Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

A Systematic Review on Sentiment Analysis for Sindhi Text

***Safdar Ali Soomro* *[1], *Siti Sophiayati Yuhaniz[1], *Mazhar Ali Dootio[2], *Ghulam Murtaza[3], *Muhammad Hussain Mughal[3]**

[1]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.
[2]Department of Computer Science, Benazir Bhutto Shaheed University, Karachi, Pakistan.
[3]Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan.
*Corresponding Author.

Received 15/02/2024, Revised 11/08/2024, Accepted 13/08/2024, Published Online First 20/11/2024

Abstract

The field of sentiment analysis has experienced significant growth in recent years due to its applications in various domains such as news headlines, online product purchase, marketing, and reputation management. With the rise of social media and online shopping platforms, there is a wealth use-generated data available. This has led manufacturing, sales, and marketing companies to seek global feedback on their practices and products from these sources. In the context of Sindhi language, millions of phrases are shared daily on news media sites, Twitter, Facebook, and other platforms. However, the exclusion of sentiment analysis for Sindhi language limits the utilization of this vast amount of data, focusing primarily on the resource-rich English language. This systematic review aims to collect and evaluate published research related to Sindhi language sentiment analysis, specifically focusing on pre-processing, feature extraction, classification methods. The study offers a comprehensive analysis of research conducted on Sindhi text for product evaluation, covering key areas, such as relevant corpora acquisition, data preprocessing, feature extraction, classification techniques, methodologies, limitations, and future directions. Each reviewed article is assessed and classified based on specified criteria. The findings of this review provide valuable insights and propose several approaches for future investigations in this area.

Keywords: Natural Language Processing, Sentiment Analysis, Sindhi Corpus, Sindhi Text, Systematic Review, Text Pre-processing.

Introduction

One of the main means of human communication is language, and natural language processing is the general term for the study, investigation, and examination of languages, which includes data capture, translation, data normalization, sentiment analysis, morphological analysis, and text analytics of languages. Sentiment analysis (SA) is a crucial branch of natural language processing that helps to identify and extract subjective data from written material [1]. It is the main tool for determining the underlying sentiment, whether good or negative, in writing. SA frequently referred as opinion mining, it is a method of assessing the polarity of comments or judgements expressed by individuals in order to score items or services. SA may be performed on a document level, assessing the whole document to identify sentiment polarity using feature extraction [1]. It might have been done at level of the sentence, with

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

contents divided into phrases and analyzed independently to identify the document's polarity. Analyzing the sentiments would need a variety of methodologies, including semantic, statistical, and linguistic. Morphological techniques pertain to the structural annotation of adjectives as well as adverbs, which have a significant impact on sentiment identification [2]. Nowadays, Facebook, news websites, e-commerce websites, and Twitter are all examples of social media platforms that create massive volumes of data on daily basis [1,3]. The majority of businesses analyze this kind of information to convert social media activities into usable business data. Although such type of data is typically unstructured which makes it hard to deal with. Unstructured data can be an obstacle, reducing the overall quality of SA process. Once compared to lengthier texts and normal documents, social media contains enormous types of noises such as jargon, errors in typography, the repetition of characters within word, mistakes in spelling, poorly structured statements, combinations of words, uncommon usage of abbreviations, diverse forms of word acronym, different syntax and an overall unstructured style of language [4]. Due to this, data pre-processing is now considered to be an essential step in SA. Mostly all such type research focuses on high resource language such as English, Chinese, Arabic and other languages.

Over 75 million people speak Sindhi [5]. Sindhi language origins may be discovered in Pakistan's province Sindh. Sindhi is the predominant language of those who live in Sindh province. Furthermore, Sindhi is spoken in certain Indian states as well [6]. Sindhi is a sophisticated language which consists 52 letters as compared to Urdu's 36, Arabic 28, English 26 and its unique linguistic peculiarities.

Therefore, SA is quickly becoming a crucial tool for monitoring and understanding sentiment in all types of information as individuals communicate their ideas and emotions more freely than ever before. Sentiments allow you to easily explore people's opinions about any person, news, item, or policy. There are no restrictions on individuals expressing their opinions on any news topic or product. People's sentiments can be divided into categories such as 'document level, phrase level, and aspect level'. Here is a sample of a Sindhi statement using Perso-Arabic script style to show somebody's sentiments for someone (آهي ماڻهو سڻو هو.) roman of that sentence is (Hoo sutho manhu aahe). In that Sindhi sentence, the word "سڻو" roman of that word is "Sutho" means "good" in English language. As a result, the sentence's polarity is examined positive. In second sentence (آهي ماڻهو بيڪار هڪ هو) roman of that sentence is (Hoo hik bekaar manhu aahe) in which the word "بيڪار" roman of that word is "Bekaar" means worthless in English language shows polarity of the sentence as negative. Likewise, some statements combine English and Sindhi sentences, such as: (Go with him آهي ماڻهو سڻو هو) a unique preprocessing method were used, in which the English and Sindhi sections were taken independently with their subjectivity. Eventually, sentence polarity was achieved by mixing both English and Sindhi sections. Furthermore, like in the preceding instances, Sindhi sentences pertaining to any product-based reviews may include such Sindhi terms that demonstrate the polarity of sentences. SA of Sindhi text can provide valuable insights into public opinion and sentiments regarding various topics and events. It can help government, media outlets and journalists understand their audience perceives news and information, and can also be useful for businesses and organizations to monitor their reputation and brand sentiment in the Sindhi-speaking community. Furthermore, SA can be used by government policymakers to gauge public opinion on various issues and make informed decisions accordingly.

Review Motivation

Now a days online shopping is growing day by day around the globe as well as in the subcontinent countries such as in Pakistan and India. Mostly people like to write the reviews regarding various online items in their native languages other than English. Sindhi is a provincial language of Pakistan and it is scheduled national language of India, which is widely spoken language in Indian states [6]. Now days a lot of Sindhi speaking persons are settled in different countries around the world. Sindhi language data has grown dramatically on internet in recent years. People wrote text in Perso-Arabic Sindhi to convey their thoughts on current events or

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

items. Sindhi opinions rapid research breakthroughs have motivated to conduct an extensive study by finding, selecting, summarizing, and analyzing pertinent papers. This study aims to present studies about text processing techniques, tools, and as well as SA approaches for Sindhi language. Furthermore, focuses on different linguistic methodologies used in SA, SA algorithms, text pre-processing, limitation of existing techniques, grammar, presentation of sentences and recommendations for future paths in SA particularly in Sindhi language.

Related Work

SA in Sindhi is still in its early stages as compared with English language which is highly resourced language. Moreover, certain research work has done. However, it is still in its infancy in terms of computer processing since, because of its complicated morphological structure and lack of language resources, it has gotten little attention from the language engineering community [7]. Therefore, it requires more work to compete with developed languages in terms of computing. In addition, this study describes a review of a few chosen research publications that present their work on other than Sindhi language.

SA is effective in determining a user's mood, emotion, attitude, feeling, temper, and personality traits based on the contents of their written messaging [8,9]. Sentiments or opinions may be defined as people's beliefs, thoughts, and attitudes based on their feelings rather than reasoning. Hence, SA recognizes the polarity of people's likes, dislikes, and behavior via written text [10]. The majority of individuals accessing websites, news and blogs is increasing every day. As a result, corporate communities, politicians, and both public and private sector organizations prefer to disseminate news, latest items and their properties on social and particular websites in order to learn about the thoughts and sentiments of various sorts of individuals or users [11].

