karachi 75270, Pakistan
* Correspondence: kbmcbm@iobm.edu.pk

Abstract: Many investigations have performed sentiment analysis to gauge public opinions in various languages, including English, French, Chinese, and others. The most spoken language in South Asia is Urdu. However, less work has been carried out on Urdu, as Roman Urdu is also used in social media (Urdu written in English alphabets); therefore, it is easy to use it in English language processing software. Lots of data in Urdu, as well as in Roman Urdu, are posted on social media sites such as Instagram, Twitter, Facebook, etc. This research focused on the collection of pure Urdu Language data and the preprocessing of the data, applying feature extraction, and innovative methods to perform sentiment analysis. After reviewing previous efforts, machine learning and deep learning algorithms were applied to the data. The obtained results were compared, and hybrid methods were also recommended in this research, enabling new avenues to conduct Urdu language data sentiment analysis.

Keywords: sentiment analysis (S.A); urdu text preprocessing; two-class classification; deep learning; opinion mining; feature extraction; hybrid algorithms

Citation: Muhammad, K.B.; Burney, S.M.A. Innovations in Urdu Sentiment Analysis Using Machine and Deep Learning Techniques for Two-Class Classification of Symmetric Datasets. *Symmetry* **2023**, *15*, 1027. https://doi.org/10.3390/sym15051027

Academic Editor: Jian-Qiang Wang

Received: 27 February 2023
Revised: 10 April 2023
Accepted: 18 April 2023
Published: 5 May 2023

1. Introduction

According to statistics derived from the internet, the top 10 most popular languages used on the Internet according to percentage share are English (25.9%), Chinese (19.4%), Spanish (7.9%), Arabic (5.2%), Portuguese (3.7%), Indonesian/Malaysian (4.3%), French (3.3%), Japanese (2.6%), Russian (2.5%), and German (2.0%), totaling 76.9%, while the rest of the languages constitute 23.1%. English's share is only 25.9%, while non-English language users comprise 74.1% out of all users. Major research has been carried out on English language processing, but a lot of work is yet to be carried out on other languages, especially South Asian languages [1]. Non-English languages include Chinese, which is used by 19.4% of people worldwide. Work has been carried out on Chinese natural language processing (CNLP), in which assessments have been performed and resources have been built [2]. Text processing and speech processing are still underway.

In this paper, a detailed study was performed to investigate Urdu text data processing, the classification of data, and the application of machine learning as well as deep learning algorithms on Urdu text to perform Urdu sentiment analysis and compare the obtained results. The key contribution of this study is the use of a hybrid technique to analyze and process Urdu language text; hence, it suggests a new method for further improvement of results.

1.1. Sentiment Analysis/Attitude Scrutiny

Sentiment analysis determines the feelings of people about a certain product or service. In social networking, people write microblogs in Urdu or other languages because they express their feelings/reviews/expressions according to various classes. These posts may

be posted on microblogs in real-time using various attributes on a myriad of subtopics and comprising discussion of contemporary issues and criticism, forming positive, negative, or neutral sentiments for objects/goods or services they use in their day-to-day routine. Manufacturing factories/companies of such products have started to collect and utilize these data to obtain a realization of the general sentiments for their products/objects or services. Microblogs are a shorter form of blogs that allow users to express their emotions in real-time. Sentiment analysis is the major aspect of the current NLP, as it is with Urdu Language Processing (ULP).

1.2. Attitude Scrutiny in Social Networks

Social media popularity impacts the preferences of customers, with Twitter and Facebook enjoying widespread popularity. This popularity not only enables people to express their opinions, but also provides guiding principles for others to follow and for companies/organizations to improve their products and services. Users write blogs as well as reviews on social media that are related to companies. Hence, customers' views are readily available for the company's use. Text mining is carried out to determine the opinions of customers. Website data and text are available and mined, providing a new dimension to the field of computer science. Sentiment analysis has recently emerged as a very important field with the development of social media. Public opinions are now available to study. Textual data are usually unstructured, and incomprehensible, vague, and unclear, making analysis difficult, but they contain a huge amount of valuable information if processed properly. Herein, we collected and processed Urdu text data on certain topics using the methods of NLP.

1.3. Research Inspiration

This work was carried out keeping in mind the following considerations:

1. People provide their views on social media to express emotions about certain products/services, especially in the Urdu language.
2. South Asian countries such as India, Pakistan, etc., use the Urdu language to express their feelings, comprising a huge online market for other countries.
3. No work has previously been performed that combines the latest deep learning algorithms to conduct such an analysis.

1.4. Contributions and Obligations

Our contributions and obligations are the following:

1. Elaborate the importance and utility of Urdu sentiments and perform an analysis thereof.
2. Investigate methods to perform S.A. in Urdu.
3. Apply machine learning and deep learning algorithms for S.A.
4. Apply a combination of algorithms to propose a hybrid approach for S.A.
5. Compare the applied algorithms to suggest the most effective ones for the Urdu language.
6. Suggest future work.

1.5. Relation to Previous Work

Research was performed to apply sentiment analysis on an Urdu tweets dataset, and five algorithms were applied and the results were compared [3]. Another study was conducted to apply machine learning and deep learning approaches such as SVM, LR, LSTM, and CNN on Urdu sentences for the analysis of polarity, with the results outperforming previous ones [4]. Research was performed by extracting a small number of tweets, both in Urdu and Roman Urdu, and the WEKA tool was applied but was not able to give results on all applied algorithms in Urdu [5]. Research was conducted as Urdu sentences were extracted. LSTM was applied for sentiment analysis and results were obtained. [6]. Sentiments from social data were analyzed using Urdu pre-existing algorithms, attaining adequate results [7]. Urdu language data were collected and some pre-processing was applied using classical machine learning along with deep learning

techniques to perform sentiment analysis [8]. The characteristic of context dependency of Urdu words was resolved to conduct Urdu sentiment analysis by assigning polarities to sentences; the accuracy of results therefore improved [9].

Sentiment analysis in Urdu was performed using a lexicon-based approach, a traditional method adopted using rules and results obtained in Urdu [9]. Multimodal datasets concerning text, audio, and videos of Urdu and comprising 1372 expressions were collected to identify context-aware sentiments, achieving a reasonable accuracy [10]. A review paper was written to examine the work conducted in Urdu and Roman Urdu for product reviews [11]. Urdu data were collected, annotated, and classified using PART, Naïve Bayes Multinomial Text, Lib SVM, decision tree (j48), and k-nearest neighbor; polarity prediction was also performed [12]. A corpus of Urdu Sentiments was developed comprising of tweets between two political parties, which were collected and tagged for polarity, and linguistics analysis was performed on them [13]. The polarity of sentences was calculated using a sub-opinion approach using orientation scores separately from sub-fragments [14].

