Stemming in Different Languages: A Survey

Register Number	Name
20BCT0350	Mohd Anas
20BDS0148	Hardeep Singh
20BDS0068	Devendra Kumar
20BDS0053	Abhishek Krishna

Corresponding Author: **Prof. Saravanakumar Kandasamy**

1. Abstract

Stemming is a method utilized to decrease words to their root frame called stem, by removing derivational and inflectional affixes. It progresses the execution of data recovery frameworks by diminishing the index size. It is a rule-based approach because it slices the inflected words from prefix or suffix as per the need using a set of commonly underused prefix and suffix, like “-ing”, “-ed”, “-es”, “-pre”, etc. It helps to understand the text, enabling machines to recognize how a particular human talk in various languages concerning the place they reside. stemming algorithms are used in a much broader way to get easier with the information retrieval system. Suffixes and affixes are usually striped of from the words to get some information but sometime over-stemming occurs in between the algorithms and due to which accuracy and other parameters faces a quick drop down in their values. A root algorithm based on Khoja's work, a light stemming (LS) method, a Machine Translation (MT) based stemmer, and many more Arabic light stemmer techniques and algorithms have previously been presented. This strategy has wide application in NLP, Content mining and data retrieval. The stemming calculations have been created using either a language-specific approach, which needs earlier information of a language's morphology, or factual strategies, which are based on probabilistic concepts to find morphological variations of a language from a collection of archives. Most of the existing stemming calculations uses fasten stripping technique. Stemming has long been used in data pre-processing to retrieve information by tracking affix words back into their root like in an Indonesian setting, existing stemming methods have been observed, and are proven to result in high accuracy level. However, there are not many stemming methods for non-formal Indonesian text processing. This study introduces a new stemming method to solve problems in the non-formal Indonesian text data pre-processing. Furthermore, this study aims to improve the accuracy of text classifier models by strengthening stemming method. The aim is also to overcome the limitation problem by decreasing word forms into their stems. This approach is fully based on unsupervised methods in agglutinative languages with some experiments on non-agglutinative languages. We proposed a fully unsupervised language-independent text stemming technique that clusters morphologically related words from the corpus of the language using both lexical and co-occurrence features such as lexical similarity, suffix knowledge. An unsupervised stemming approach to discovers potential suffixes from the ambient corpus and uses them in the suffix stripping process to obtain the stem of the word.Building corpora for low-level languages, updating them, and employing alternative platforms, as well as machine learning methods for text processing, can help scale the effort. The impact of using the corpus as a stemming method is that it can improve the accuracy of the classifier model. The major objective is to minimize all morphological variants to the word's base form. Morphological processing of inflectional suffixes in French verbs independently of the root and stem processing, addressing the abstract morphosyntactic features activated by the tense and agreement suffixes in the hierarchical morphological structure. The main challenges such as transparency of words as words are mainly composed by three letters, orthographic variations, a very good morphology, for Arabic information retrieval system. In addition to that, it also gives a brief idea about concepts such as Lemmatization and Stemming. It helps in understanding their working, the algorithms that come under these processes, and their applications. Stemming improves the ability of an IR system to retrieve more relevant documents by solving the problem of vocabulary mismatch in queries and documents at the time of indexing and searching. For the coming future, there are some sort of limitations i.e., under stemming and over stemming which needs to be covered and in future some other unsupervised algorithm can also be used in order to improve accuracy by applying some lemmatization rules, as we have seen lemmatizer has given better yield as compared to others.

2. Introduction

Word stemming is a broadly utilized mechanism in Information Recovery (IR) and Natural Language Processing (NLP) frameworks to convert the morphological variation of word shapes to their base shapes. For illustration, the variation word shapes abandon, abandoned, abandoning, abandonment, abandonments, etc. are changed to their base shape desert through stemming. It can be characterized as a handle that extracts stem or root words from a given arched word. The method decreases assorted syntactic word shapes to their root, stem, or base shape but most imperatively it is connected to the Information Recovery Framework. The key guideline is that in Information Recovery frameworks, words with comparable meanings and morphologies ought to be treated as proportionate. As a result, the major objective is to play down all morphological variations to the word's base frame. Stemming oversees words that might show up in numerous morphological shapes, and so matches the terms of records and questions that have comparative implications. The key concept behind our proposed approach is to identify just those versions of words from the ambient corpus that fit the original intent of the query terms. Indeed, the work is not either widely accessible or well-explored for Indonesian language. There have been no comparative studies that look at the relative efficacy of different stemming procedures for this language. An improved stemmer has not been developed to address problems in the graphene sequence. We considered stripping suffixes and all possible suffixes were manually collected taking the help of an Assamese linguistic expert. The suffixes that are collected are divided into eight different suffixes such as plural words, case words, definitive words, pleo words, extra words, in-definitive words, verbs, and kinships.

The stemming algorithms were created using either a language-specific approach, which necessitates prior knowledge of a language's morphology, or statistical methods, which are based on probabilistic or statistical concepts and are used to discover morphological variants of a language from a collection of documents. In this paper, stemming related problem while working on highly inflected language has been discussed. Morphologically rich languages give some sort of issues while under stemming processors. So, this problem has been tried to solve in this paper. The primary goal of our proposed research is to develop a completely unsupervised, language independent, cognitively based corpus-based stemming algorithm that can perform stemming without any language-specific information or human intervention and can function similarly to the human brain. Proposed method comprises three steps: The proposed work's first step is to estimate structural distance and lexical distance between the word pair. The distance function is responsible for mapping the word-pairs. Then removing the stop words. With the help of this unsupervised stemmer, we can normalize inflected words of Sindhi language. These normalized word forms can be used in various applications of Natural Language Processing NLP such as in Information Retrieval, Text summarization, Machine translation and Topic Identification. Morphological awareness improves with age and in each subsequent grade is more predictive of both reading and spelling achievement in children exposed to different orthographies. Morphological awareness usually predicts unique variance in addition to phonological awareness and has different degrees of association with word recognition (and spelling) in different scripts. The ability to discern morphological ties between the phonological forms of the word in SpA and StA might help prevent readers from becoming distracted by differences in phonetic form and in so doing, compensate for and alleviate the disruptive impact of phonological distance. The typological regularity in morpheme order we target here concerns number and case morphology, specifically, how these two classes of morphemes are ordered when there is a clear morphological boundary between them.

Root based stemmers use morphological analysis to extract the root of a given Arabic word. The solution of the above problem will be to propose new lists of suffixes and prefixes. The lists will be used by the proposed algorithms that use word length rules to remove suffixes and prefixes from target words. There are wide range of morphological variants formed through different linguistic processes such as affixation, compounding, conversion, etc. Getting a problem of vocabulary mismatch in queries and documents at the time of indexing and searching. due to the morphological variants retrieval accuracy not getting improved. The problem which is answered in this paper is regarding Part-of-speech PoS tagging which hold up in important application like sentiment analysis, question answering, text summarization, and machine translation.

In today's world, natural language processing (or computational linguistics) is becoming the state of the art. It has evolved over many years, beginning in the 1960s. The goal of NLP is to interpret real human utterances in terms of voice or text, take them as input, and respond appropriately. Natural language processing is used in Text Analytics to convert unstructured corpora into standard and standardised documents, or databases, for further analysis using Artificial Intelligence methods and Machine Learning algorithms. Various semantic and grammatical norms for defining language structure, lexical and syntactical constructions of multilingual came into force in the mid-1970s to 1990s, resulting in significant and effective improvements in NLP. The first stemmer was developed by Ramanathan and Rao in 2003. In this approach, they have manually extracted the suffix list and used a matching approach for finding of an inflected word. After that, a Derivational stemmer for Guajarati language was proposed by Suba et. al. by using a rule- based approach and achieved an accuracy of 70.7%. They have also worked on Inflectional Stemmer by using hybrid approach with 90.7% accuracy. Some research on Indonesian text processing had been done by using the existing Indonesian stemmer, i.e., “Sastrawi”. The research and improvements of “Sastrawi” algorithm were also done by using dynamic affixes to process non-formal Indonesian. However, the result of their research had limitations on non-formal Indonesian expressions, which were formed into abbreviated and deformed words.

Stemming algorithms can be seen under two categories, first one is rule based and second one is statistical approaches. Lovin and Porter stemmers are defined as two previous rule-based stemmers which usually targeted towards suffix removal of English language. Suffix stripping can also be applied to different languages but beware of over stemming. But this problem can be sort out using a predefined lexicon of stems which will avoid over-stemming while performing stemming over a highly inflected language. One of the proposed methods is known as 'Morfessor' which works on MDL-based morphology induction tool and has also given good results in English, Finnish and in Turkish datasets. So, after this, a new version of 'Morfessor' is proposed in this paper. This paper proposed a new method for evaluation of long texts and short texts based on LDA and DMM, respectively evaluate CmTLB against the studied weighting schemes by the quality of the learned topics (topic visualization and topic coherence), classification, and clustering tasks. Gibb’s sampling is used to train all models sparsity of short texts. use of CmTLB outperform the baseline models in most cases for both long and short document datasets. For example, the improvement over LDA-CEW is about 1%–6%. It is noticeable that topics learned by DMM-CmTLB are cleaner and more coherent. stemming process significantly improves the performance of topic modelling and that Farasa stemmer outperforms the other stemmers with statistically significant enhancements. A smoothing algorithm is performed to eliminate local minima because the main problem in using this method, though, is that the difference identified between cohesion scores can be, at times, caused by local minima that do not correspond to a topic boundary.