SA may be divided into three types: sentence level SA, document level SA, and aspect level SA [12]. Lexicon-based techniques are important for various types of SA systems. The lexicon-based technique is feasible since it does not offer training to the machine, while machine learning does. Lexicon-based investigations use dictionaries that give a list of sentiment-based terms [13]. These terms are crucial for SA based on a lexicon. In addition, machine learning techniques may provide greater results when compared to it.

Training is the foundation of the machine learning-based SA approach. The machine is trained using opinionated terms so that an appropriate test may be done successfully [14]. Because sentiment-based text corpora and data sets are important for this strategy, they are created before employing the machine learning algorithms.

According to Tripathy et al. [12], the use of machine learning algorithms for SA operates on two levels: supervised and unsupervised. Furthermore, researchers said that in supervised learning for SA, target classes are labelled in order to give training so that acceptable output may be obtained. The unsupervised machine learning approach does not utilize labelled data; instead, data is sorted into clusters and unsupervised algorithms are used to do SA. To execute accurate aspect-based SA, the machine learning method for feature analysis detects the characteristics of items. Furthermore, it needs to pre-processes the data to remove stop words from data. After pre-processing the data machine learning models such as the n-gram model, the TF-IDF model focuses on the significance of a term or words [15].

Fig. 1 depicts the technique or approach to evaluate similar articles for this study.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

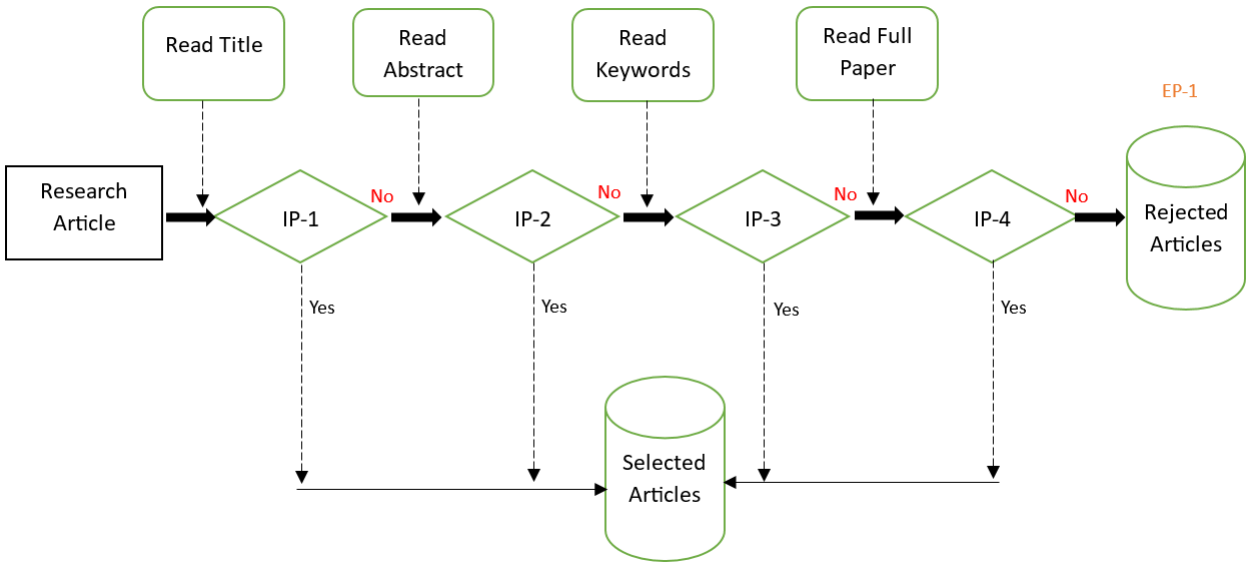

Figure 1. Diagram of Searching and Filtering Research Articles

Research Survey Methodology

This survey is carried out by looking for relevant publications in four most prominent electronic repositories such as IEEE Xplore, Scopus, ScienceDirect, and Web of Science. ScienceDirect and IEEE Xplore are the prominent publishers around the globe to cover most of the research areas of multidisciplinary. As well as Scopus and Web of Science are the prominent scholarly databases that includes papers from various publishers related to vast research areas from the globe. By using these databases researchers were able to analyze publication trends and development of the respective research area. The number of obtained articles is then filtered using inclusion and exclusion criteria. Finally, appropriate publications are chosen based on research questions (RQ), and findings are provided after thorough review.

RQ1: What text pre-processing techniques are employed for Sindhi SA and what approaches are employed by researchers, as documented in published articles?

RQ2: What exactly does feature-extraction entail? Which feature-extraction methods are employed for Sindhi corpus in chosen research papers?

RQ3: What exactly does sentiment categorization entail? How do different categorization methods assess Sindhi text sentiments in the chosen articles?

To discover the most relevant research articles, a structured keyword based search was conducted by submitting several search queries. Various search queries were provided by using keywords including ‘‘sentiment analysis in Sindhi”, ‘‘opinion mining in Sindhi”, “Sindhi sentiment classification method”, ‘‘sentiment classification of Sindhi text”, ‘‘subjectivity analysis in Sindhi text”, ‘‘preprocessing in Sindhi sentiment analysis”, “Sindhi preprocessing”, “Sindhi part of speech tagging”.

The criteria for article acceptance using Inclusion Principle (IP) and rejection, were followed as specified in [16-18]; see Fig. 1.

IP1: Include an article that is totally or partially linked to a specified title.

IP2: Include an article with an abstract linked to some sentiment categorization approach for a product evaluation in Sindhi Language.

IP3: Include an article demonstrating some new approaches in SA for the Sindhi Language.

IP4: Include an article demonstrating the acquisition of corpus of Sindhi sentiment using a pre-processing approach.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

The Exclusion principle (EP) is expressed as follows. EP1: Exclude any article that do not match the acceptance criteria in IP1 to IP4.

Research Survey Quality Assessment

To maintain the assessment quality, the process described in articles [16-18] was used for selected publications. Every research paper was evaluated using the research questions listed in 'Research Survey Methodology'. Each quality assessment question was entered into Excel spreadsheet and assigned a pre-determined rating: value "1" was assigned to research papers with completely explained answers, value "0.5" was assigned to questions with partially explained answers, and value "0" was assigned to questions with no explanation. There were three research questions utilized for quality assessment.

Table 1 shows the results of assessed questions on 4 selected research papers addressed in this section of the study. From overall score of 3, research paper S3 scored a 3.0 with a normalized value of one, and article S4 scored a 2.5 with a normalized score of 0.83, but research papers (S1, S2) received a 2 with a normalized score of 0.67. The threshold for the quality score was 0.5. Any articles below this score were rejected as they did not fulfill the quality criteria.

Table 1. Sample of Quality Assessment Scores of Selected Studies

Quality Assessment Criteria	Questions	Example Studies				Remarks
		S1 Ali et al. [19]	S2 Surahio et al. [20]	S3 Ali et al. [21]	S4 Sodhar et al. [22]	
QA1	The article delivers report of one or more pre-processing approaches used for Sindhi SA.	1	0.5	0	0	Study S3and S4 elucidated no pre-processing approaches in their study
QA2	The article provides details of feature extraction techniques used for Sindhi SA.	1	1	1	1	All researchers discussed feature extraction techniques in their study
QA3	The article clearly states sentiment classification of Sindhi content using approximately new technique.	1	1	1	1	All researcher one or more classification technique along with proper platform
Aggregate (out of 3)		3	2.5	2	2	Cumulating the scores in the preceding rows
Normalized Score (0-1)		1	0.83	0.67	0.67	To normalize results, divide the preceding row's scores by three

Research Survey Execution and Classification

Only twenty-five (25) research papers were obtained according to the criteria indicated in segment 'Research Survey Criteria' from four prominent research paper databases, including IEEE Xplore, Scopus, ScienceDirect, and Web of Science. In the first phase, twenty (20) papers were chosen on the first inclusion criteria; in the second phase, researchers read the abstract of the papers and evaluate papers according to the second inclusion criteria, in third and fourth phase researchers read the complete papers and applied the (third and fourth) inclusion criteria to evaluate the papers and at the end rejection criteria were used. At the end, the survey in this study included ten papers.