Research was performed on Urdu language intensifiers by applying rules to assign polarities to them for S.A, and an Urdu sentiment analyzer was developed improving the accuracy [15]. People's feelings and behavior toward a product/service or brand has gained importance in the field of opinion mining, therefore opinion mining on Urdu comments on websites was performed [16]. A rule-based approach was used to handle negations in Urdu sentences, thus improving the accuracy of S.A [17]. A word-level translation scheme was introduced to create an Urdu lexicon as well as polarity scores to assign sentiments and performed evaluation [18]. A review of previous work on sentiment analysis of Urdu was conducted using machine learning to investigate improvement [19]. Reviews from various channels on sports, food, software, politics, and entertainment were collected to create a corpus in Urdu, and some machine learning models were applied for S.A [20]. English sentiments were used to form Urdu sentiment lexicons and ML was used to compare their performance [21]. Urdu opinion mining was performed by first collecting Urdu tweets as a form of pre-processing and then forming a feature vector, identifying positive and negative words as well as using a decision tree used for classification [22]. Another study was performed to identify Senti-units from Urdu text using parsing; a lexicon-based approach was used and performance was evaluated on various texts [23]. Another study was performed to classify tweets data into three classes using a Markov Chain, applying prediction improving results from previously used lexicon and traditional machine learning approaches [24]. A large scale dataset was created in the Urdu language by applying pre-processing, and emoji's were used to identify sentiments [25].

Songs were selected from YouTube and online reviews were compared for S.A [26]. A review was performed for this study using various ML algorithms, and their contribution toward opinion mining was evaluated [27]. An attempt was made to perform cross-domain S.A in Urdu using ML and DL classifiers [28]. An identification of the challenges and future directions of research to be performed in the Urdu language toward emotion detection was conducted [29]. A study was conducted during the Pakistan General Elections and during the promotion of political parties and public opinions on social media; we also observed the effects on the elections' results [30]. Urdu poetry was analyzed using the work of many poets over a long time period, exploring the main features of Urdu ghazal popularity and public liking [31]. The identification of characteristic-based sentiment analysis was performed with the help of a dataset for this purpose [32]. Another study was conducted to develop Urdu language communication on Twitter and observe its use in opinion mining [33]. Research was conducted to detect sarcasm and S.A relationship by using a cognitive approach to perform classification [34]. Sentences in English were extracted and translated to Urdu, grammatical errors were corrected using natural language processing, and an emotional analysis was performed using ML approaches [35]. S.A was performed on Urdu Blog data using structural correspondence learning (SCL), POS tagging, and the approach was validated using the supervised learning method [36]. Urdu news text was classified using 12 approaches of machine learning and compared [37]. Another study

was performed to study Urdu and English tweets/news data from websites in conjunction with the dengue epidemic, and S.A was performed to get insights using ML algorithms, with the data then being evaluated [38]. A hybrid feature selection approach was used to classify the Urdu text of news articles, giving improved results [39].

Efforts were made to develop an annotated corpus from the Urdu Nastalique Emotions Dataset (UNED), and machine learning and deep learning approaches were used to conduct emotion detection [40]. Using an Urdu dataset comprising of many labels, emotions were classified using machine learning [41]. Research was performed to reduce the gaps in pre-processing tools in Urdu by designing a stemmer, tokenizer, and preparing a stop-words list to perform Urdu news classification [42]. Urdu text document classification was performed using DL models in comparison with ML models in product manufacturing data [43]. The intent detection of users data in Urdu was performed after information retrieval using Bidirectional Encoder Representation from Transformers (BERT) [44]. To detect threatening content in the Urdu language, a stacking model was developed using Naïve Bayes and was applied for learning, while Logistic Regression was used for meta-learning; the stacked model compared well with other approaches [45].

The natural language processing and Urdu NLP tasks performed are shown in Figure 1 below:

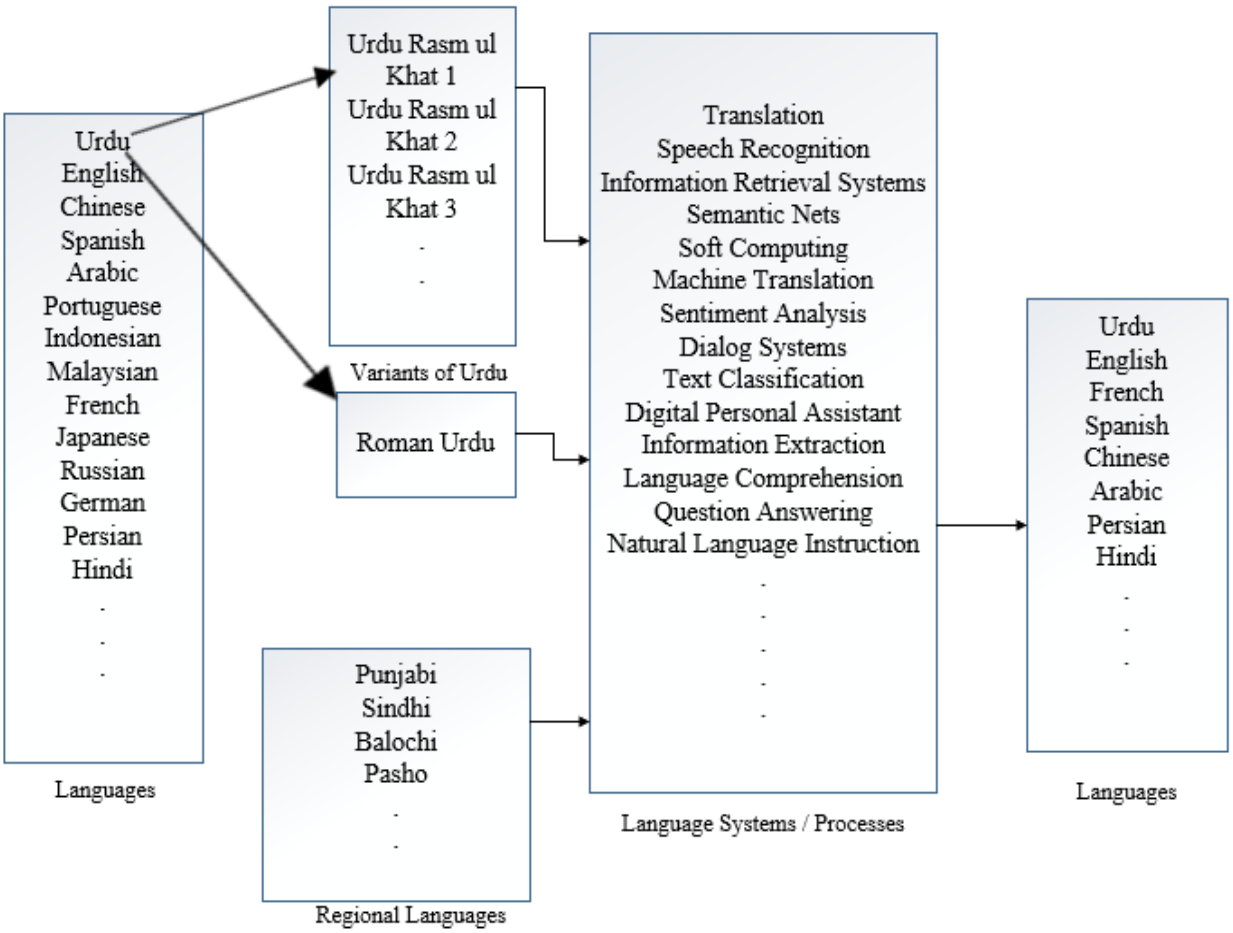

Figure 1. Natural language processing (NLP) of languages.

2. Material and Methods

This research was conducted to explore the following research questions:

2.1. Research Questions

The survey was conducted using the following research questions.

Research Question 1: Is the pre-processing required for Urdu text justified?
Research Question 2: How can the ML and DL methods be applied on Urdu text for extracted data?
Research Question 3: Can we apply more than one algorithm simultaneously on the Urdu data?
Research Question 4: Which algorithm was found to be more effective and accurate for Urdu language processing?

2.2. Technique and Criteria for Acceptance and Rejection

To obtain the most relevant research articles, different research queries were given using keywords such as "sentiment analysis"," Urdu sentiment analysis", "opinion mining", "Urdu opinion mining", "Urdu opinion mining for review," "Urdu tweets sentiment analysis ", "Urdu preprocessing", "sentiment classification in Urdu", "Urdu part of speech tagging", and "Urdu Classification method".

The selection and rejection criteria of the papers is as follows:

1. IP1: Inclusion Principle: All articles to be selected matching title or some words.
2. IP2: Inclusion Principle: Papers matching in abstract used to classify Urdu sentiments.
3. IP3: Inclusion Principle: Papers including methods used in Urdu sentiment analysis.
4. IP4: Inclusion Principle: Papers using some preprocessing for Urdu sentiment analysis.

Exclusion Principles (EPs) are:

1. EP1: Exclusion Principle: Remove all papers not following the above inclusion principles.
2. EP2: Exclusion Principle: Do not include any Roman Urdu papers.

2.3. Study Quality Evaluation

The quality of the study was maintained by adopting only selected papers for the study. Additionally, research questions mentioned in Section 2.1 were used to assess and address the quality of research. Two researchers analyzed all selected papers and only those selected qualified by meeting the criteria.

2.4. Survey Execution

A reasonably sized Urdu dataset was developed and used in this research to perform sentiment analysis. Initially, some pre-processing was performed on the data, and then the algorithm's application was studied and applied. Finally, the results were compared. The survey was conducted and categorized according to the research questions.

2.4.1. Data Collection

The data were collected from various blogs, websites, and tweets comprising of more than 10,000 sentences. A tedious exercise was conducted to obtain 10,009 sentences, which were manually annotated by 3 annotators in terms of being positive, neutral, or negative, as shown in Figure 2.

	A	B	C	D	E	F
1	**Language**	**Sentences**	**1**	**2**	**3**	**Labels**
2	**Urdu**	جہاں تک بہترین سوفٹویئر کی بات ہے تو بہترین سوفٹویئر وہ ہے جو آپ کو استعمال کرنا آتا ہو	positive	pos	pos	pos
3	**English**	As far as the best software is concerned, the best software is the one you know how to use				
4						
5	**Urdu**	معذرت چاہتا ہوں میں کسی اور خیال میں تھا میں یہ دونوں سوفٹویئر استعمال کر چکا ہوں لیکن صرف انگلش کے لئے یہ اردو سپورٹ نہیں کرتے	neutral	neu	neu	neu
6	**English**	Sorry I was thinking of something else I have used both these softwares but for English only they don't support Urdu				
7						
8	**Urdu**	اردو کے لئے سب سے آسان مائکروسوفٹ پبلشر ہے اور شاید سب سے سستا بھی	positive	pos	pos	pos
9	**English**	The easiest is Microsoft Publisher for Urdu and probably the cheapest				

Figure 2. Urdu text data annotated with labels.

Nearly 7000 sentences were found to lie in the above three categories by the annotators. Finally, two classes of positive and negative sentences were selected for this work. Hence, a gold standard dataset was formed to perform the analysis. It was divided into 3737 negative labels and 2815 positive labels; remaining sentences were undecided to fit in these two criteria, as shown in the figure below. The data were trained for two-class classification; before this, the data needed cleaning.

2.4.2. Data Cleaning

Data cleaning is an important step towards pre-processing, hence the removal of URLs from the data was performed; then, the data ere tokenized, removal of emails was performed, new line characters were removed, distracting single quotes were also removed, all the punctuation symbols were removed, then detokenization of the data was performed; finally, the list of texts was converted into a 'numpy' array as this is the requirement for the application of machine learning to the sentence needs to be converted into a list, as shown below:

'جہاں تک بہترین سوفٹوئر کی بات ہي تو بہترین سوفٹوئر وہ ہي جو آپ کو استعمال کرنا آتا ہو'

Then, the data were converted as follows:

'ہو'،'اتا'،'کرنا'،'استعمال'،'کو'،'اپ'،'جو'،'ہي'،'وہ'،'سوفٹویر'،'بہترین'،'تو'،'ہي'،'بات'،'کی'،'سوفٹویر'،'بہترین'،'تک'،'جہاں'

After being detokenized, the data were as follows:

'جہاں تک بہترین سوفٹویر کی بات ہي تو بہترین سوفٹویر وہ ہي جو اپ کو استعمال کرنا اتا ہو'

Initially, the machine learning algorithm Naïve Bayes and a support vector machine were applied. Naïve Bayes is a probabilistic-based machine classifier, while support vector machines also belong to a set of supervised machine learning methods for classification, with these methods being highly effective. Then, deep learning algorithms such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional LTSM, Bidirectional GRU, and Attention-based Bidirectional GRU were applied on the data. Then, a combinational algorithm was also tested on the data.