3) Keywords

Assamese – An Official language of Assam.

Corpus - A collection of authentic text organized into datasets.

Suffixes - A suffix is a letter or group of letters, for example '-ly' or '- ness', which is added to the end of a word in order to form a different word.

WSD - Word-sense disambiguation is an open problem in computational linguistics concerned with identifying which sense of a word is used in a sentence.

Text Tiling - A technique for subdividing texts into multi-paragraph units that represent passages, or subtopics.

Morphology - The study of words, how they are formed, and their relationship to other words in the same language.

Linearisation - A mathematical process of finding the linear approximation of inputs and corresponding outputs.

Typology - The process of describing the various linguistic types found across languages for some grammatical parameter, such as grammatical number or the formation of relative clauses.

Phonotactic Rules - Rules and restrictions concerning the ways in which syllables can be created in a language.

LDA - Latent Dirichlet allocation is used to classify text in a document to a particular topic.

Infixes - An affix inserted inside a word stem.

Text Mining - Text mining (also referred to as text analytics) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

Inflection - A process of word formation in which a word is modified to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, mood, animacy, and definiteness.

NLTK - The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

Lexicon - A lexicon is the vocabulary of a language or branch of knowledge.

Orthographic - An orthography is a set of conventions for writing a language, including norms of spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation.

Unsupervised - It is a type of algorithm that learns patterns from untagged data.

Agglutinative - It is a type of synthetic language with morphology that primarily uses agglutination.

Lemmatizer - Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

Smoothing – The term smoothing refers to the adjustment of the maxi- mum likelihood estimator of a language model so that it will be more accurate.

Semantic – A subfield of Natural Language Processing (NLP) that attempts to understand the meaning of Natural Language.

Arabic Diglossia – In a strict definition, is distinct in that the "high" version of a language isn't used for ordinary conversation and has no native speakers.

SVM – A supervised machine learning algorithm that can be used for both classification or regression challenges.

G2P – Grapheme-to-phoneme conversion. This is the process of using rules to generate a pronunciation for a word (for creating a pronunciation dictionary).

MDL – Minimum Description Length provides a criterion for the selection of models, regardless of their complexity, without the restrictive assumption that the data form a sample from a 'true' distribution.

LUCENE – It Provides a Java-based search and indexing platform.

FARASA – The state-of-the-art full-stack package to deal with Arabic Language

HMM – Hidden Markov Model (HMM) is a statistical model which is also used in machine learning. It can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable.

ANLP – The Association for Neuro-Linguistic Programming (ANLP) is a UK organisation founded in 1985 by Frank Kevlin to promote neuro-linguistic programming (NLP)

BCPL – BCPL ("Basic Combined Programming Language") is a procedural, imperative, and structured programming language

MLE – Maximum Likelihood Estimate, it is the choice of parameters that gives the highest probability to the training corpus.

4. Related work

Stemming is basically an important part of Part-of-Speech (PoS) tagging ,an application in natural language processing (NLP). Word stemming is a linguistic process in which the various inflected word forms are matched to their base form. Stemming is the basic thing which is done before proceeding with the text data in machine learning section. In today's world, there are a number of stemmers for different types of languages like Urdu, Arabic, Assamese, Sindhi and many more. A rule-based Comprehensive stemming approach for Urdu text has been proposed. Use data mining techniques to extract relevant information from such a massive amount of potentially interesting base data [1]. In a paper, Suyanto et al. [2] researchers discussed about a Phonemicization, also known as grapheme-to-phoneme conversion (G2P), is a method of translating a word into its pronunciation. The key principle is that in Information Retrieval systems, words with comparable meanings and morphologies should be treated as equivalent. As a result, the major goal is to minimize all morphological variants to the word's base form [3]. Researchers have offered a third stemming method that utilizes lexical resources to validate stems more thoroughly [5]. Stemming words to (usually) remove suffixes has applications in text search, machine translation, document summarization, and text classification [26]. Multilingual Natural Language Processing Applications is the first comprehensive single-source guide to building robust and accurate multilingual NLP systems [37]. Several ANLP tools are developed such as morphological analysers, syntactic parsers, etc. and are characterized by their diversity in terms of development languages used, inputs/outputs manipulated, internal and external representations of results, etc. [44]. Arabic language needs a very specific stemmer which can work around corners in each stemming processor and has discussed all the present approaches in a hope to proposed a better error free stemmer in the process [45]. Mustafa et al. [46] says a different approach to stemming is defined in this paper which is related to two unique stemming techniques based on light stemming techniques. In Kannada language, morpheme boundary detection is a problematic task due to highly inflectional and agglutinative morphology of this language. The algorithm which is used is also trained with highly inflected languages such as Finnish and Bengali. The algorithm is evaluated with a F-measure of 73% [51]. A new stemming algorithm FINDSTEM is proposed to analysis it and evaluate it with the previous present stemmers namely "AF" and "LM" algorithms. FINDSTEM and AF algorithms uses inflectional and derivational stemmers but LM works only on inflectional rules [64]. Yin et al. [69] proposed a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model for short text clustering (abbr. to GSDMM). GSDMM can also cope with the sparse and high-dimensional problem of short texts, and obtain the representative words of each cluster. The problem of high-frequency words can be dealt with more elegantly with the use of appropriate weighting schemes comparable to those sometimes used in Latent Semantic Indexing (LSI) and latent dirichlet allocation [70]. Conventional topic models suffer from a severe sparsity problem when facing extremely short texts such as social media posts. So, Dirichlet multinomial mixture (DMM) can handle the sparsity problem, however, they are still very sensitive to ordinary and noisy words [72]. Paik's et al. [82] algorithm uses the statistics collected on the basis of certain corpus analysis based on the co-occurrence between two-word variants a simple co-occurrence measure that reflects how often a pair of word variants occurs in a document as well as in the whole corpus.One more adopted approach was applied along with a rule engine that is capable of generating all the likely suffix sequences. A list of root-word- sized 20,000 approximately was added in this research which yielded the result with 82% accuracy [88]. Contextual similarity-based method for identifying stems or root forms of Bangla words using N-gram language model and also have fairly gained 40.18% accuracy [90]. Awareness of derivations was also found to distinguish typical readers from poor readers by Abu-Rabia et al [93]. Learners are exposed to a miniature artificial language with nouns, and case (accusative) and number (plural) markers [95]. Whether apparent cross-linguistic differences are important for theory building, or simply reflect methodological differences and/or mere sampling variability [96].

5.1 Information Retrieval (IR): Under Information Retrieval (IR), The automatic removal of suffixes from English words is of particular interest in the field of information retrieval. The algorithm for suffix stripping is described, and it has been implemented as a short, fast programme in BCPL. Despite its simplicity, it outperforms a much more complex system with which it was compared [21]. Many news providers used to share their news headlines on various web sites and web blogs before the development of web blogs and Social Networks. These data may contain a wealth of useful information relevant to a variety of social research fields [22]. A framework for constructing probabilistic models of information retrieval in which the models are nonparametric IR models derived from the language model approach. Term-weighting models are derived by measuring the divergence of the actual term distribution from that obtained by a random process [30]. A method proposes and evaluate a statistical graph-based algorithm for stemming. Considering that a word is formed by a stem (prefix) and a derivation (suffix), the key idea is that strongly interlinked prefixes and suffixes form a community of sub-strings [31]. In an article, Hindi, Bengali, and Marathi languages form an Information Retrieval (IR) perspective through describing the key elements of their inflectional and derivational morphologies, and suggest a light and more aggressive stemming approach [33]. The inflectional structure of a word influences the retrieval accuracy of Latin-based information retrieval systems in which two stemming algorithms for Arabic information retrieval systems are presented [40]. According to Chen et al. [42] an approach was proposed to the cross-language retrieval was to translate the English topics into Arabic using online English-Arabic machine translation systems A model for identifying the verb root made in a tool by root retrieval. This method was able to extract suffixes and prefixes without using any linguistic rules [47]. A new light stemming technique is proposed to improve search effectiveness [48]. A rule-based and an unsupervised Marathi stemmer is also proposed by Majgaonker et al. [49] to solve the problems related to Marathi Internet queries and Information Retrieval. To learn about the impact of stemming techniques, called as Information Science Research Institute (ISRI), Tashaphyne, and ARLStem on Arabic Document Classification DC [54]. There is a necessity to improve Arabic Information Retrieval IR technique using the Stemmer and Lemmatizers, so an stemmer has been proposed to deal with it by Zeroual et al. [58]. Paik et al. [60] proposed an algorithm which is a novel-based language-independent stemming algorithm , well suited for information retrieval .Main features of this algorithm are retrieval effectiveness, generality, and computational efficiency. This algorithm has been tested on a number of languages varies with morphological complexity. After that applying a same modification to query terms, a technique has proposed two unique properties for Web Search. First, based on statistical language modelling , second, this approach performs a context sensitive document [61]. Researchers developed several light stemmers based on heuristics and a statistical stemmer based on co-occurrence for Arabic retrieval and also compared the retrieval effectiveness of our stemmers and of a morphological analyzer on the TREC-2001 data [74]. The proposed solution is to evaluate the performance of a number of the Arabic root extraction methods. Because Root extraction gives valuable support to many natural language processing application such as information retrieval, machine translation.