A full overview of selected research publications on Sindhi sentiments was included in the survey assessment. The goal of this study was to list most of

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

the relevant tasks that can fulfill the knowledge gap and identify methods to execute Sindhi SA in a far more precise manner. Study was conducted on the way to investigate the concepts of many researchers on pre-processing, feature-extraction, and the use of numerous classifiers with their benefits and drawbacks in Sindhi text.

Findings

The survey was carried out and classified in accordance with the study questions listed Below.

RQ1. What text pre-processing techniques are employed in Sindhi SA and what approaches are employed by researchers, as documented in published articles?

Text pre-processing in sentiment analysis indicates to the process of formulating input data for the resulting level of assessment and verification. Text pre-processing is a challenging process, and it becomes more complex for right hand written languages besides English. Because of scarcity of available resources, since each language has its own word segmentations, POS tags, and structural challenges.

In this section, investigators focused on the Sindhi preprocessing approaches presented in selected papers. Preprocessing for Sindhi sentiments analysis is often separated into three processes such as “Sindhi Word Segmentation, Tokenization, and POS Tagging”, which are outlined below and graphically shown in Fig. 2.

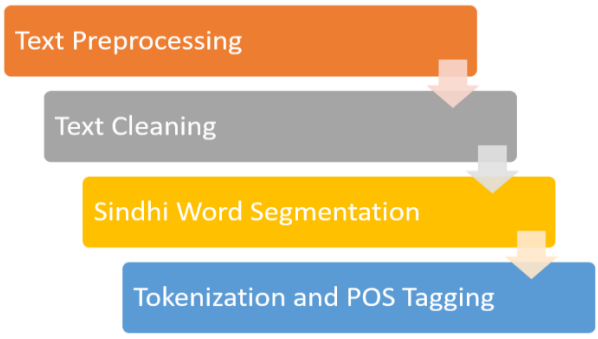

Figure 2. Flow of Preprocessing Techniques

Text Cleaning

Special characters, HTML elements, scripts, punctuation, emails, URLs, and adverts are some of the most common types of noise found in online text. Removing them entirely reduces noise in the text, which enhances the performance and accuracy of text classification algorithms to some amount. Pre-processing is an extremely critical phase in Sindhi sentiment analysis. Fig. 3 depicts the detail of the text cleaning process.

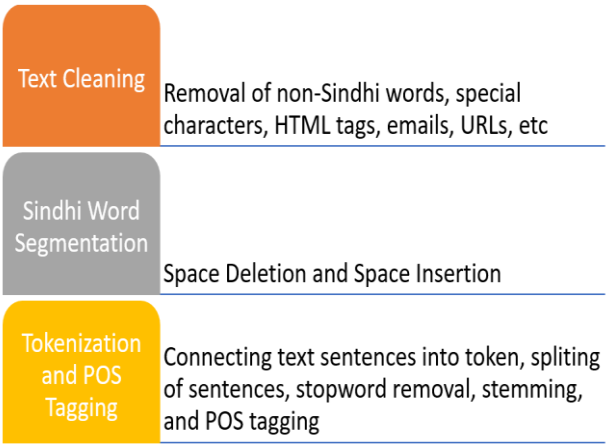

Figure 3. Detail of Each Step Involved in Pre-processing Techniques

Data pre-processing is an essential step before it is carried on for further analysis. To normalize the dataset, Hammad et al. [23] used a variety of strategies. Researcher did the following process to delete unnecessary whitespace before and after the text, as well as within words, with simply one space between them. Secondly, eliminating hashtags because the majority of tweets content comprises hashtags, hashtags are a very widespread social media trend. Third, text punctuation marks are being removed such as ([|@#/:.!"#$%&()*+,-./:;=>?@[] '|]). Fourth, tweets are being cleaned of pictograph, emoticon, symbols, map symbol, and flags. At the end, duplicate tweets are removed by comparing both the content and the tweet ID. Moreover, matching text since many tweets were reposted and appeared often in the dataset. Afterwards pre-processing steps, the cleaned dataset got which only have tweets in Sindhi with no additional text. Dataset left with 4236 tweets after removing duplicates. According to Wazir Ali et al. [20] pre-processing is accomplished by several text cleaning stages on acquired data. Construct a pre-processing workflow for removing unnecessary data from crawled tweets in order to achieve the desired

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

content for annotation, that consist of three steps first removed unnecessary punctuation marks at the starting and ending of tweets. Second noise filtering, like special-characters, non-Sindhi text, HTML elements, emails, and URLs. Finally, normalization, including the elimination of identical copies, numerous spaces, and tweets containing just user references. Moreover, eliminate phrases that are more than 80 words but less than 5 words. Mazhar Ali Dootio et al. [24] explain pre-process method in their research that for the sake of model building and performance analysis, the text corpus is pre-processed and standardized. However, the procedure of identifying and removing stop words was done carefully since pronouns, particularly personal pronouns, might be valuable for text corpus analysis. Surahio et al. [20] used pre-processes as a model step in which input consists of manually annotated corpora based on a tag set of 67 tags. Input passed to the pre-processing phase, in which tagged lexicon for each work is extracted, relevant input is converted to vector construction, and a training file for each tag is generated. Mazhar Ali et al. [24] in their research discussed that mostly corpus includes some undesired information. Thus, filtering and normalizing such data is critical in order to acquire a more accurate vocabulary for the annotation process. The following process comprise the pre-processing steps such as multiple punctuation marks at the beginning and ending of phrases are removed. Filtering out invalid data such non-Sindhi words, special characters, HTML elements, emails, URLs, and so on. Then tokenization is used to normalize the text, as well as the deletion of duplicates and multiple white spaces.

Word Segmentation

For unsegmented languages, the segmentation approach is the first and most important stage [7]. This study is significant in Sindhi Language Processing since Sindhi text segmentation is a requirement for numerous applications such as discretization, text-to-speech synthesis, and text recognition. Other NLP tasks may benefit from the most accurate separated text. Using a lexicon-driven method, Mahar [25] developed an algorithm for Sindhi word segmentation. The experiment began with 16,601 words that had prefixes, suffixes, and stems. The proposed approach was evaluated on 3,984 words and had a cumulative segmentation error rate (SER) of 9.54%. Narejo et al. [26] recently created and implemented a method for compound and complicated word segmentation. Their technique is capable of segmenting words into potential morphemes. Meanwhile, the system was evaluated on a modest quantity of data, with 109 compound words, 179 prefix, 1343 suffix, and 50 prefix-suffix words utilized for testing, and a satisfactory cumulative SER of 5.02% was reported by applying the suggested approach to both compound and Complex Sindhi terms. The suggested algorithm is capable of dealing with all fundamental grammatical components of the language, such as root words, prefixes, and suffixes. The work described here serves as a base for segmenting difficult and compound words into their simplest free forms. Furthermore, it is stated in their studies that strengthening the context of lexicons might lower the SER much further.

Tokenization

It is the process of converting lengthy text into words [27]. Word tokenization is a prerequisite for NLP activities for example POS recognition, parsing, NER and so on, as well as standalone NLP tasks. The Sindhi language processing (SLP) researchers use a variety of strategies to address various Sindhi word tokenization challenges. Researchers have accomplished outstanding outcomes and made significant contributions to the SLP research field. Dictionary or lexicon, and statistical or machine learning are the most often used methodologies for Sindhi word tokenization. As compare with the right-hand written language such as Arabic POS tagging applied on each term to get the tags named “noun, verb, adverb and etc” [28]. This process will help to find the adjectives and adverbs.