3. Results

The following answers were found for the research questions stated above.

3.1. Research Question 1

Is the preprocessing required for Urdu text justified?

The data were collected from various blogs, websites, and tweets comprising of more than 10,000 sentences. Urdu being a low-resourced language, this task proved to be a tedious one. Annotators were able to classify 7010 sentences into 3 classes: positive, neutral, and negative. A gold standard dataset was formed to perform the analysis. To perform two-class classification, two classes were selected, including 3737 negative labels and 2815 positive labels; remaining sentences were excluded, resulting in a balanced data.

Sentiment analysis cannot be performed before this, since the processing of text in any language, especially Urdu, contains many unavoidable words/symbols, etc., those used which have no significance in forming opinions. To process text in any language is difficult, but Urdu has its own challenges, which surpass all other languages, making this task more daunting. Urdu text/data preprocessing includes some steps to be followed. The processing of the Urdu dataset was performed as shown in Figure 3 below:

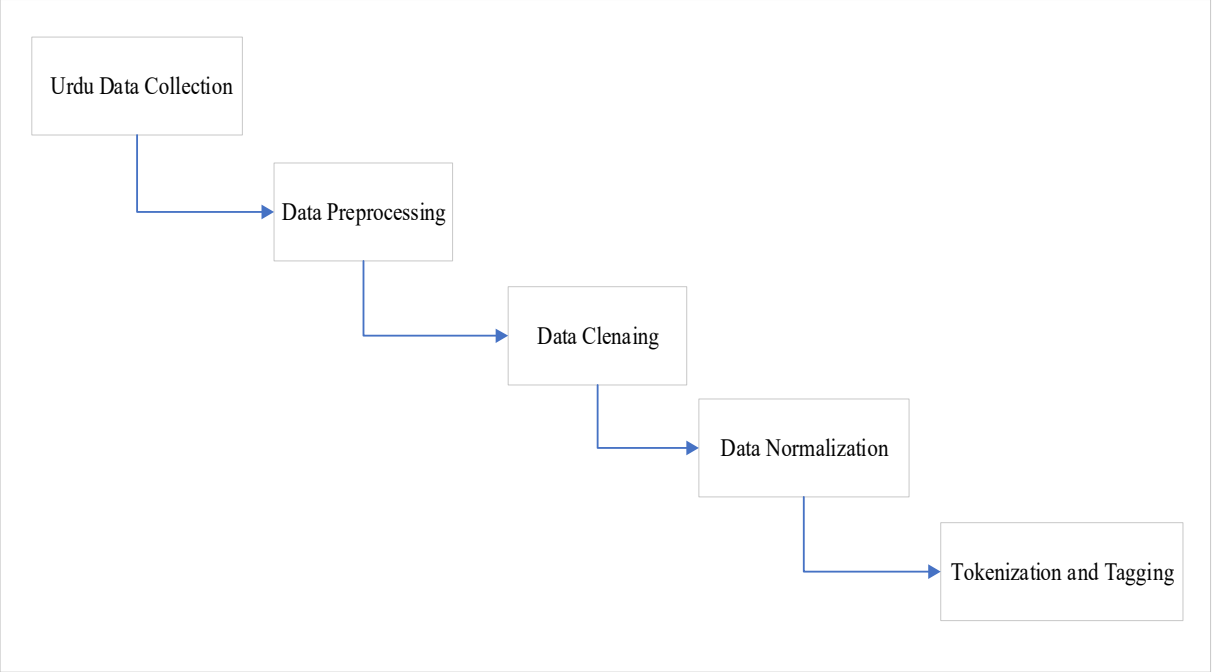

Figure 3. Processing of Urdu dataset before analysis.

Urdu Text Pre-processing

a. Data cleaning/noise removal;
b. Data normalization;
c. Tokenization and tagging.

3.1.1. Data Cleaning/Noise Removal

The data were to be trained for two-class classification, but before this the data needed to be cleaned for processing. The Urdu text from tweets comprised some additional 'noise' in the form of URLs, emails, new line characters, punctuation/signs; removing these from the data makes the data more suitable to use and more meaningful. Additionally, this improves the results in the classification and accuracy of the implementation of algorithms.

Typical steps identified and followed were as follows.

- Remove URLs from the tweets.
- Tokenize text.
- Remove email addresses.
- Remove new line characters.
- Remove all punctuation signs.
- Detokenize text.
- Convert list of text to 'Numpy' array.
- Apply text cleaning process to the data.

The steps involved, from data pre-processing until the application of the algorithms, are shown in Figure 4 as follows:

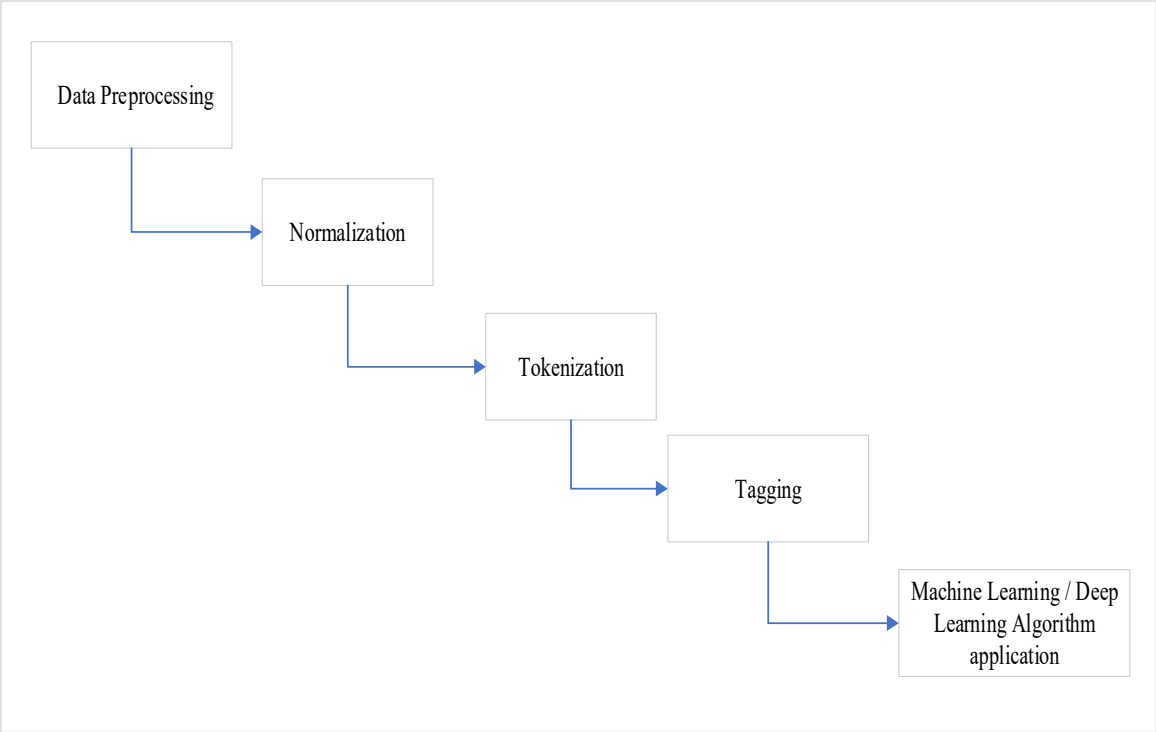

Figure 4. Steps involved toward algorithm application.