5.2 Text Classification (TC): According to Krishnakumar [23] , because of the rapid growth of online information, there is an increasing demand for tools that aid in the discovery, filtering, and management of high-dimensional data. Building a text classifier by hand is time consuming and expensive, so automated text categorization has grown in popularity. The process of assigning various input short texts to one or more target categories based on their contents is known as short text classification [24]. For low density languages with little linguistic description, such as many Bantu languages, unsupervised morphological segmentation is appealing , paper proposes a weighted similarity measure that calculates Ordered Weighted Aggregator (OWA) operator weights using the normal distribution [32]. Miner et al. [35] presents comprehensive how-to reference that shows the user how to conduct text mining and statistically analyse results. An application which is proposed here in this paper is finding the likeness between course descriptions of the same subject for credit transfer among various universities or within same academic programs [56]. Three unsupervised models have been used in this paper which were based on word embeddings. Pearson correlation is used evaluation metric with a best performance of 0.6887 [57] .Abuaiadah et al. [71] describe the impact of dataset characteristics on the results of Arabic document classification algorithms using TF-IDF representations. A problem related to Arabic documents classification discussed. Vector Space Model (VSM) is a typical method to describe the text feature in text classification at present. It adopts TF-IDF weights to compute the term weighting in each dimension of the text feature [73]. The effects of the light stemming technique on feature extraction where Bag of Words (BoW) and Term frequency- Inverse Documents (TF-IDF) are employed for Arabic document classification. K-nearest Neighbour (kNN), Support Vector Machine (SVM) classifiers are used to build the classification model

[77]. An algorithm for suffix removal is proposed and GRAS is suitable only for suffixing languages as it clusters morphological variants from the corpus using suffix knowledge [81]. A problem needs improvement of "Sastrawi" algorithm by using dynamic affixes to process non-formal Indonesian which were formed into abbreviated and deformed words, also while using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is checked [83]. Indonesian language in the accuracy of text classification, also getting problem in online conservation, while using flexible Affix Classification improves the accuracy for reduplicated words confix-stripping [84]. Due to lack of accuracy in text classification, there is need of Non-formal Affix Stemming Algorithm proposed by Putra et al. [85], used to get the basic word of a non-formal word. However, this algorithm has limitation in non-formal affixed word stemming. Thus, this research aims to focus on the modification of non-formal affix algorithm to increase accuracy in non-formal affixed words stemming. A new stemming method to solve problems in the non-formal Indonesian text data pre-processing. Furthermore, this study aims to improve the accuracy of text classifier models by strengthening stemming method [86]. Lexical cohesion, which refers to the cohesive effect achieved by the selection of vocabulary by A. El-Shayeb et al. [91]. TopSeg starts with a random initialization of PLSA, which might yield a different error rate of segmentation. Therefore, Brants et al. [92] proposed an improved version of TopSeg named TopSeg-C.

5.3 Neural Network: In this paper, the performance of various sequence to sequence neural networks on the task of grapheme to phoneme (G2P) conversion. G2P is a very important component in applications like text-to-speech, automatic speech recognition etc. according to Achanta et al. [25]. Emiru et al. [27] specifies that an automatic speech recognition is investigated using deep neural network (DNN) acoustic modelling method for Amharic language at syllabic acoustic units. A paper analyses the effectiveness of neural sequence to sequence models in grapheme to phoneme conversion for Myanmar language. The first large Myanmar pronunciation dictionary is introduced, and it is applied in building sequence to sequence models [28]. Recurrent neural networks are extremely appealing for sequence-to-sequence learning tasks, and thus performance suffers when dealing with long sequences, a simple yet effective approach was proposed to overcome this shortcoming [29].

5.4 Part-of-Speech (PoS)Tagging: Quranic Arabic Corpus is an annotated linguistic resource with multiple layers of annotation including part-of-speech tagging, and syntactic analysis using dependency grammar [43]. Alnaied et al. have proposed a different way to extract the stem from Arabic text by applying some sort of rules using the AMIR dictionary [55]. Abstract Arabic Information Retrieval AMIR is basically a solution for the problems related to web and social media networks in Arabic text. Maximum Likelihood training , a procedure is proposed that estimates hidden Markov models parameters from training data but that will not guarantee the improvement in the tagging accuracy [59]. Stemming is performed by calculating the most probable path, through the HMM states ,this method finds the HMM parameters which are used to estimates the most probable stem for an arbitrary [62]. A common approach is proposed to define a model and increasing the probability of the hidden structure given the observed data. This is performed using the maximum-likelihood estimation (MLE) of the model parameters by Goldwater et al. [63]. With the help of dictionary lookup and customized rules for Gujarati language , more attention is directed toward prefix removal of this language by Desai et al. [68]. In Indonesian POS tagging problems, Yuwana et al. [86] evaluate models with deep architectures for Indonesian POS tagging problems to find the best structures for Indonesian POS tagging. Models with various number of hidden layers are investigated.

5.5 Sentiment Analysis: According to Balahur et al. [38] , Sentiment analysis is the natural language processing task dealing with sentiment detection and classification from texts.

5.6 Morphological Analysis: Motlani et al. [52] to understand the morphology of Sindhi language , an open-source finite-state morphological analyser is proposed in this paper. Apertium's lttoolbox is used as finite-state toolkit to introduce the transducer. A solution has been proposed for handling of highly inflected and compounding languages called 'Morfessor' where words usually consist of lengthy sequence of morphemes. A lexicon of word segments called morphs, is also retrieved from the data by Creutz et al. [66]. Arabic finite-state morphological analysis and generation is performed. The analysis shows the root, pattern and all other affixes combined with feature tags directing part-of-speech, person, number, mood, voice etc [67].Getting problem due to the highly inflected and complexity of Arabic language morphological structure, Arabic Language has complex morphology; this led to unavailability to standard Arabic morphological analysis tools until now. Also implement and integrate Arabic morphological analysis tools into the leading open-source machine learning and data mining tools, Weka and RapidMiner [76]. A model is implemented in a task of unsupervised morpheme segmentation of Finnish and English words, with this almost as good results are obtained in the English task [78]. Whereas ,a generative probabilistic model is applied to segment word forms into morphs. The morphs are assumed to be generated by one of three categories, namely prefix, suffix, or stem that make easy to understand [79]. Baroni et al. [80] proposed a fully unsupervised language-independent text stemming technique that clusters morphologically related words from the corpus of the language using both lexical and co-occurrence features such as lexical similarity, suffix knowledge. A list of root words is prepared manually along with a rule list to determine the inflectional and derivational class of Assamese words. The result of this approach was impressive as the morphological analyzer provided 84.64% (approximately) accuracy as per the test case [89]. The second grade is a critical milestone in the development of word reading in Arabic because readers begin to use morphological cues in their word reading at this age, and transition from a grapheme-based (letter and diacritic) phonological recoding mechanism to a letter-based morpho-orthographic mechanism as a natural response to the transparent representation of morphology in the written word [94].

5.7 Lemmatization: A lemmatizer for Sindhi language in devanagari script has been proposed by Nathani et al. [53].

5.8 Text Mining: Link detection-a rapidly evolving approach to the analysis of text that shares and builds upon many of the key elements of text mining-also provides new tools for people to better leverage their burgeoning textual data resources [39]. An unsupervised program that learns the structure of words itself was proposed in which data (raw untagged Hindi corpus) will be fed to this program helps it to discover all the possible options in morphology of a word [50].