Sentence Splitting

The procedure that establishes sentence boundaries. It is an essential preprocessing step for several language analysis procedures, including tokenization, POS tagging, NER, parsing, and information retrieval. It is a difficult process in Sindhi since the language utilizes several markings for sentence boundaries.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Stop Word Removal

Stop words are words that appear often but have no important significance in any particular language. Normally, removing these terms enhances the whole outcome of SA methods [24]. Sindhi stop-words were taken from a list of prepositions or adverbs, determiners, verbs, conjunctions, interjections, and articles. Stop-words have no significance in content summarization and tend to be perceived as worthless. They are just purposeful terms in any language. Because these terms have no significance, they are removed from the corpus to minimize its size. Because every language has a set of specified stop-words, they have been deleted through that list. The deletion of stop words has the primary benefit of returning only relevant documents in information retrieval.

Stemming

This procedure converts terms into their basic form; such as the term "drinking" is converted to the basic term "drink". Stemming is fundamental text analysis pre-processing method. Stemming, a dictionary-based approach, was used to organize similar phrases with comparable meanings [29]. The goal of stemming aims to reduce the token to its original or base word. Stemming is a popular practice while working with written texts before information retrieval and NLP [24]. The process of tumbling a given term to its stem, origin or basic form is known as stemming. For example, the stem of Sindhi word "مددگار" in English (helpful) is "مدد" in English (help).

RQ2. What exactly does feature-extraction entail? Which feature-extraction methods are employed for Sindhi corpus in chosen research papers?

Text phrases are really a great resource of features for feature-extraction. The techniques of extracting such characteristics are known as feature-extraction. Text feature-extraction is a concept used in SA to describe the process of creating results from extracted contents and converting that results into feature combinations that can be utilized by a classification algorithm.

Sindhi Corpus

Mazhar Ali Dootio et al. [24] proposed Sindhi text corpora to provide NLP specialists and academics with text resources. Sindhi text corpora have been created in order to undertake more study on Sindhi language variants, feature extractions, stemming, lemmatization, POS tagging, syntactic analysis, SA, language modelling, and information retrieval. As a result, the Sindhi text corpus generated is a sentiment-based text corpus that characterizes content with +ve and -ve polarity. The corpus comprises 11864 +ve polarity documents and 3924 -ve polarity documents. The produced Sindhi text corpus has 15788 items in total. In the Sindhi text corpus, there are 23728 words with +ve polarity and 7848 words with -ve polarity. Additionally, the Sindhi text corpus comprises 7 characteristics of 2 electronic gadgets, as well as the opinion polarity of each feature. The 2 electronic gadgets are laptop computers and mobile phones from various manufacturers. These two gadgets have the following features: a battery, a microphone, speakers, a camera, a display screen, memory, and a price.

Mazhar Ali et al. [24] created a Sindhi annotated corpus utilizing a UPOS (Universal Parts of Speech) tag set and a SPOS (Sindhi Parts of Speech) tag set for the goal of analyzing linguistic characteristics and disparity. The Term Frequency and Inverse Document Frequency (TF-IDF) approach is used to extract the features. The supervised model is created for testing the annotated corpus in order to determine the structural annotation of the Sindhi text. This approach evaluated on 20% of the test set after training on 80% of the annotated corpus. The model's accuracy was reviewed and verified using the cross-validation procedure by utilizing 10-folds. The model's findings indicate improved performance and validate the correct annotation to the Sindhi corpus.

The Sindhi annotated corpus is a corpus dataset with several classes and features. Sindhi annotated data collection has four characteristics such as 'UPOS tagging', 'SPOS tagging', 'lemmatization', and 'stemming' are all techniques used. All of these characteristics are critical for Sindhi text analysis for the reason that Sindhi text corpus named ('UPOS' and 'SPOS') tagging supports in syntactical, lexical, and sentiment analysis. The lemmatization procedure makes Sindhi words are distinct in terms

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

of their basic morphology or syntax, this helps in determining the structural position of each stated lexicon. Sindhi lexicons' root terms are revealed through stemming words. As a result, stemming words may be used to create hierarchical lexicon trees. Each ('UPOS' and 'SPOS') tag is allocated a numerical number to be processed by the model. The Sindhi annotated dataset has six attributes such as 'WordID', 'UPOS', 'SPOS', 'WORD', 'STEM', and 'LEMMA' [24]. The 'WordID' displays the numerical number of a Sindhi words or lexicon. Attributes 'UPOS' and 'SPOS' display the commonly used words and SPOS sets. The 'Stem' attribute displays the stemming words of related Sindhi language. The 'Lemma' attribute displays Sindhi lexicon lemmas.

Subjective Sentiment Lexicon

Using a bilingual English-Sindhi dictionary, Wazir Ali et al. [19] created a Sindhi subjective words by expressing the emotion polarity value across all English opinion terms. Researchers used Bing Liu's opinion terms with the NRC vocabulary to produce the Sindhi subjective lexicon. Afterwards, sentimental polarity was assigned using SentiWordNet (SWN 3.0) and a multilingual dictionary is used to translate to Sindhi. The process of construction of Sindhi subjective lexicons are as follows:

- The NRC lexicon is a collection of English opinion words connected with fundamental sentiments such as nervousness, anger, sorrow, contempt, excitement, and happiness. There are 2,312 +ve and 3,324 -ve terms in the vocabulary.
- Bing Liu's terminology is all-purpose. The English sentiment lexicon has 2,036 +ve 4,814 -ve terms.
- SentiWordNet 3.0 includes an English WordNet synset of 117,659 words. Each word is assigned numerical sentiment value ranging from (0.0, 1) to signify whether the attitude is positive, negative, or neutral.
- Sindhi modifiers strengthen or weaken the mood of sentiment lexicon. As a result, researchers collected 173 Sindhi modernizers and allocated polarity by using SWN 3.0 and person judgement. In the absence of an English translation of Sindhi modernizers in the SWN 3.0 data set, researchers manually give the score to modifiers.
- Using a comprehensive online English to Sindhi dictionary, each English opinion term is translated to the appropriate Sindhi word. Whenever a bilingual dictionary yields many meanings for an emotion term, the first or precise meaning is used while discarding less prevalent interpretations.

To create a sentiment lexicon, Wazir Ali et al. [19] combined Bing Liu's and NRC lexicons and removed redundancies. SWN 3.0 is used to give polarity scores to each word in the list.

The key contributions of [19] are as follows:

- The creation of a polarity allocated Sindhi subjective lexicons via the integration of current English materials.
- Obtaining and annotating a multi-domain tweets corpus for Sindhi Sentiment Analysis using the Doccano text annotation tool.
- A subjectivity analysis test is used to assess the suggested lexicon's coverage, and the sentiment annotated dataset is assessed by using SVM, LSTM, BiLSTM, and CNN models.