3.1.2. Normalization

The Urdu text data extracted required normalization to follow proper Urdu grammatical rules and to making the data meaningful as well. They required some normalization; grammatical and lexical rules also needed to be followed. This process must be applied before any algorithm to be applied for sentiment analysis in order to give better results.

3.1.3. Tokenization

Tokenization is performed using treebank tokenizer, as it is a must to break long sentences into understandable separate words for analysis; also, wordnet lemmatizer is used to convert words into their base forms, with this process being called stemming; stop-words are also removed from the text. The process of stemming in Urdu is shown below with the help of an example in Figure 5:

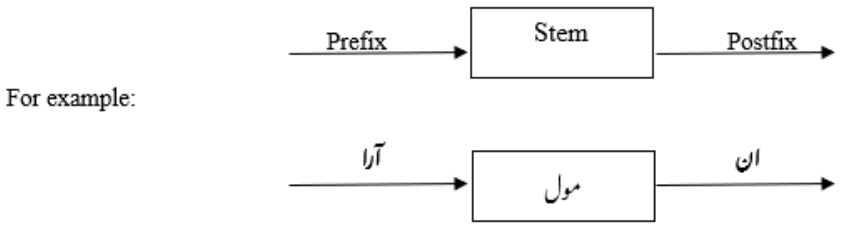

Figure 5. Stemming in Urdu.

3.1.4. Tagging

The Urdu text is then tagged and labelled manually by 3 annotators in the form of positive or negative sentiments to perform two-class classification with 0 and 1 labels, thus improving the quality of data.

3.2. Research Question 2

How can the ML and DL methods be applied on Urdu text for extracted data?

The extracted Urdu data needed to be converted into a readable form via the machine learning algorithms as well as deep learning, therefore the above-mentioned steps in Figure 3 were applied. Model building was performed using Keras API in Python. Initially, machine learning algorithms were applied to the Urdu dataset. Naïve Bayes was applied, and the model was trained, giving an accuracy result of 0.78265. The support vector machine applied on the same data gave an accuracy result of 0.56075.

Deep learning algorithms require all data to be converted to similar lengths, and so a maximum length of 139 words was selected based on the available data. A Convolutional Neural Network (CNN) was applied on the given data firstly by using sigmoid as the activation function and binary cross entropy as the loss model to train and test the model. A Simple Recurrent Neural Network (RNN) was applied on the data using the same perimeters; Single-Layered Long Short-Term Memory (LSTM) was also applied on the same data, and 20 epochs were processed for these models as the data were large, but the results obtained from the special GPU processing machine were used for model building. Bidirectional LSTM and a Bidirectional Gated Recurrent Unit (GRU) were also applied on the data. A complete depiction of Urdu text collection until the comparison of the sentiment analyses is shown in Figure 6 below:

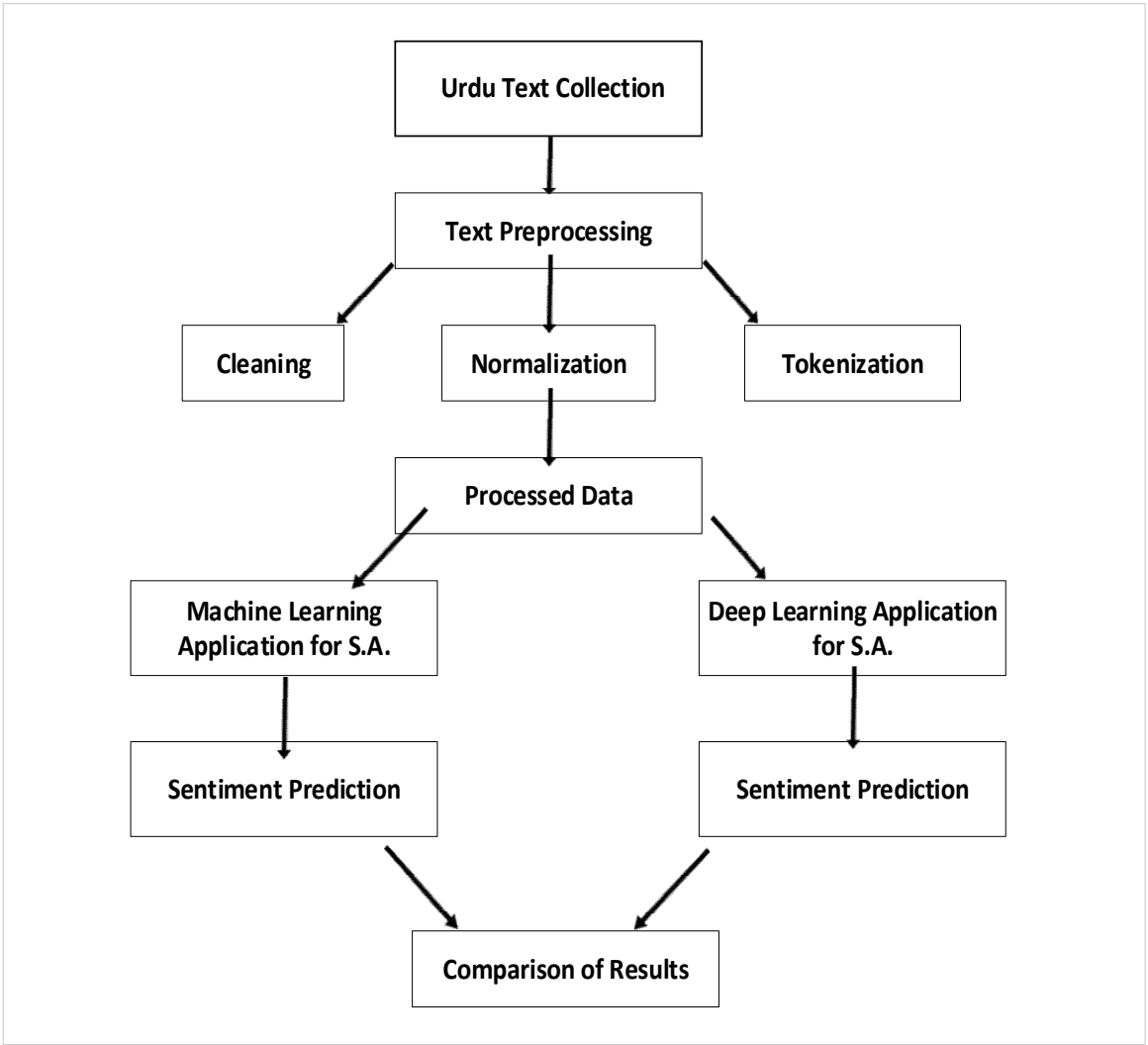

Figure 6. Process of sentiment analysis.