Table 1:

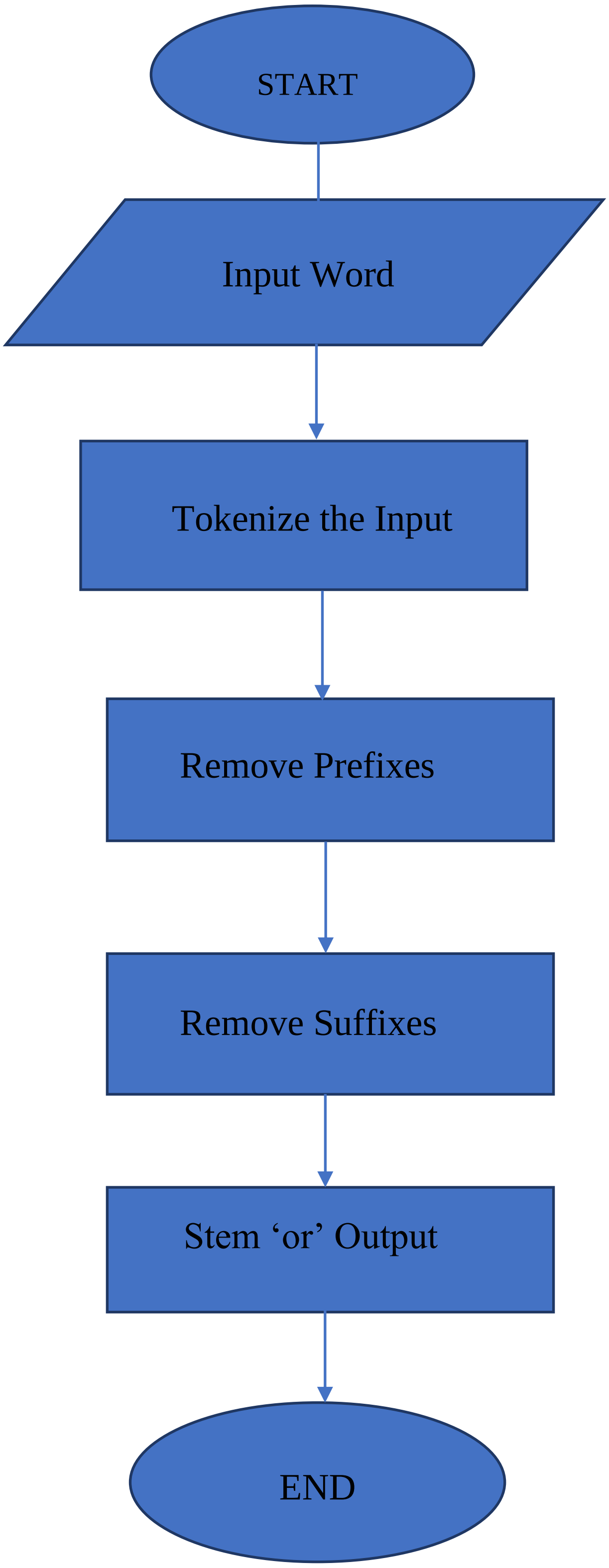
Some comparisons with an old survey paper [97] on Stemming

Features	**This Survey Paper**	**Old Survey paper**
Language	A number languages are covered such as : Assamese, Urdu, Arabic, Turkish, Finnish, English etc.	Only Urdu language
Based Applications	Information Retrieval (IR), PoS tagging, Text Classification, Morphological Analysis, Neural Networks, Lemmatizer.	Information Retrieval (IR), PoS tagging, Text Classification, NER, Machine Translation
Architecture (Diagram)	Present	Present
Evaluation Measures	Accuracy, Precision, Recall, F-measure, TF-IDF, Normalized mutual information (NMI), Pearson Correlation score, Direct Approach	Accuracy, Information Retrieval Accuracy, Precision, Recall, F-measure
Advantages/ Disadvantages	Present	Only Limitations

So, the above table consists of the features which are different in the cited survey paper. The old survey paper is basically more focused on Urdu Language in which Jabbar et al. [97] have described about approaches and challenges in Urdu Stemmers. A well-defined architecture is also shown and described. Whereas on the other hand, this survey paper is concerned about a number of languages in stemming. Different evaluation measures are used to get the accurate results. Advantages and disadvantages are also included. Multilingual text processing is useful because the information content found in different languages is complementary, both regarding facts and opinions. While Information Extraction and other text mining software can, in principle, be developed for many languages, most text analysis tools have only been applied to small sets of languages because the development effort per language is large [97]. Improved information retrieval techniques have become necessary because of the huge amount of information people have available, which continues to increase rapidly due to the use of new technologies and the Internet. Stemming is one of the processes that can improve information retrieval in terms of accuracy and performance [98]. Stemming is a critical stage in many natural languages processing applications, including information retrieval, part-of-speech tagging, syntactic parsing, and machine translation. It is a morphological procedure that attempts to convert a word's inflected forms to its root form [99]. Stemming is a pre-processing step in text mining application and commonly used for Natural Language Processing (NLP). A stemmer can execute operation of altering morphologically identical words to root word without performing morphological analysis of that term [100].

5. Architecture

Stemming is an Information Retrieval pipeline characteristic that is extensively employed in text mining and natural language processing. The primary goal of stemming is to reduce an inflectional or derivational word to its root form. The difficulty in building a stemming algorithm is identifying and removing affixes because each language has distinct traits and grammatical norms.



6. Used methodologies in the base papers

In this survey, different types of stemming algorithms are proposed. Many of them are working fine but there is some of the work required in areas like while dealing with more inflected languages some of the algorithms not able to perform better with accuracy and other metrics. In the below defined table, all the proposed methods are shown with respect to their type, and language holdings.

Table 2 : Comparison of the proposed methods

Reference	**Proposed Method**	**Type**	**Description**	**For which language it is proposed**
[1]	Urdu Stemmer and classification of short text	Supervised Algorithm	The majority of unstructured Urdu textual data is available around the world. Data mining techniques can be used to extract relevant information from such a huge, potentially interesting base data set.	Urdu and Arabic
[2]	Phonotactic and Stemmer for Indonesian phonemicization	Supervised Algorithm	Proposed method discussed about a Phonemicization, also known as grapheme-to-phoneme conversion (G2P), is a method of translating a word into its pronunciation.	Indonesian
[3]	Language-independent Stemmer for information retrieval	Unsupervised	The key principle is that in Information Retrieval systems, words with comparable meanings and morphologies should be treated as equivalent. As a result, the major goal is to minimize all morphological variants to the word's base form	Multi-lingual
[4]	Text Analytics and Text Mining Approach for Multi-lingual Processing	Meta Classification	There is propose work to Morphology, which is an important aspect of natural language processing, also known as Morphological parsing, in this multilingual environment	Multi-lingual
[5]	Improving Arabic stemming method	Unsupervised	Offering a third stemming method that utilizes lexical resources to validate stems more thoroughly.	Arabic
[6]	Arabic Morphology Information Retrieval Algorithm	Supervised Learning Model	The proposed solution is used to extract the Arabic stem on a checking of the letter before removing affixes by building AMIR dictionary.	Arabic Language
[7]	'Linguistica 5’ which is based on the Minimum Description Length (MDL) Algorithm	Unsupervised Learning Algorithm.	This stemmer will focus on defined parameters such as Maximum affix length, Minimum Stem length and Minimum Signature count. Through a List of Signature, 'Linguistica 5' tool will try to find new signature which will contains a stem and pair of suffixes.	Sindhi Language
[8]	Semantic Text Similarity (STS) at sentence level	Supervised Algorithm	The proposed solution is combination of the different Arabic stemming, stemmer and lemmatization algorithms which are around 10 in numbering to describe the effect of stemming and lemmatization on	Arabic Language

			Arabic sentences.	
[9]	Bayesian HMM model for joint learning of PoS tag and stems	Unsupervised framework	The proposed solution contains overcoming of out-of-vocabulary (OOV) issue and proposing a new idea of combination learning of PoS tagging and stemming in a fully unsupervised environment.	Turkish, Finnish, Hungarian, Basque, and English Languages.
[10]	Morfessor 2.0	Unsupervised Statistical Model	The proposed method is used to segment word into sequence of morphemes to be represented in a smaller number of bits and in this way using the Minimum Description Length MDL technique, it will help in solving the problem related to stemming.	Turkish
[11]	CmTLB	Classification and clustering tasks.	Proposed a new method for evaluation of long texts and short texts based on LDA and DMM, respectively evaluate CmTLB against the studied weighting schemes by the quality of the learned topics (topic visualization and topic coherence).	Arabic
[12]	Arabic light-based stemmer	Light-based stemmers and artificial intelligence stemming	The proposed method used to improve the Arabic light-based stemmer that can extract the expected stems by effectively removing all affixes with exception to the affixes that are a part of the original word	Arabic
[13]	lexical and co-occurrence features, lexical similarity	Unsupervised	The proposed method used for solving the problem of vocabulary mismatch in queries and documents at the time of indexing and searching	Marathi and Hungarian
[14]	Support Vector Machine Algorithm, existing algorithm	Text Classifier Model	Existing stemming methods have been proposed, and the existing stemming methods are proven to result in high accuracy level.	Indonesia
[15]	Morphological Operation	Morphosyntactic features	The morphological operation effect would reflect the morphosyntactic features checking in the combination of morphemes for word verification and recognition. word recognition is mediated by sub-lexical processing driven by morphological operations	French
[16]	Assamese linguistic	Unsupervised	In this method, we considered stripping suffixes and all possible suffixes. The suffixes that are collected are divided into eight different suffixes such as plural words, case words, definitive words, pleo words, extra words, in-definitive words, verbs, and kinships.	Assam Language
[17]	Smoothing Algorithm	Supervised	It is performed to eliminate local minima because the main problem in using this method, though, is that the difference	Multi-lingual

			identified between cohesion scores can be, at times, caused by local minima that do not correspond to a topic boundary.	
[18]	Arabic Diglossia	Classification	It encourages young readers to rely on information beyond the phonological word in their reading, such as morphemes, which can help in accessing meaning from the orthographic letter string.	Arabic
[19]	Semantic	Classification	The semantic analysis of natural language content starts by reading all of the words in content to capture the real meaning of any text. It identifies the text elements and assigns them to their logical and grammatical role.	Multi-lingual
[20]	Rule-based Algorithm	Classification Model	In this method, it stores and manipulate knowledge to interpret information in a useful way. The term rule-based system is applied to systems involving human-crafted or curated rule sets.	Multi-lingual