Fig. 4 is a graphical illustration of the various stages involved in sentiment analysis from survey findings.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

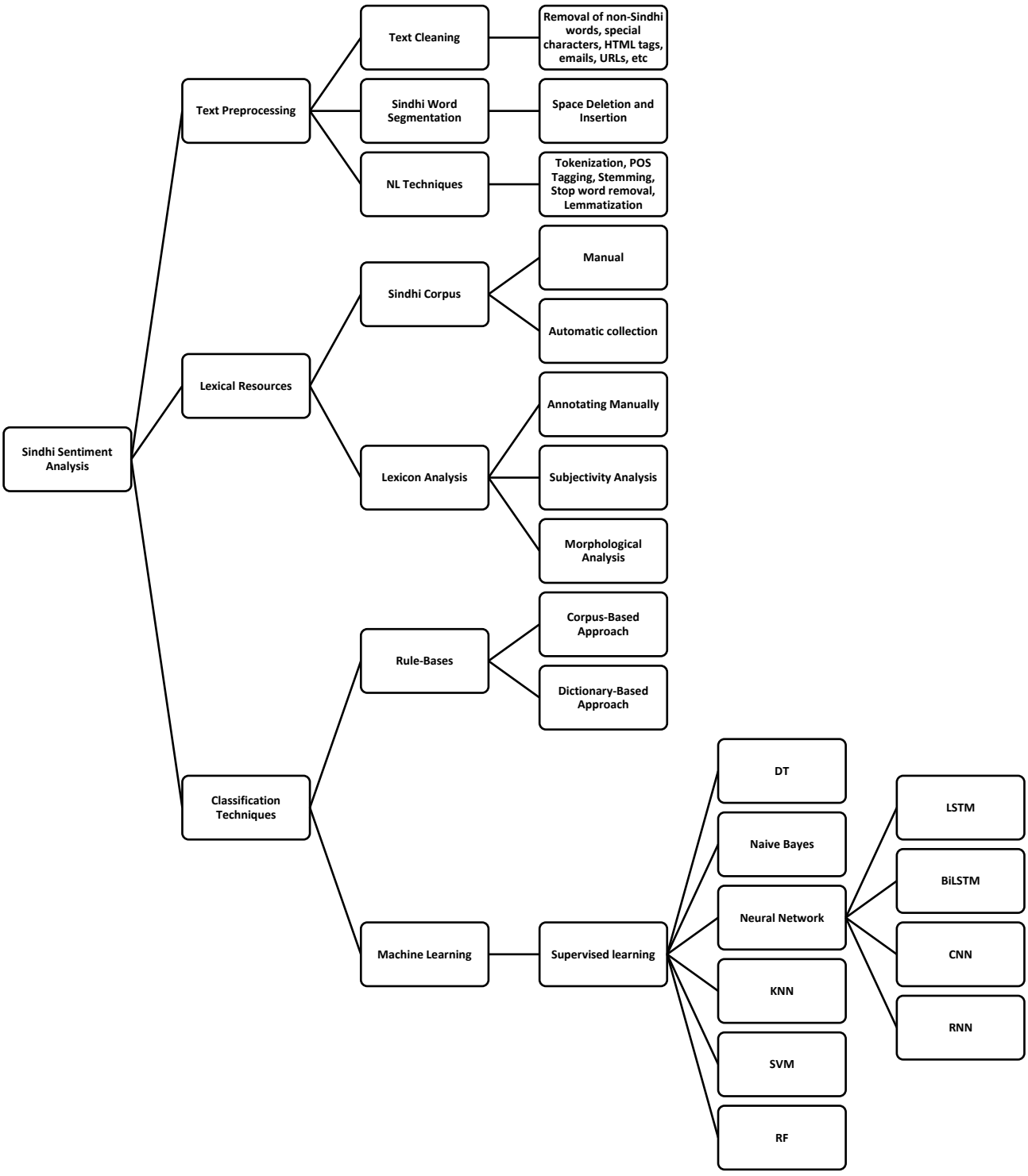

Figure 4. Graphical Illustration for the Sentiment Analysis Processes from Selected Research Articles

RQ3. What exactly does sentiment categorization entail? How do different categorization methods assess Sindhi text sentiments in the chosen articles?

SA, as the term indicates, is the process of determining the point of view or opinion underlying a scenario. It simply involves analyzing and determining the emotion or purpose behind some

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

kind of writing, speech, or any other form of communication. As humans, everyone interacts in a wide range of languages, and each language is simply a medium through which every person attempts to reveal ourselves. But all each one say is tinged with emotions. It may be good or bad, or it could be neutral.

Assume there is a restaurant chain that sells a range of various food products. They have built a website to offer their meal, and clients may now purchase any type of food via the website, in addition to offer feedback, such as whether they liked or disliked the meal.

NLP (natural language processing) interprets subjective data, which facilitates in understanding consumer sentiments about any item, services, or branding. The fundamental purpose of NLP is to solve problems by employing computer methods and semantics to translate person-provided content into such a structure understood through machine.

There are two techniques of classifying sentiments involves in the selected articles such as:

- Rule-based
- Machine Learning

Rule-Based Technique:

This method includes a set of rules used for all categories. A rule-based technique focuses on each language's vocabulary, which contains both positive and negative terms. The polarity of any document is found by counting the amount of positive and negative terms in the phrase. A sentence is considered positive if it includes more positive terms than negative terms [19,27]. Unfortunately, this method has numerous limitations, including the inability to add new terms and determine the polarity of complicated type phrases.

The rule-based approach has been further divided into two categories such as corpus-based and dictionary-based.

Corpus-Based Technique

This is a data-driven method that not only retrieves opinions via tags, but also tries to make use of information in machine learning techniques. When employing a large corpus, a corpus-based technique employs feed words connected to opinion to locate more opinionated terms. Such technique is classified into two subgroups such as statistical-based and semantic-based. First method is to gather the polarity of terms by number of observations, which is also known as lexical items. Identical sentiment values are assigned to semantically adjacent phrases in a semantic-based method [19,30].

Dictionary-Based Technique

The technique is used to discover polarity at the phrase or text level, either independently or via application like WordNet. Begin by calculating words with indicating sentiments, such as negations, and then calculate the occurrence of these terms. This is considered a basic strategy because each term is gathered manually based on its polarity. Despite the fact that this approach is deemed less efficient, the algorithm's performance is dependent on the number of efforts made for the gathering of terms for the given language.

Machine Learning Technique:

The word "learning" in machine learning refers to the technique through which computers study existing data and acquire new abilities and knowledge from it. Machine learning techniques use algorithms to find patterns in datasets which might contain structured large datasets, unstructured text data, and numeric data.

To forecast attitudes, this method incorporates machine learning techniques. Certain predictions are made using a trained dataset. The machine learning algorithm turns text input into vectors based on the polarity of the phrase and finds a predetermined pattern associated with every variable which is negative, positive, or neutral [19,31]. The system grows knowledgeable as it processes data and begins to make predictions for categorization. By delivering more comprehensive datasets, the system's efficiency increases.

Furthermore, ML techniques are classified as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In this survey the selected research articles only use supervisor learning techniques.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

For SA, this approach employs labelled data, also known as trained datasets. Supervised machine learning employs many strategies, some supervised techniques are used in research articles are as follows:

- Decision Tree Classification
- Naïve Bayes
- K-nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Neural Network-based Classification
    - Long Short-Term Memory (LSTM)
    - Bidirectional Long Short-Term Memory (BiLSTM)
    - Convolutional Neural Network (CNN)
    - Recurrent Neural Network (RNN)

Classification Approaches Used by Authors

Mazhar Ali et al. [24] presented a novel Sindhi annotated corpus in which researchers used supervised classification on it to evaluate the accuracy of conventional ML algorithms for solving Sindhi NLP challenges. ML algorithms are tested and appraised using 10-fold cross verification. The Sindhi labeled corpus is divided into two parts as training and testing. The classifier has been trained using 80% of the training data set. According to the results, the research concludes that the RF machine learning approach outperforms then the SVM non-linear method. Ali et al. [19] developed a Sindhi subjective lexicon that utilized multiple resource and an opinionated annotated corpus, that would be a starting point for further developments. The subjectivity analysis experiment is used to identify statements either objective or subjective in order to examine the scope of the proposed vocabulary. If a sentence includes one or more subjective words, it is categorized as subjective; probably, it is categorized as objective. After that, all categories are integrated for sentence level classification by utilizing SVM, LSTM, BiLSTM, and CNN supervised classifiers. The SVM is the least effective baseline classifier. It has a 67.86% accuracy, 68.00% precision, 69.00% recall, and a 68.00% F1-value. The LSTM network outperforms then SVM network, producing 81.42% precision, 82.59% recall, 81.76% F-value, and 79.83% accuracy. The BiLSTM has the highest F-value (83.11%) and accuracy (82.37%). The CNN network performs extremely similarly to BiLSTM, with precision 83.26%, recall 82.67%, F1-value 82.54%, and 81.68% accuracy. The findings show that the BiLSTM and CNN perform much better than the SVM and LSTM. Nonetheless, the BiLSTM outperforms SVM, LSTM, and CNN models. Mazhar ali et al., [24] proposed Sindhi text corpora to provide NLP specialists and academics with text resources. Sindhi sentiment-based text corpus is created and analysed utilising DTM and TF-IDF models based on the n-gram model's 2-gram approach. Table 2 gives a brief summary of selected studies on SA.