3.3. Research Question 3

Can we apply more than one algorithm simultaneously on the Urdu data?

An innovation in the application of models in this study was achieved, and a combination of models was also applied to the Urdu language data. LSTM was used in combination with GRU, resulting in the attention-based GRU (LSTM + GRU) algorithm; the results were then noted. Additionally, RNN was applied in combination with CNN resulting in a combined model (RNN + CNN), and the results for the Urdu data were noted. Hence, this showed that combinational models can also be applied on data, resulting in a Hybrid model that can be used for data analysis.

3.4. Research Question 4

Which algorithm was found to be more effective and accurate for Urdu language processing?

In order to evaluate the effectiveness and accuracy of various algorithms applied in the Urdu language data, the measures used are accuracy, precision, and recall. The accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions made. We calculate it by dividing the number of correct predictions by the total number of predictions, as shown in Equation (1):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

where, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

The precision is calculated as the ratio between the number of positive samples correctly classified to the total number of samples classified as positive (either correctly or incorrectly). The precision measures the model's accuracy in classifying a sample as positive, as shown in Equation (2):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2)$$

The recall is calculated as the ratio between the numbers of positive samples correctly classified as positive to the total number of positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3)$$

Various machine learning and deep learning algorithms were applied on the data. Naïve Bayes performed better than SVM, while deep learning algorithms performed efficiently; simple RNN gave acceptable results; LSTM improved results further than GRU; attention-based GRU proved even better in this regard; CNN also gave good results; a combination of two algorithms, namely RNN plus CNN, outperformed all other applied algorithms. A detailed discussion of the algorithm results obtained is conducted in Section 4.

The following answers were found for the research questions stated above.

4. Discussion

The extracted Urdu Data needed to be converted into a readable form by the machine learning algorithms as well as deep learning, therefore the above-mentioned steps in Figure 3 were applied. Model building was performed using Keras API in Python. Initially, machine learning algorithms were applied to the Urdu dataset. Firstly, the Naïve Bayes algorithm was applied, and the model was trained, giving an accuracy result of 0.78265. Another machine learning algorithm support vector machine applied on the same data gave an accuracy result of 0.56075; this shows that SVM did not perform well on the Urdu dataset.

Deep learning algorithms require all data to be converted to similar lengths, and so a maximum length of 139 words was selected in each sentence based on the available data. A Convolutional Neural Network (CNN) was applied on the given data firstly by using sigmoid as the activation function and binary cross entropy as the loss model to train and test the model; 20 epochs were run using the GPU machine. The final obtained accuracy was 85.8%, with a precision of 0.8799 and a recall value 0.8901. A Simple Recurrent Neural Network (RNN) was applied on the same dataset using the same parameters, resulting in an accuracy of 82.8%, a precision of 0.9144, and a recall value of 0.7931. The dataset was further tested on Single-Layered Long Short-Term Memory (LSTM), resulting in an accuracy of 84.0%, with a precision of 0.8757 and a recall value of 0.8160. Bidirectional LSTM was also tested, obtaining an accuracy result of 82.99%, with a precision of 0.9187 and a recall value of 0.7927. A Bidirectional Gated Recurrent Unit (GRU) was also applied, giving an accuracy of 84.4%, with a precision of 0.8960 and a recall value 0.8412.

An innovation in the application of models in this study was achieved, and a combination of models were also applied to the Urdu language data. LSTM was used in combination with GRU resulting in the (LSTM + GRU) algorithm, achieving an accuracy of 84.5% with a precision of 0.8889 and recall value of 0.8593. Additionally, the RNN was applied in combination with the CNN resulting in a hybrid model (RNN + CNN), and the results showed an accuracy of 85.8% with a precision of 0.8799 and a recall value of 0.8901 for the dataset. Comparing these results shows that a combination of the two models can also be applied on Urdu datasets, resulting in a Hybrid model for data analysis. Additionally, the results have improved in terms of the accuracy being above 85%, and both precision and recall being above 90% makes them superior to the previous figures achieved, as shown in Figure 7 below:

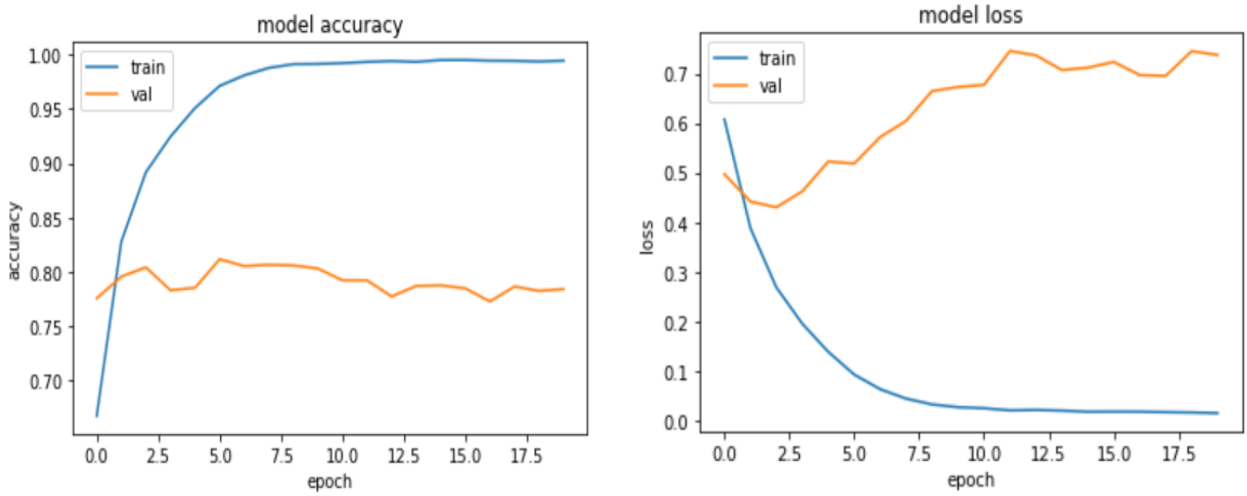

Figure 7. Graphs show (RNN + CNN) model accuracy and model loss.