6. 1 Supervised Learning Algorithms:

Alnaied et al. [6] proposes a method to produce a powerful tool to extract Arabic root words by deploying a morphological analysis with the help of some linguistic regulations. AMIR approach is basically performing better in Arabic stems extraction as compared to previous methods such as LUCENE and FARASA methods. These existing methods only perform better in affixes removal but when it comes to understanding then they are not able to recognize whether the removed letters are actually core letters of the root word or not.FARASA algorithm was proposed by Darwish and this also helps in segmentation of the Arabic text into words. This mainly handles prefixes and suffixes in the words and in our AMIR solution infixes are also handled in addition. 'Trec_eval' software is the standard tool used by the TREC community for evaluating Information Retrieval IR system [6]. The use of statistical metrics like TREC_EVAL tool to measure the precision @ 10, precision @20 and Mean Average Precision MAP as evaluation metrics. Term Frequency Inverse Document Frequency TF.IDF is also important in terms of frequency metrics as it counts on word's used frequency and used to evaluate the quality of our scheme performances retrieval.In 50 queries, AMIR attained MAP values by 0.34% on the other hand LUCENE, FARASA and No stemmer are 0.27%, 0.28% and by 0.21, respectively by using MB25 model.When comparing of the AMIR with existing two stemmers, FARASA and LUCENE, and No stemmer using BM25 model and language model LM with Dirichlet technique in order to check the quality of scheme performance, we take an example of reading a word of Arabic language by both side (AMIR and FARASA,LUCENE), the word مساجد (Mosques) indicates to plural, AMIR method is able to reduce this to a singular form مسجد) Mosque) (AMIR rule No 3) in place of the previous word but on the other hand FARASA and LUCENE, unable to reduce it to singular form as they cannot handle plural using infixes. Alhawarat et al. [8] basically focuses on semantic similarity of Arabic texts at sentence level and explores mainly the effect of using stemming and lemmatization on semantic text similarity. This method is a combination of the different Arabic stemming, stemmer and lemmatization algorithms which are around 10 in numbering to describe the effect of stemming and lemmatization on Arabic sentences. First, we have to select stemming algorithms and prepare new training and testing text documents after pre-processing, after that choose machine learning algorithm to use in the study, after that selection of the features to use in experiments by applying String and Character-Based, Term-based and Distance-Based similarity measures. At last, setting up of the experiment by choosing ML algorithm. Pearson correlation for cosine similarity is used in the baseline here to evaluate. Pearson correlation score is calculated between each pair of the gold test sentence using Binary encoding, TF-IDF encoding and word embedding vectors as 62.32%,63.04% and 55.56% respectively.Using ARLSTem stems, Tashaphyne roots, ISRI stems and Qalsadi lemmas, we received an enhancement in Pearson correlation in the range of 3.24%-7.13% while using Word Embedding representations the statistical similarity receives lower results with the range of -1.42%-5.5% [8]. Ma, T., Al-Sabri [11]. This paper proposed a new method for evaluation of long texts and short texts based on LDA and DMM, respectively evaluate CmTLB against the studied weighting schemes by the quality of the learned topics (topic visualization and topic coherence), classification, and clustering tasks. Gibb's sampling is used to train all models sparsity of short texts. use of CmTLB outperform the baseline models in most cases for both long and short document datasets. For example, the improvement over LDA-CEW is about 1%–6%. It is noticeable that topics learned by DMM-CmTLB are cleaner and more coherent. Stemming process significantly improves the performance of topic modeling and that Farasa stemmer outperforms the other stemmers with statistically significant enhancements. The CmTLB performs about 1%–4% better than DMM-CEW and 1%–9% better than standard DMM. We also observe that term weighting models achieve better classification accuracy scores than the standard DMMFarasa, LDA, DMM algorithm outperforms the other stemmers to helps in segmentation of the Arabic text into word. Text mining, text classification, automatic data extraction, stemming, custom data analysis used to confirm that the stemming process has increased the accuracy of classification. Khoja Stemmer is a root-based stemmer that uses a root dictionary [11].Alshalabi [12]we aim to improve the Arabic light-based stemmer that can extract the expected stems by effectively removing all affixes with exception to the affixes that are a part of the original word. light-based stemmers and artificial intelligence stemming approaches. Root based stemmers use morphological analysis to extract the root of a given Arabic word. The solution of the above problem will be to propose new lists of suffixes and prefixes. The lists will be used by the proposed algorithms that use word length rules to remove suffixes and prefixes from target words. We proposed a fully unsupervised language-independent text stemming technique that clusters morphologically related words from the corpus of the language using both lexical and co-occurrence features such as lexical similarity, suffix knowledge. Natural Language Processing (NLP) systems to transform the morphological variant word forms to their base forms. For example, the variant word forms abandon, abandoned, abandoning, abandonment, abandonments, etc. are transformed to their base form abandon through stemming. Stemming improves the ability of an IR system to retrieve more relevant documents by solving the problem of vocabulary mismatch in queries and documents at the time of indexing and searching. the performance of stemmer in improving retrieval accuracy increases with the increase in the morphological complexity of the language. The supporting tools are LDA,DMM,and The Stem-Based Approach, referred to as Light Stemmers, focus on removing the common prefixes and suffixes of given Arabic words. D1ight increases the stem F- measure approximately by 34% more than the light stemmer, 31% more than the Cond light stemmer, and 25% more than the ARLST.One of the most popular root-based stemmers is that of Khoja and Garside, which removes suffixes, infixes and prefixes. This stemmer is largely based on pattern matching and pattern strength techniques to extract the root of words. To evaluate our proposed light-based Arabic stemmer, we compared our stemmer against existing Arabic stemmers, namely, Light10, Cond light and ARLST.Dataset used for designing and testing an Arabic stemmer, the standard dataset Text Retrieval Conference (TREC 2002) is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER style SGML and was transcoded to Unicode (UTF-8) for testing an Arabic stemmer.TREC dataset is used for question classification and divided into broad semantic categories. Text Retrieval Conferences is used to encourage research in information retrieval from large text. Its main purpose was to support research within the information retrieval community by providing the infrastructure necessary for large –scale evaluation of text retrieval methodologies [12]. [15] Estivalet proposed an unsupervised stemming method that helps us to discovers morphologically related words from the corpus based on both lexical and semantic knowledge has also been tested. MORFESSOR and HPS performed better for morphologically and inflectionally rich languages Marathi and Hungarian as compared to English, and Bengali. stemmer is largely based on pattern matching and pattern strength techniques to extract the root of words. To evaluate our proposed light-based Arabic stemmer, we compared our stemmer against existing Arabic stemmers, namely, Light10, Cond light and ARLST. Dataset used for designing and testing an Arabic stemmer, the standard dataset Text Retrieval Conference (TREC 2002) is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER style SGML and was transcoded to Unicode (UTF-8) for testing an Arabic stemmer. The corpus analysis helps in reducing all such erroneous conflation. As a results, the study should focus on distinct stemming approach HPS, YASS, MORFESSOR etc. for different languages [15].Smoothing algorithm is performed to eliminate local minima because the main problem in using this method, though, is that the difference identified between cohesion scores can be, at times, caused by local minima that do not correspond to a topic boundary [17]. The performance of the bisect K-means clustering