Table 2. Summary of Selected Studies on Sindhi Sentiment Analysis

Study No	Study	Objective(s)	Techniques Utilized	Dataset(s)	Results (%)	Future Work and Limitations
1	Sodhar et al. [22]	Text preprocessing Statistical study of Aspect Based Sentiment Analysis based on 3 types Confidence level, +ve and -ve polarity	Text Cleaning, Tokenization Sindhi CL Tool	5 sentences and 152 words from Newspaper website	97.41	Use Sindhi CL tool with large dataset of Sindhi Did not mention type of sentences belongs to which category of News
2	Dootio et al.[24]	Sindhi sentiment based corpus developed,	Tokenization, UPOS tagging, lemmatization, stemming,	15788 total document in which 11864 +ve polarity documents and 3924	-	study is the fundamental research study on the generation and analysis of Sindhi text corpus using

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

		preprocessing applied	aspect based sentiment analysis, sentiment analysis, morphological analysis Corpus analyzed by using Document term matrix (DTM) and TF-IDF using n-gram model	-ve polarity documents. Total number of documents contains 23728 +ve polarity words and 7848 -ve polarity words		Arabic-Persia script; nevertheless, additional research work is needed for Sindhi text corpus analysis using Word2vec, term similarity analysis and so on.
3	Ali et al. [32]	Developed SiPOS dataset for tagging, SiPOS annotated by 2 experienced native annotators, character-level representation using Neural network supervised machine learning technique	Text cleaning, Tokenization, manually annotation using Doccano text annotation tool, CRF and BiLSTM encoder	293K Tokens	96.25	For Sindhi text categorization, researchers aim to pre-train transformers such as (BERT and GPT) languag e models. For the performance boost, they use CRF as the initial baseline and the BiLSTM model by adding CRF as a decoder and self-attention mechanism.
4	Ali et al.[21]	Developed Sindhi annotated corpus using 'UPOS' tag set and 'SPOS' tag set Assess annotated corpus using supervised machine learning model	TF-IDF using N-gram, SVM non-linear and RF, confusion matrices, accuracy, precision, recall and F1-score	-	SVM non-linear (89.16 and 89.1) RF (99.57 and 99.89)	Apply some other supervised classifier for assess model, Sindhi annotated corpus consists 6 attributes
5	Hammad et al.[23]	Developed dataset of Sindhi language tweets and assigning polarity to tweets, Preprocessed the dataset,	Automated approach using predefines lexicons, Text cleaning and Tokenization applied on dataset, SVM, Naïve Bayes, DT, K-NN by	4236 tweets	69, 65.4, 71.2, 70.3	Further study is needed to analyse the semantics and sentiments of the Sindhi language with large number of tweets. One of the study's key weaknesses was the use of an automated approach to identify the dataset,

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

		Assessed by using supervised machine learning techniques	precision recall and F-score			which was not 100% correct.
6	Dootio et al.[33]	Developed a dataset on grammatical and morphological structure of Sindhi language text, besides that developed a Sindhi lexicon with sentiment polarity	Unicode-8 based dataset as a tool for Sindhi linguistics	6841 categorical records with 19 attributes	-	The data set was created using the outputs of a Sindhi online natural language processing (NLP) tool that was used for parsing, tagging, morphological and SA, stemming, and lemmatization of Sindhi text.
7	Surahio et al.[20]	Developed corpus for Sindhi POS using supervised machine learning technique	Text cleaning, manually annotated corpora, SVM using precision recall and f-score	28000 words	97.86	Only one classifier used. Furthermore, no information found related to classes of tag set used in this study.
8	Ali et al.[34]	Developed Sindhi NLP toolkit, Sindhi sentimentally structured and analyzed corpus assessed by supervised machine learning techniques	DTM and TF-IDF using n-gram model, SVMs and K-NN by using precision recall and f-score	9779 records	Average of all classes f-score of SVM is 62.5 Average of all classes f-score of K-NN is 55.5	Sentiment structurization has addressed the challenge of linguistic SA for opinion tracking and sentiment summarization. There is a greater need to concentrate on Sindhi SA for feature extraction so that businesses may get accurate thoughts and ratings on their items.
9	Ali et al.[19]	Sindhi subjective lexicon, annotated corpus assessed by supervised machine learning	Merge NRC lexicon and Bing Liu's lexicon, Polarity assigned by SentiWordNet 3.0 translated Sindhi using bilingual dictionary, Corpus annotated manually, SVM, LSTM, BiLSTM, CNN ,	173 Sindhi modifiers	83.11 and 82.37	Subjectivity analysis experiment is used to identify statements as subjective in order to examine the coverage of the proposed vocab. The suggested vocab may be extended in the future by employing a corpus-based technique to collect language-specific terms.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

			F-score and Accuracy			
10	Mahar et al.[27]	Developed Sindhi corpus, Rule based Sindhi POS using supervised approach	Data collect from comprehensive Sindhi dictionary, Tokenization, tagging, lexicon of Sindhi words, linguistic rules, manually tag the tagset	26366 tagged words which includes (15091 noun, 137 pronoun, 4656 verb, 5328 adjective, 979 adverb, 98 preposition, 18 conjunction,59interjection)	96.28	This study's future work will focus on statistical methods to SPOS and then compare the outcomes to a rule-based approach. When researchers evaluated poetry material and phrases in the future tense, they discovered that SPOS's accuracy was poor.