5. Conclusions

With the growth in businesses and development of online strategies for e-commerce, it has become vital to understand public opinions and feel the pulse of customers in terms of products or services. Urdu, being a widely used language of South Asia, is of great importance and yet research on this language is limited, hence a lot of work is required. The primary objective of this work was to explore and implement the techniques for Urdu text preprocessing and develop a new source for the usage of data analysis. Various algorithms of machine and deep learning categories were implemented on the Urdu text data. One-class and two-class classification were applied on a large dataset, and the results were obtained. Machine learning algorithms showed good results, while the majority of deep learning algorithms improved the results further. A combination of machine learning algorithms was applied on the Urdu text data and led to the further improvement of results,

opening a new avenue for researchers. It is recommended that other combinations of algorithms be tested on different datasets of Urdu in the hope of leading to even better results for sentiment analysis, which can be used to make software that responds toward customer sentiments better than the existing ones.

6. Future Work

Future work can be conducted by applying various combinations of algorithms on Urdu text data to products or service reviews. Additionally, similar work can be conducted on other languages such as the Sindhi, Panjabi, Pashto, and Balochi languages.

7. Human and Animal Rights

This study did not involve any experimental research on humans or animals.

Author Contributions: Conceptualization S.M.A.B.; methodology, K.B.M. and S.M.A.B.; formal analysis, K.B.M. and S.M.A.B.; investigation, S.M.A.B.; writing—original draft preparation, K.B.M. and S.M.A.B.; writing—review and editing, K.B.M. and S.M.A.B. visualization, K.B.M.; supervision, S.M.A.B.; project administration S.M.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used for the work is available at: https://docs.google.com/spreadsheets/d/1eRmLmSVHlk-eHIu7gYm-qY9M6mRjvwnW/edit?usp=share_link&ouid=115369265597659036496&rtpof=true&sd=true (accessed on 10 January 2023).

Acknowledgments: The authors acknowledge all possible administrative and partial financial support provided by our institute, i.e., "College of Computer Science and Information Systems, Institute of Business Management" to conduct and publish this research work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weber, G. Top languages. *World* **2008**, *11*, 2009.
2. Tao, J.; Zheng, F.; Li, A.; Li, Y. Advances in Chinese Natural Language Processing and Language Resources. In Proceedings of the 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, Urumqi, China, 10–12 August 2009; pp. 13–18.
3. Ahmad, W.; Edalati, M. Urdu Speech and Text Based Sentiment Analyzer. *Comput. Lang.* **2022**. [CrossRef]
4. Sehar, U.; Kanwal, S.; Dashtipur, K.; Gogate, M.; Khan, F. *A Hybrid Dependency-Based Approach for Urdu Sentiment Analysis*; Research Square: Durham, NC, USA, 2022.
5. Rehman, I.; Soomro, T.R. Urdu Sentiment Analysis. *Appl. Comput. Syst.* **2022**, *27*, 30–42. [CrossRef]
6. Masood, M.; Azam, F.; Anwar, M.; Rahman, J.U. Deep-Learning Based Framework for Sentiment Analysis in Urdu Language. In Proceedings of the 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), Rawalpindi, Pakistan, 24–26 May 2022; pp. 1–7. [CrossRef]
7. Mashooq, M.; Riaz, S.; Farooq, M. Urdu Sentiment Analysis: Future Extraction, Taxonomy, and Challenges. *VFAST Trans. Softw. Eng.* **2022**, *10*. [CrossRef]
8. Khan, L.; Amjad, A.; Ashraf, N.; Chang, H.-T. Multi-class sentiment analysis of Urdu text using multilingual BERT. *Sci. Rep.* **2022**, *12*, 5436. [CrossRef] [PubMed]
9. Mukhtar, N.; Khan, M.A.; Chiragh, N.; Nazir, S.; Jan, A.U. An intelligent unsupervised approach for handling context-dependent words in Urdu sentiment analysis. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–15. [CrossRef]
10. Sehar, U.; Kanwal, S.; Dashtipur, K.; Mir, U.; Abbasi, U.; Khan, F. Urdu Sentiment Analysis via Multimodal Data Mining Based on Deep Learning Algorithms. *IEEE Access* **2021**, *9*, 153072–153082. [CrossRef]
11. Khan, I.U.; Khan, A.; Khan, W.; Su'ud, M.M.; Alam, M.M.; Subhan, F.; Asghar, M.Z. A Review of Urdu Sentiment Analysis with Multilingual Perspective: A Case of Urdu and Roman Urdu Language. *Computers* **2021**, *11*, 3. [CrossRef]
12. Mukhtar, N.; Khan, M.A.; Chiragh, N. Effective Use of Evaluation Measures for the Validation of Best Classifier in Urdu Sentiment Analysis. *Cogn. Comput.* **2017**, *9*, 446–456. [CrossRef]
13. Khan, M.Y.; Nizami, M.S. Urdu sentiment corpus (v1.0): Linguistic Exploration and Visualization of Labeled Dataset for Urdu Sentiment Analysis. In Proceedings of the 2020 International Conference on Information Science and Communication Technology (ICISCT), Karachi, Pakistan, 8–9 February 2020; pp. 1–15. [CrossRef]
14. Hassan, M.; Shoaib, M. Opinion within opinion: Segmentation approach for Urdu sentiment analysis. *Int. Arab. J. Inf. Technol.* **2018**, *15*, 21–28.