algorithm compared to the standard K-means algorithm in the analysis of Arabic documents. The experiments included five commonly used similarity and distance functions (Pearson correlation coefficient, cosine, Jaccard coefficient, Euclidean distance, and averaged Kullback-Leibler divergence) and three leading stemmers. Using the purity measure, the bisect K-means clearly outperformed the standard K-means in all settings with varying margins. For the bisect K-means, the best purity reached 0.927 when using the Pearson correlation coefficient function, while for the standard K-means, the best purity reached 0.884 when using the Jaccard coefficient function. Removing stop words significantly improved the results of the bisect K-means but produced minor improvements in the results of the standard K-means [17]. we proposed two Arabic topic segmenters, ArabC99 and ArabTextTiling. The main advantage of these segmenters is that they are based on the two most well-known and used segmenters, C99 and Texttiling. Moreover, each segmenter is based on a different segmentation unity: sentences for ArabC99 and blocs for ArabTextTiling. However, these segmenters have a severe limitation in terms of choosing the stemming algorithm that was not based on a systematic analysis of Arabic stemming algorithms. Moreover, these segmenters follow an endogenous approach. For these reasons, we opted to study these two topic segmenters on two levels. First, we will focus on the pre-processing level through studying the choice of stemming algorithms for the two segmenters ArabC99 and ArabTextTiling. Then, we will focus on the segmentation level by adding semantic knowledge to the topic segmentation and by studying the impact of using sentences and blocs on the segmentation process [17]. This segmenter is based on C99, and it goes through two important steps. The first is pre-processing, which is dedicated to the extraction of words, elimination of stop words, and stemming by using the Khoja stemmer algorithm. The second step is segmentation. This step includes four operations. The first one corresponds to the construction of the frequency dictionary by linking each word with its stem and its frequency. The second one is the similarity matrix construction by using the cosine measure between sentences the third operation is the rank matrix construction by using the similarity matrix. The rank is the number of neighboring elements having a lower similarity value using a rank mask, and it is presented as a ratio to avoid normalization problems. Finally, the fourth one corresponds to topic boundaries identification, which is based on the density calculation. In relation to this density, the detection of boundaries is realized by making use of Reynar algorithms (Reynar 1989) that look for the distribution with the highest density [17]. TextTiling segmenter goes through two important steps. First one is pre-processing, which is the same as ArabC99, and segmentation, which includes three operations. This corresponds to the construction of blocs by the means of a sliding window, whereas the second one is related to the similarity matrix calculation by employing the cosine measure such as in ArabC99. However, the difference between both segmenters is that in ArabC99, the similarity is calculated between sentences while in ArabTextTiling, the similarity is calculated between bloc's sim (b1, b2), where b1 and b2 correspond to blocs. The third operation refers to the topic boundaries identification, and we could observe that the major difference between ArabC99 and ArabTextTiling is located in this step. In fact, for ArabTextTiling a cohesion score is calculated to quantify the similarity between neighboring blocs. Thus, a topic boundary is identified if the cohesion values of the previous and following blocs are very different from the current bloc. The main problem in using this method, though, is that the difference identified between cohesion scores can be, at times, caused by local minima that do not correspond to a topic boundary. To overcome this problem, a smoothing algorithm is performed to eliminate local minima. How morpheme order is determined in language more generally, has been proposed that semantic relationships among morphemes, sometimes called scope, determine linear order. And Related theories argue that universal syntactic hierarchies, potentially reflecting semantics, determine order. On one formulation, morphemes which more directly affect or modify the semantic content of the base have narrower scope. Wider scope morphemes modify the larger semantic constituent which includes any lower scoping morphemes [19]. We test whether participants learning a miniature artificial language are indeed biased in favor of placing number morphemes closer to noun stems than case morphemes. These experiments are also designed to investigate the mechanism underlying any such bias—in particular whether the bias is driven by (absolute and relative) frequency, or by a cognitive preference, for example, for scopeisomorphic ordering. To preview, we uncover clear evidence for biased ordering across two populations (English and Japanese speakers). We also find that it holds independent of morpheme position (prefixal or suffixal), degree of boundedness (free or bound morphology), frequency, and which particular case/number feature values are instantiated in the overt markers (accusative or nominative, plural or singulative). All things equal, this suggests that the typology may reflect frequency independent cognitive biases of learners. However, we also find that the presence of case allomorphy conditioned on the stem (which strengthens the dependency between case markers and the stem) can reverse participants' preferences. We interpret this as a competing bids for local dependencies. This result adds to the growing body of work using experimental methods to investigate how learning and use shape typological patterns in morphology and word order. A novel integrated methodology that leverages multiple computational techniques to extract heterogeneous American-English data terms used in different highway agencies and their semantic relations from design manuals and other technical specifications. The proposed method implements Natural Language Processing (NLP) to detect data elements from text documents, and employs machine learning to determine the semantic relatedness among terms using their occurrence statistics in a corpus. The study also consists of developing an algorithm that classifies semantically related terms into three different lexical groups including synonymy, hyponymy and meronymy. The key merit in this technique is that the detection of semantic relations uses only linguistic information in texts and does not depend on other existing hand-coded semantic resources. A case study was undertaken that implemented the proposed method on a 16-million-word corpus of roadway design manuals to extract and classify roadway data items [19]. A more equivocal finding was the interaction of surface-form frequency and phonological neighborhood density, which (whether in its class- or form-based variant) was observed only for Polish individually (Though also, in an exploratory analysis, when collapsing across all languages). Again, this finding – if subsequently confirmed – is difficult to explain under rule-based accounts, which posit no mechanism that would yield such an effect, but it falls naturally out of connectionist and exemplar accounts: Recall from the Introduction that connectionist models predict this interaction and observe it, because high-frequency pairs (e.g., stem-output) develop their own dedicated input–output mappings, while low-frequency pairs rely more on the abstract representations stored in the hidden units [20]. We can define "a word embedding" as content representation such that words of similar meaning also receive the same representation. This method deals with representing documents and words and it may be seen as one of the keys in the procedures ahead of deep learning when testing natural language processing (NLP) problems. Furthermore, it is a category of methodologies in which vectors that are real-valued are used in a predefined vector space to represent single words. Every word is determined to be a vector and the values of the vectors are discovered following a neural network method. Later, the technique is usually grouped into a profound learning field. The key to the approach is to use a densely dispersed representation for each word. A real-valued vector is utilized to represent each word, frequently tens of, or many, measurements. This is divided into hundreds and thousands or matched to larger numbers of dimensions required to represent a word, such as one-hot encoding [20].