Hammad et al. [23] developed a dataset of tweets in Sindhi language. To assign positive and negative polarity to tweets, an automated technique based on a predetermined vocabulary is utilized. Pre-processing include removing non-Sindhi words, unnecessary white spaces, and punctuation, followed by text tokenization. Researchers used SVM, NB, DT, and K-NN supervised machine learning algorithms and for assess the algorithms researchers used precision, recall, and f-score. A total of 4236 tweets were pre-processed and tokenized. Researchers created four distinct dataset sizes, 500, 1500, 3000, and 4236 tweets, to test how algorithms perform on various dataset sizes. The accuracy grows as the number of tweets increases, with a 66% accuracy on the whole dataset. Precision is consistent across the sample, peaking at 83.9% on 500 tweets. With more data, recall remained about 70% and f-score improved. The mean accuracy is 64.6%. As contrasted with other algorithms, Nave Bayes performed poorly on any number of tweets. On the maximum number of tweets, the decision tree performed well, with accuracy at 68%, precision at 70%, recall at 72%, and f-score at 71%. The KNN (K-nearest neighbour) worked well on Sindhi dataset offering over 70% score on precision, recall and f-score though giving 67% accuracy. The findings shown that Decision Tree (DT) and (KNN) k-nearest neighbour worked well on the Sindhi tweets dataset. Surahio et al. [20] demonstrated a POS tagger that is entirely focused on Sindhi text. A number of 28000 words have been gathered in the corpus, which includes poetic content derived from primary text books, newspapers, and narratives. SVM is used to tag Sindhi language phrases. Their proposed tagging technique achieved 97.86% accuracy for ambiguous and unidentified tagged terms. Mahar et al. [27] developed a rule-based POS tagger for Sindhi language. Algorithms for POS tagging and tokenization were created and implemented. Sindhi spelling is complicated owing to the lack of diacritic characters. As a result, the supervised technique was utilized to create a SPOS tagging system. SPOS delivered an accuracy of 96.28%. Throughout the studies, it was discovered that SPOS's accuracy was poor when examined poetry material and phrases in the future tense. Similarly, when present and past tenses were examined, the accuracy was quite high. Sindhi linguists conducted the validation assessment. Researchers were completely happy with SPOS's outcomes. A few Sindhi POS tagging issues were resolved as a result of this enhancement. Several papers on Sindhi and other Indo-Pak local languages have been written in recent years to analyze opinions in those languages and examine people's perspectives to support decision making.

Challenges And Gaps

In this study researchers find very few works done on Sindhi text as indicating in the findings. Researchers found only one dataset [21] for product-based SA. The majority of publications considered for this study mention an acute lack of datasets to conduct research on Sindhi Text corpus [19,20,22,24,27,34]. SA for Sindhi text, as for many other languages, unveils multiple challenges and gaps due to linguistics peculiarities, availability of data and cultural variations. Some instances as follows:

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

- Data availability is one of the main issues such as labeled Sindhi corpus or dataset for SA. Limited dataset makes it challenging to develop reliable and consistent SA models for Sidhi text.
- Another challenge is the language complexity. Sindhi is a sophisticated and complicated language with a specific alphabet, grammar, and syntax. Its morphological richness, which includes accent and compound words, affects sentiment detection techniques, that requires advance research in linguistics. Sindhi speakers frequently use words or phrases from other language such as English, Urdu and Arabic. SA algorithms must be able to handle bilingual material in order to appropriately assess sentiment.
- Sindhi, like any other language, has slang, idioms, and informal phrases that may not be directly translated or interpreted. SA algorithms must be directed to identify and evaluate such statements in context.
- SA accuracy varies across domains, such as news articles, social media and product evaluations. Developing domain driven SA models for Sindhi text needs domain driven datasets or corpus which is lacking. As well as lacking of language specific sentiment lexicons, pre trained models, and annotation tools for building significant SA models.
- To address these gaps, linguistics, data scientists and researchers must work together to develop and collect datasets, build language specific models, and tools specifically designed for Sindhi text SA.

Results and Discussion

Nowadays, individuals often prefer to read reviews before making purchases or sales. Such type of data is hugely available on social media. Most likely, individuals around the world likes to read reviews regarding to products, food items, places to visited, restaurants, games, entertainment places, workout places, sports, sports clubs, health tips, hospitals, and government policies. When individuals want to make a decision, they often rely on reviews obtained from social media or other authentic online sources. As well as government, organizations and businesses need to know about the public and consumer opinion about their services and products. Therefore, most of the work is done on the pre-processing techniques, Sindhi parts of speech, lexicon, universal parts of speech for Sindhi text. Additionally, efforts have been made to create Sindhi corpus text for SA, supporting NLP researchers for further contribution in the NLP field. There is a significant gap in Sindhi NLP field to reach at least the level of recognition that other similarly written languages such as Arabic and Urdu receive around the globe.

Conclusion

In light of the Internet's global spread and persons numerous pieces of opinions regarding different online products and events. It has evolved into an unavoidable requirement for businesses or governments to take into account online opinions given in any locally used language and manage those certain opinions for insuring decisions, to enhance the value as well as the norms of the brands. The major purpose of this article was to focus on numerous advanced approaches utilized for analyzing text, pre-processing contents, feature-extraction, lexicon information, and Sindhi classification algorithms. This research examined the Sindhi sentiments issues and went into depth on pre-processing, feature extraction, lexicon inputs, and sentiment algorithms. For different kinds of datasets, many algorithms were discussed in terms of their accuracy, precision, recall, and F-score. This study reviewed the limitations of all previous works. Almost all of the Sindhi research listed above employed a lexicon-based technique or ML to determine sentence or phrase polarity. It is suggested that future research looks at a system that uses a mix of machine and deep learning approaches to deliver better results.

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Future Work

The aim of this study is to investigate that the machine learning and deep learning classifiers performs better for sentiments of product reviews that may be performs better for Sindhi newspaper headlines.

Acknowledgment

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at Universiti Teknologi Malaysia.

Authors' Contribution Statement

S. A. S, S. S. Y., and M. A. D. designed and develop the Idea. S. A. S., S. S. Y., and M. A. D. designed the methodology. S. A. S., G. M., and M. H. M. performed search queries, acquisition of data and analysis. S. A. S. writing original draft preparation. S. A. S. writing review and editing. S. S. Y. and M. A. D. reviewed, proofread and critically analyzed the manuscript. All authors read and agreed to the published version of the manuscript.

References

1. Al-Bakri NF, Yonan JF, Sadiq AT, Abid AS. Tourism Companies Assessment via Social Media using Sentiment Analysis. Baghdad Sci J. 2022; 19(2): 422–9. https://doi.org/10.21123/BSJ.2022.19.2.0422
2. Al-Jumaili ASA, Tayyeh HK. A Hybrid Method of Linguistic and Statistical Features for Arabic Sentiment Analysis. Baghdad Sci J. 2020; 17(1): 385-390. https://dx.doi.org/10.21123/bsj.2020.17.1(Suppl.).0385
3. Mutasher WG, Aljuboori AF. New and Existing Approaches Reviewing of Big Data Analysis with Hadoop Tools. Baghdad Sci J. 2022; 19(4): 887–98. https://doi.org/ 10.21123/bsj.2022.19.4.0887
4. Zaki UHH, Ibrahim R, Abd-Halim S, Kamsani II. Prioritize Text Detergent: Comparing Two Judgement Scales of Analytic Hierarchy Process on Prioritizing Pre-Processing Techniques on Social Media Sentiment Analysis. Baghdad Sci J. 2024; 21(2): 0662-0683. https://doi.org/10.21123/bsj.2024.9750
5. Motlani R. Developing language technology tools and resources for a resource-poor language: Sindhi. In Proceedings of the NAACL Student Research Workshop, 2016; 51–58. https://doi.org/10.18653/v1/N16-2008
6. Mukherjee S. Sindhi language and its history. L D, Kolkata, 2018.
7. Jamro WA. Sindhi Language Processing: A Survey. Conference: 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies. (ICIEECT), 2017. https://doi.org/10.1109/ICIEECT.2017.7916560
8. Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Springer, 2015. https://doi.org/10.1007/s12559-014-9316-6
9. Bhadane C, Dalal H, Doshi H. Sentiment analysis: measuring opinions. Procedia Comput Sci, 2015; 45: 808-814. https://doi.org/10.1016/j.procs.2015.03.159
10. de Albornoz JC, Plaza L, Gervás P. SentiSense: An easily scalable concept-based affective lexicon for