15. Mukhtar, N.; Khan, M.; Chiragh, N.; Nazir, S. Identification and handling of intensifiers for enhancing accuracy of Urdu sentiment analysis. *Expert Syst.* **2018**, *35*, e12317. [CrossRef]
16. Rehman, Z.U.; Bajwa, I.S. Lexicon-Based Sentiment Analysis for Urdu Language. In Proceedings of the 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 24–26 August 2016; pp. 497–501.
17. Mukhtar, N.; Khan, M.; Chiragh, N.; Jan, A.; Nazir, S. Recognition and effective handling of negations in enhancing the accuracy of Urdu sentiment analyzer. *Mehran Univ. Res. J. Eng. Technol.* **2020**, *39*, 759–771. [CrossRef]
18. Asghar, M.Z.; Sattar, A.; Khan, A.; Ali, A.; Kundi, F.M.; Ahmad, S. Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Syst.* **2019**, *36*, e12397. [CrossRef]
19. Liaqat, M.I.; Hassan, M.; Shoaib, M.; Khurshid, S.; Shamseldin, M.A. Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. *PeerJ. Comput. Sci.* **2022**, *8*, e1032. [CrossRef]
20. Safder, I.; Mahmood, Z.; Sarwar, R.; Hassan, S.; Zaman, F.; Nawab, R.M.A.; Bukhari, F.; Abbasi, R.A.; Alelyani, S.; Aljohani, N.R.; et al. Sentiment analysis for Urdu online reviews using deep learning models. *Expert Syst.* **2021**, *38*, e12751. [CrossRef]
21. Khan, M.Y.; Emaduddin, S.; Junejo, K.N. Harnessing English Sentiment Lexicons for Polarity Detection in Urdu Tweets: A Baseline Approach. In Proceedings of the 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 30 January–1 February 2017; pp. 242–249.
22. Bibi, R.; Qamar, U.; Ansar, M.; Shaheen, A. Sentiment Analysis for Urdu News Tweets Using Decision Tree. In Proceedings of the 2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), Honolulu, HI, USA, 29–31 May 2019; pp. 66–70. [CrossRef]
23. Syed, A.Z.; Aslam, M.; Martinez-Enriquez, A.M. Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. In *Advances in Artificial Intelligence: 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, 8–13 November 2010, Proceedings, Part I 9*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 32–43.
24. Nasim, Z.; Ghani, S. Sentiment analysis on Urdu tweets using markov chains. *SN Comput. Sci.* **2020**, *1*, 269. [CrossRef]
25. Batra, R.; Kastrati, Z.; Imran, A.; Daudpota, S.; Ghafoor, A. A large-scale tweet dataset for Urdu text sentiment analysis. *Comput. Sci. Math.* **2021**, 2021030572. [CrossRef]
26. Asif, M.; Qureshi, M.; Abid, A.; Kamal, A. A Dataset for The Sentiment Analysis of Indo-Pak Music Industry. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–6.
27. Devi, G.D.; Kamalakkannan, S. Literature review on sentiment analysis in social media: Open challenges toward applications. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 1462–1471.
28. Altaf, A.; Anwar, M.W.; Jamal, M.H.; Hassan, S.; Bajwa, U.I.; Choi, G.S.; Ashraf, I. Deep Learning Based Cross Domain Sentiment Classification for Urdu Language. *IEEE Access* **2022**, *10*, 102135–102147. [CrossRef]
29. Azam, N.; Tahir, B.; Mehmood, M.A. Sentiment and emotion analysis of text: A survey on approaches and resources. *Lang. Technol.* **2020**, *87*.
30. Soomro, T.R.; Ghulam, S.M. Current status of Urdu on Twitter. *Sukkur IBA J. Comput. Math. Sci.* **2019**, *3*, 28–33. [CrossRef]
31. Rabbani, S.; Qureshi, Z.A. Exploratory Data Analysis of Urdu Poetry. *Sci. Stud. Read.* 2021. [CrossRef]
32. Rani, S.; Anwar, W. Resource Creation and Evaluation of Aspect Based Sentiment Analysis in Urdu. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Suzhou, China, 4–7 December 2020; pp. 79–84.
33. Ghulam, S.M.; Soomro, T.R. Twitter and Urdu. In Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (ICOMET), Sukkur, Pakistan, 3–4 March 2018; pp. 1–6.
34. Khan, M.Y.; Ahmed, T.; Wasi, S.; Siddiqui, M.-m.S. Enhancing sarcasm and sentiment analysis with cognitive relationship: A context-aware approach for Urdu-a resource poor language. *Comput. Intell. Neurosci.* **2022**, *8*.
35. Chhajro, M.A.; Arshad, A.; Luhana, K.; Wagan, A.; Muneed, M.; Umrani, A.I. Electronic Ledger Management: A mobile-enabled sentiment reviews analysis of Urdu Language. *J. Tianjin Univ. Sci. Technolo* **2022**, *55*, 6.
36. Mukund, S.; Srihari, R.K. Analyzing Urdu Social Media for Sentiments Using Transfer Learning with Controlled Translations. In Proceedings of the Second Workshop on Language in Social Media, Montreal, QC, Canada, 7 June 2012; pp. 1–8.
37. Malik, K.I. Urdu news content classification using machine learning algorithms. *Lahore Garrison Univ. Res. J. Comput. Sci.* **2022**, *6*, 22–31. [CrossRef]
38. Ali, M.Z.; Javed, K.; Tariq, A. Sentiment and emotion classification of epidemic related bilingual data from social media. *Comput. Lang.* **2021**. [CrossRef]
39. Rasheed, I.; Banka, H.; Khan, H.M. A hybrid feature selection approach based on LSI for classification of Urdu text. *Mach. Learn. Algorithms Ind. Appl.* **2021**, *907*, 3–18.
40. Bashir, M.F.; Javed, A.; Arshad, M.; Gadekallu, T.; Shahzad, W.; Beg, M.O. Context aware emotion detection from low resource Urdu language using deep neural network. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**. [CrossRef]
41. Ashraf, N.; Khan, L.; Butt, S.; Chang, H.-T.; Sidorov, G.; Gelbukh, A. Multi-label emotion classification of Urdu tweets. *PeerJ Comput. Sci.* **2022**, *8*, e896. [CrossRef]
42. Farooq, A.; Noreen, Z.; Batool, S.; Naz, F. Urdu News Classification: An Empirical Study Using Machine Learning Techniques. In Proceedings of the 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 27–28 October 2022; pp. 1–7. [CrossRef]

43. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.; Abdelmajeed, M.; Fayyaz, M. Exploring deep learning approaches for Urdu text classification in product manufacturing. *Enterp. Inf. Syst.* **2022**, *16*, 223–248. [CrossRef]
44. Shams, S.; Sadia, B.; Aslam, M. Intent Detection in Urdu Queries Using Fine-Tuned BERT Models. In Proceedings of the 2022 16th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 14–15 December 2022; pp. 1–6. [CrossRef]
45. Mehmood, A.; Farooq, M.S.; Naseem, A.; Rustam, F.; Villar, M.G.; Rodríguez, C.L.; Ashraf, I. Threatening URDU Language Detection from Tweets Using Machine Learning. *Appl. Sci.* **2022**, *12*, 10342. [CrossRef]