6.2 Unsupervised Learning Algorithms:

Nathani et al. [7] designed an unsupervised learning of morphology using a tool called 'Linguistica 5' which is based on the Minimum Description Length (MDL) algorithm. This approach is somewhat differed from the previous rule based, time consuming approaches. This stemmer will focus on defined parameters such as Maximum affix length, Minimum Stem length and Minimum Signature count. Through a List of Signature, 'Linguistica 5' tool will try to find new signature which will contain a stem and pair of suffixes with every iteration till no new modification found in signature. The MDL algorithm which is used in this scenario is basis of statistical modeling, pattern identification and machine learning. Minimum Description Length holds better explanation of given set of observed data. It basically helps in smallest description of the given data.There is various command line interface of 'Linguistica' with the different parameters such as Maximum affix length =4, Minimum Stem length=4 and Minimum Signature count i.e., minimum number of stems for a valid signature is 5. In our evaluation process, our algorithm performs well with a great accuracy but with some minor ups and downs in the form of under stemming and over stemming [7]. Bölücü et al. [9] proposed a solution contains overcoming of out-of-vocabulary (OOV) issue and proposing a new idea of combination learning of PoS tagging and stemming in a fully unsupervised environment. For joint learning of PoS tag and stems, Bayesian HMM model is used. In this approach, PoS tagging is executed in combination with other methods such as morphological segmentation and morphological disambiguation. Hidden Markov Model is a tool for finding probability distribution over a sequence of observation. It is a Markov model in which the system being modelled is assume to be a Markov process with hidden states. The proposed system is evaluated by four metrics in case of PoS tagging results: many-to-one, one-to-one, normalized mutual information (NMI), and variation of information (VI). Semantic similarity is also used between the stems and words to find the inflectional morphology. Neural word embeddings are also used so that semantic similarity between the word and stem can be observed. The results which have been received show that a joint model for PoS tagging and stemming increases on an independent PoS tagger and stemmer in agglutinative languages. Comparison of other PoS tagging and stemming models with our models shows better results. Our joint model for PoS tagging and stemming perform better as compared to word-based Bayesian HMM model, Brown Clustering and Anchor HMM model for three agglutinative languages: Turkish, Finnish, and Hungarian [9]. Özbey et al. [10] proposed solution is basically 'Morfessor 2.0' which is an MDL- based technique in that it mainly looks upon Expectation Maximization (EM) to update the probabilities of segments and it takes a lexicon as a input. This proposed solution is somewhat differed from Morfessor, first in case of codebook, it has two dictionaries to store length of segments: one for prefixes and other one is for suffixes whereas in previous method we got only one single codebook. Second, in 'Morfessor 2.0', we can initialize dual dictionaries by a heuristic in a deterministic way whereas random differentiating is applied in 'Morfessor'. Third, during the EM stage, the dictionaries can be updated with M different prefix suffix pairs with their corresponding probabilities whereas only segments in the most likely segmentation are considered, and their parameters are too determined irrespective of probability. Fourth, the time complexity is O(M) for stem boundary within a word of size M whereas it will be O(M^2) in case of 'Morfessor'. SDDM Stemmer is evaluated on the test sets with different thresholds for pruning and without pruning it gives the same yield of NDM stemmer of 37.25% and 18.59% accuracy rates for mixed set and highly inflected verbal set respectively. PDDM Stemmer is evaluated by checking its accuracy results on y-axis after applying EM on x-axis, while in no EM, PDDM stemmer increases its accuracy rate by 16.25%, 21.72% and 41.54% for mixed, highly inflected verbal, and for uninflected sets respectivelyOur proposed solution 'Morfessor 2.0' is also evaluated with these three: mixed, highly and uninflected sets and accuracy score are compared with the SDDM and PDDM Stemmers. PDDM Stemmer has performed with better results when trained with threshold =2 by 10 EM [10]. Singh et al[13]. compared the retrieval performance of various stemmers method using Mean Average Precision (MAP) as a primary evaluation measure. R–Precision (R–P) and Precision@10 (P@10) values have also been reported in order to analyse the precision enhancing capability of the stemmer. for evaluating the performance of various stemming methods in the text classification task, text classification is the process of assigning text documents to one or more pre-defined classes. High Precision Stemmer (HPS),an unsupervised stemming method that discovers morphologically related words from the corpus based on both lexical and semantic knowledge has also been tested. The F-score metric considers both precision and recall and is hence used as a primary evaluation metric to compare the performance of stemmers in these experiments denotes the number of times the morphological class of the stemmer contains incorrect word Yet Another Suffix Stripper (YASS), an unsupervised stemmer that uses lexical information between word pairs to group morphologically related words has been used. MORFESSOR, the unsupervised recursive Minimum Description Length based morphological analyzer has been used. Cross-validation, it is a common method to evaluate the performance of a text accuracy and classification. "Sastrawi" algorithm improvement of "Sastrawi" algorithm by using dynamic afxes to process non-formal Indonesian which were formed into abbreviated and deformed words, also while using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is checked [13]. [14] Rianto, R proposed a method in an Indonesian setting, existing stemming methods have been proposed, and the existing stemming methods are proven to result in high accuracy level. The results of proposed solution to develop a text classifier model. "Sastrawi" algorithm by using dynamic afxes to process non-formal Indonesian which were formed into abbreviated and deformed words, also while using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is checked. The existing Indonesian stemming methods are still oriented towards Indonesian formal sentences this phenomenon underlies the suggestion of developing a corpus by normalizing Indonesian non-formal into formal to be used as a better stemming method. The impact of using the corpus as a stemming method is that it can improve the accuracy of the classifer model. corpus is important for the stemming process, so it can produce a classifer model that has high accuracy. The sample words like example "berlari" (running) into "lari" (run), "menulis" (writing) into "tulis" (write), "memakan" (eating) into "makan" (eat) are formal words that do not have a problem in stemming, but getting problem in non-formal, the words which are deviated from Indonesian standard words, for example in sentences like "sistemnya lambreta bingit" (the system is very slow). "Lambreta" and "bingit" are not formal Indonesian words, so we getting a problem in the accuracy of text classification, online conservation with the nonformal word in stemming process. In linguistic computation, the non-formal language comes to problems in data preprocessing which mostly comprises tokenizing, removing, stemming, and normalisation. The solution of this problem need improvement of "Sastrawi" algorithm by using dynamic afxes to process non-formal Indonesian which were formed into abbreviated and deformed words, also while using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is checked. In an Indonesian setting, existing stemming methods have been proposed, and the existing stemming methods are proven to result in high accuracy level. inflectional suffixes, we investigated the processing of inflectional suffixes in French verbs, obtaining significant effects in the processing of a different number of morphological operations in tense and agreement morpheme.The main objective is to propose an efficient fully unsupervised corpus-based stemming approach which can serve as a multi-purpose tool in a number of Information Retrieval and Natural Language Processing applications. stemming approach exploits the lexical, semantic and cooccurrence knowledge of the words to group morphologically related words appearing in the corpus. lexical function that gives a high score to morphologically related words without any knowledge of the language [14]. We considered stripping suffixes and all possible suffixes. The suffixes that are collected are divided into eight different suffixes such as plural words, case words, definitive words, pleo words, extra words, in-definitive words, verbs, and kinships [16]. Saharia et al. [16] an HMM-based hybrid approach to classifying the mismatched last character. For each word, the stem is extracted by calculating the most probable path in four HMM states. They have obtained 94% accuracy for Assamese and Bengali and 87%, and 82% for Bishnupriya Manipuri and Bodo, respectively, using the hybrid approach. Assamese stemming might be characterized as a handle that strips off a set of postfixes from words. But this handle also has certain set back such as vocalization ambiguity, incorrect expulsion, single arrangement [16]. Assamese is another Indian language rich in morphology with inflectional and derivational variants forms of words and so applying stemming is quite difficult for this language. But, in this paper, we have not only attempted to address such problems but also proposed a few key contributions, and the same is summarized as follows: (a) Identification of the appearance of varied suffixes including their pattern classifications and (b) Improvement of a rule-driven stemming algorithm for Assamese words that is capable of stripping suffix without seeking the help of any databases. The algorithm proposed in this paper is efficient enough to extract stem words from inflected words of any length. And the results obtained after evaluation proposed stemming algorithm is capable to achieve 86.16% accuracy [16]. The facilitation induced by morphology on the reading performance of less skilled readers may reflect access to lexical reading units (morphemes) that are shorter than the whole word when this reading unit is too long and complex for the reader. Morphemes (specifically roots and suffixes) can be efficient reading units because they have an intermediate grain size between single letters—which entail extremely slow and analytical sub lexical processing—and the word—which for beginning readers and children with dyslexia is usually too large a unit to be processed as a whole. By contrast, for skilled readers who master lexical reading of familiar word units, recourse to morphemic units is beneficial only for low frequency words. In this case, morphemes (roots and affixes) usually have a higher frequency than the word in which they occur. Therefore, access to morphemes may facilitate lexical reading for a low-frequency word that otherwise would probably not be represented as a whole in the mental lexicon [18]. Arabic allows inflection through both linear and nonlinear morphological procedures. Linear word formation is usually used to create inflected forms indicating, for example, number (singular, dual, plural), in which distinctions are marked by adding suffixes to word stems resulting in regular (so-called sound) plurals. However, pluralization is also marked in Arabic using non-linear procedures, which involve irregular stem-internal vocalic changes as well, as in broken plurals. For example, kura 'ball' +the plural feminine suffix—at < ا ت > results in krat:t, but daftar 'notebook' gets the broken plural form dafa:ter, not daftarat, in which the consonants of the stem noun are inserted within a broken plural vocalic pattern CaCCaC . Inflectional categories in Arabic also include possessive forms. These are only linear but they also include regular and less regular forms. The default possessive forms which are added to the ends of nouns (e.g., i 'my', na 'our') receive an allomorph with the sound t preceding the possessive pronoun when the stem noun ends in so-called ta? marbuta, a grammatical gender marking letter-morpheme. For example, the possessive form that would parallel the English phrase 'my notebook' is daftari = daftar-i. However, the same possessive form of the phrase 'my ball' is kurati = kura -ti. This affixational procedure is less regular because in writing it would involve deleting the last letter from the noun, ta? marbuta < ة >, and replacing it by the letter < ت >, also called ta? maftuha, and then adding the possessive suffix. Similarly, in speech, speakers must note the final sound in the noun and identify it as a feminine gender marker, in which case possession should involve the use of the allomorph ti [18]. Arabic's non-linear formation, which is mainly characterized by derivation formations, is created by the combination of a consonantal root, indicating the semantic family, with a vocalic verbal or nominal pattern, composed of vowels and affixal consonants, indicating the word's prosodic structure as well as its syntactic category and related grammatical properties. For example, the word katab 'write' is created by a derivation in which the root KTB is interdigitated into the pattern CVCVC (where the C indicates a slot for the root consonant). The same root consonants may be mounted onto the pattern maCCu:C to create the word/maktu:b/'is written' and onto the pattern Ca:CeC to create the word/ka:teb/'writer', etc. This derivational procedure is non-linear because the root morpheme is inserted into slots within a fixed prosodic pattern instead of being linearly attached, as is common in European languages like English [18]. In Arabic, these two processes seem to develop in parallel at a rather young age, and their development might also depend on variations among children in mastering basic lexical skills in standard Arabic. A closer look at the results reveals differences in reading comprehension scores between readers with high and low morphological awareness. We predicted that the contribution of morphological awareness at the beginning of second grade to reading comprehension at the end of the school year would differ for these two groups of readers. In line with our predictions, high morphological awareness readers earned significantly higher scores in reading comprehension than low morphological awareness readers, at the beginning and the end of the school year. Also, whereas inflectional awareness predicted reading comprehension among both examined groups of readers, derivational awareness predicted reading comprehension only in readers with high morphological awareness. Generally, although scores in derivational awareness weren't high, they did predict success in reading comprehension and differentiated strongly between the two groups of readers. This is true at least for the specific forms targeted in this study (i.e., deverbal nouns constructed from basic verbs, which are predominant in the Arabic language, both SpA and StA) [18].

7. Evaluation of the proposed methods

7.1 Methods / Metrics used:

There are numerous methods for determining how well a stemmer performs. Methods differ depending on the type of stemmer. Some of the evaluation methods are discussed further below.