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

sentiment analysis. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), 2012.
11. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, 2011.
12. Tripathy A, Agrawal A, Rath SK. Classification of sentiment reviews using n-gram machine learning approach. Expert Sys Appl. 2016; 117-126. https://doi.org/10.1016/j.eswa.2016.03.028
13. Peng H, Cambria E, Hussain A. A Review of Sentiment Analysis Research in Chinese Language. Springer, Aug. 2017; 9(4): 423–435. https://doi.org/10.1007/s12559-017-9470-8
14. Manek AS, Shenoy PD, Mohan MC, Venugopal KR. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 2017; 135-154. https://doi.org/10.1007/s11280-015-0381-x
15. Erra U, Senatore S, Minnella F, Caggianese G. Approximate TF–IDF based on topic extraction from massive message stream using the GPU. Inf Sci. Jan. 2015; 292: 143–161. https://doi.org/10.1016/j.ins.2014.08.062
16. Nazir S, Nawaz M, Adnan A, Shahzad S, Asadi S. Big data features, applications, and analytics in cardiology—A systematic literature review. IEEE Access. 2019; 7: 143742–143771. https://doi.org/10.1109/ACCESS.2019.2941898
17. Nazir S, Shahzad S, Mukhtar N. Software birthmark design and estimation: A systematic literature review. Arab J Sci Eng. 2019; 44: 3905–3927. https://doi.org/10.1007/s13369-019-03718-9
18. Keele S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Version 2.3, EBSE Technical Report, Keele University and Durham University Joint Report; EBSE: Keele, UK, 2007; p. 1–57.
19. Ali W, Ali N, Dai Y, Kumar J, Tumrani S, Xu Z. Creating and Evaluating Resources for Sentiment Analysis in the Low-resource Language: Sindhi. Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. April 19, 2021; 188–194.
20. Surahio FA, Mahar JA. Prediction System for Sindhi Parts of Speech Tags by Using Support Vector Machine. International Conference on Computing, Mathematics and Engineering Technologies. iCoMET, 2018. https://doi.org/10.1109/ICOMET.2018.8346331
21. Ali M, Wagan AI. AN Analysis of Annotated Corpus using Supervised Machine Learning Methods. Mehran Uni Res J Eng Technol. Jan 2019; 38(1): 185-196. https://doi.org/10.22581/muet1982.1901.15
22. Sodhar IN, Sulaiman S, Buller AH, Sodhar AN. Aspect-Based Sentiment Analysis of Sindhi Newspaper Articles. Int J Comput Netw Secur. May 2022; 22(5). https://doi.org/10.22937/IJCSNS.2022.22.5.54
23. Hammad M, Anwar H. Sentiment Analysis of Sindhi Tweets Dataset using Supervised Machine Learning Techniques. 22nd Int Multitopic Conf. (INMIC), 2019. https://doi.org/10.1109/INMIC48123.2019.9022770
24. Dootio MA, Wagan AI. Development of Sindhi Text Corpus. J King Saud Univ – Comput Inf Sci. 33, 2021; 468–475. https://doi.org/10.1016/j.jksuci.2019.02.002
25. Mahar JA, Memon GQ, Danwar SH. Algorithms for Sindhi Word Segmentation Using Lexicon-Driven Approach. Int J A R. 2011; 3(3).
26. Narejo WA, Mahar JA, Mahar SA, Surahio FA, Jumani AK. Sindhi Morphological Analysis: An Algorithm for Sindhi Word Segmentation into Morphemes. Int J Comput. Sci Inf. Sec. (IJCSIS).14, June 2016; 14(6):293-302.
27. Mahar JA, Memon GQ. Rule Based Part of Speech Tagging of Sindhi Language. Int Conf. Signal Acquisition and Processing, 2010. https://doi.org/10.1109/ICSAP.2010.27
28. Al-Jumaili ASA, Tayyeh HK. A Hybrid Method of Linguistic and Statistical Features for Arabic Sentiment Analysis. Baghdad Sci J 2020, 17(1): 385-390. https://dx.doi.org/10.21123/bsj.2020.17.1(Suppl.).0385
29. Noureen, Huspi SH, Ali Z. Sentiment Analysis on Roman Urdu Students' Feedback Using Enhanced Word Embedding Technique. Baghdad Sci J. 2024, 21(2): 0725-0739 https://doi.org/10.21123/bsj.2024.9822
30. Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining. Int J Sci Res., 2016; 5: 2094–2097.
31. Yang H, Fong S. Optimized very fast decision tree with balanced classification accuracy and compact tree size. In Proceedings of the 3rd Int Conf. on Data Mining and Intelligent Inf Tech Appl., Vienna, Austria, 29–31 August 2014; 57–64.
32. Ali W, Xu Z, Kumar J. SiPOS: A Benchmark Dataset for Sindhi Part-of-Speech Tagging. Proceedings of the Student Research Workshop associated with RANLP- Sep 1-3, 2021; 22–30.
33. Dootio MA, Wagan AI. Unicode-8 based linguistics data set of annotated Sindhi text. Data in Brief, 2018;

Published Online First: November, 2024
https://doi.org/10.21123/bsj.2024.10954
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

19: 1504–1514. https://doi.org/10.1016/j.dib.2018.05.062
34. Ali M, Wagan AI. Sentiment Summarization and Analysis of Sindhi Text. (IJACSA) Int J Adv Comput Sci Appl. 2017; 8(10): 296-300. https://doi.org/10.14569/ijacsa.2017.081038

مراجعة منهجية لتحليل المشاعر للنص السندي

صفدار علي سومرو[1]، ستي صوفياتي يوهانيز[1]، مظهر علي دوتيو[2]، غلام مرتضى[3]، محمد حسين موغال[3]

[1]كلية رزاك للتكنولوجيا والمعلوماتية، الجامعة التكنولوجية الماليزية، كوالالمبور، ماليزيا.
[2]قسم علوم الحاسوب، جامعة بينزير بوتو شهيد، كراتشي، باكستان.
[3]قسم علوم الحاسوب، جامعة سوكور IBA، سوكور، باكستان.

الخلاصة

نظرًا لتطبيقه في مجالات مثل عناوين الأخبار، وشراء المنتجات عبر الإنترنت، والتسويق، وإدارة السمعة، فقد ارتفعت أنشطة رفع الوعي في مجال استخراج الرأي بشكل ملحوظ .أصبحت مدونات الإنترنت والمواقع الاجتماعية ومتاجر التسوق الإلكترونية مرجعًا مهمًا للمعلومات التي ينتجها المستخدم .تتطلع شركات التصنيع والمبيعات والتسويق بشكل متزايد إلى هذا المورد للحصول على تعليقات عالمية حول ممارساتها وعناصرها .تتم مشاركة ملايين العبارات السندية يوميًا على مواقع الوسائط الإخبارية و Twitter و Facebookو Snapchat. إن تجاهل آراء الناس في اللغة السندية والتركيز فقط على اللغات الغنية بالموارد في العالم الغربي يؤدي إلى خسارة فادحة لهذه الكمية الكبيرة من البيانات .تركز هذه الدراسة على جمع وتقييم المنشورات المرتبطة باللغة السندية استجابة لمناهج التصنيف واستخراج الميزات والمعالجة المسبقة .تقدم الدراسة الحالية فحصًا شاملاً للعمل المنجز على كلمات اللغة السندية للعناصر أو تقييم العلامة التجارية .تركز الدراسة الحالية على الاستحواذ القائم على المجموعة، وتقنيات التصنيف، واستخراج الميزات، والمعالجة المسبقة للبيانات، والمنهجيات، والقيود .تم تقييم كل مقالة تمت مراجعتها وتصنيفها على أساس معايير معينة محددة .وبناء على النتائج، سوف تقترح هذه الدراسة عدة طرق مفيدة للتحقيق في المستقبل.

الكلمات المفتاحية: معالجة اللغات الطبيعية، تحليل المشاعر، مجموعة النصوص السندية، النص السندي، المراجعة المنهجية، المعالجة المسبقة للنص.