Among the approaches that are frequently used are frequency count, n-gram, Hidden Markov Models, and link analysis. The n-gram tagger is evaluated in the phonotactic domain without the use of a stemmer or phonotactic rules. Without the use of a stemmer or phonotactic rules, the n-gram tagger can be evaluated. The PERs produced by all G2P models when those optimum parameters are used, along with a comparison to the Transformer-based G2P model. Many methods, such as the TERRIER retrieval system, can be used to evaluate a single stemmer. 2. Weighting model IFB2 3. Statistical divergence 4. Retrieval precision on average 5. Accuracy in recall 6. Precision @ 10 7. The T-test. Comparing the proposed stemmers to previous works in order to demonstrate their contributions and improvements. Buckwalter morphological analyzer and Tri-literal root extraction are two evaluation methods used in various experiments. Precision @10, Precision @20, and

Mean Average Precision (MAP) are all used in the field of information retrieval. The Pearson correlation coefficient between machine scores and human judgments is sometimes used to assess performance. Two evaluation matrices, normalised mutual information (NMI) and Purity, are used in Document Clustering evaluation. Cross-validation is a common method for evaluating the accuracy and classification of a text. Inflectional suffixes are used to investigate the processing of inflectional suffixes in languages such as French verbs, yielding significant effects in the processing of a variety of morphological operations in tense and agreement morphemes.

7.2 Similar approaches:

To ensure that the proposed method is superior and provides higher accuracy, it must be compared to its competitors. Several approaches and alternative methods for comparing are discussed further below.

The proposed minimum word length rule was examined first, followed by the prefix rule, and finally the infix rule was evaluated. A comparison of the proposed method with a lightweight Urdu stemmer. Experiment 1 used the same news headline, and it was discovered that incorrect interpretation of prefixes and postfixes has a significant impact on stemming accuracy [1]. In order to classify Approach for evaluating proposed short Urdu text classification The following news headlines were used: politics, sports, terrorism, and weather. Their new infix stemming strategy, along with the concepts of prefix and postfix stemming, clearly improves classification accuracies. In comparison to Naive Bayes and SVM-based Urdu text categorization algorithms. The proposed complete Urdu text pre-processing and classification system outperforms the competitor's efforts significantly[1].

In terms of PER and WER for Indonesian phonemicization, the model is compared to the state-of-the-art Transformer-based G2P model. There are more requirements[3] when working with a tool that is language independent and works for many languages, such as the TERRIER retrieval system, IFB2 weighting model, Probabilistic divergence, Average retrieval precision, Recall precision, Precision @10, and T-test. Stanford's Core NLP Suite is a better choice for multilingual work as well.

There are several light stemmers available for Arabic. Lights 1, 2, 3, 4, 5, 6, 7, 8, and 10 are among them. In terms of retrieval information, the Light10 stemmer outperforms the other light stemmers[5]. For Arabic information retrieval, FARASA and LUCENE are used, and Jaccard similarity is used to measure similarities between sets. It is defined as the intersection size divided by the union size of two defined sets.

When it comes to agglutinative languages, there are primarily three methods for evaluation[8]. 1)Word-based Bayesian HMM, Word-based Bayesian Hidden Markov Model the Markov Model is a tool for determining the probability distribution over a sequence. It is a word-based Markov model with some states hidden. 2)Brown Clustering, Brown clustering is a method for creating clusters of similar words. It is an example of a clustering algorithm that generates a hierarchical cluster of words. 3) In the anchor HMM model, each hidden state has at least one observation that can only be created by that state. For example, "they"" can only be tagged as a determiner and cannot be tagged in any other way.

SDDM (Strict Dual Dictionary Model) is a stemming algorithm that validates suffixes to reduce the number of false positive stems found by the Naive Dictionary Model (NDM)[9]. Furthermore, the Probabilistic Dual Dictionary Model (PDDM) defines a mechanism for determining the most significant stem boundary by calculating the probabilities of all valid prefix-suffix pairs. "Sastrawi" algorithm improvement by using dynamic afxes to process non-formal Indonesian which were formed into abbreviated and deformed words, as well as while using the Support Vector Machine algorithm, a text classifier model is developed, and its accuracy is tested.

7.3 Dataset:

There are many datasets related to Urdu and Arabic, such as 1) An Urdu news headline corpus. It includes news from two main areas, namely politics and weather, and it also serves as an Urdu headline news corpus. It is divided into two news categories: sports and terrorism, and it contains one-of-a-kind Urdu words. It evolved through the use of several grammar books and Urdu dictionaries, and 4) combining corpus 1, corpus 2, and corpus 3 produced a full headline news corpus. There is a Great Dictionary of 50, 000 Indonesian words, or Kamus Besar Bahasa Indonesia (KBBI). Using high-quality datasets such as the Wall Street Journal document collection for English and the FIRE 2010 document collection for Bengali and Marathi.

EveTAR (2016) SemEval2017 datasets and datasets can be used for both training and testing, and the language used in these datasets is Modern Standard Arabic (MSA). There is also a large lexicon that has been used, which consists of 363,370 words created through tokenization of Turkish Wikipedia for the Turkish language. The Text Retrieval Conference (TREC 2002) standard dataset is used, which is made up of articles from the Agence France Presse (AFP) Arabic Newswire. The WebKB dataset is a collection of web pages that have been searched for the most frequently used words, allowing the model to train more efficiently.

8. Comparison setup

In this survey paper , a number of base papers are seen through different parameters. Some of the parameters can be seen in the below table based on the list of the method used to evaluate, similar approaches, comparison of the results of the proposed solution with the similar approaches and keywords used for comparison.

Table 3: Comparison of the Base Papers over different parameters

Reference	**List of Methods used to Evaluate**	**Similar Approaches**	**Results comparison of proposed approach with similar approach**	**Keywords used for comparison**
[1]	Minimum word Length rule, Frequency count, N-gram, Hidden Markov Models.	1. Evaluation of Proposed Urdu Stemmer 2. Light weight Urdu Stemmer 3. Short Urdu text classification	1. Using minimum word length rule 2. Naive Bayes and SVM.	Prefix Rule, True Positive, False Positive, Postfix Rules.
[2]	N-gram tagger, NGT, NGTP, NGTS,	1. Transformer based G2P model	1. Using the illustration in figures and using the frequency 2. G2P ,with a substantially higher mean PER (up to 1.14 percent).	NGPT, NGT, NGTSP, Standard Deviation
[3]	TERRIER retrieval system,IFB2 weighting model, Probabilistic divergence .	1. TERRIER retrieval system 2. GRAS 3. PORTER	1. The retrieval studies were carried out with the use of the IFB2 weighting model, which is based on the probabilistic divergence from randomness paradigm.	Linguistica, YASS, GRAS, HPS, IFB2, Probabilistic divergence, Random Paradigm, Light Bengali Stemmer, Light Marathi Stemmer, High Precision Stemmer.
[4]	Evaluation done through some computation of linguistics.	1. Stanford's Core NLP Suit 2. Part of speech tagging	1. Natural Language Processing methods such as Text Mining and Text Analytics are used to leverage natural language processing.	Multilingual, Meta classification, NLTK, morphological parsing.
[5]	Buckwalter morphological analyser and Tri-literal root extraction.	1. Light stemmers for Arabic 2. Motaz Stemmer 3. Tashaphyne ,a light stemmer	1. Researchers chose three light stemmers, Light10, Motaz stemmer, and Tashaphyne, and compared their results with our SAFAR-Stemmer.	Stemming results, Annotated evaluation corpus, and Gs-Score.

[6]	Precision @10, Precision @20, Mean Average Precision (MAP)	1. FARASA 2. LUCENE 3. No stemmer	1. The values of MAP, Precision@10 and Precision@20 have been found out to compare it them with the values obtained by AMIR algorithm. 2. AMIR algorithm performs better with 0.34%, 0.63% ,0.59% values for MAP,Prec@10 and Prec@20 respectively using BM25 model.	Mean Average of Precision MAP, Precision @10, Precision @20, TF.IDF values
[7]	Accuracy, Direct approach by an Language Expert	1. Rule Based Stemmer 2. Lemmatizer	1. Our proposed approach has given an accuracy of 87% which is greater than Rule- Based stemmer but less then rule based Lemmatizer which is around 90%.	Accuracy , Total Number of Stems ,Total Signature , Output Stem , Actual Stem
[8]	Pearson Correlation	1. Bi-gram 2. Tri-gram 3.Jaccard similarity	1. Pearson correlation score is calculated between each pair of the gold test sentence using Binary encoding, TF-IDF encoding and word embedding vectors as 62.32%,63.04% and 55.56% respectively.	Pearson correlation ,Alkhalil Stems ,ARLSTem Stems ,Assem Stems ,,Tashaphyne Stems
[9]	Many-to-one, One-to-one, Normalized mutual information , Variation of information	1. Word-based Bayesian HMM 2.Brown Clustering 3.Anchor HMM model	1. Comparison of other PoS tagging and stemming models with our models shows better results. 2. Joint model for PoS tagging and Stemming perform better as compared to word-based Bayesian HMM model, Brown Clustering and Anchor HMM model for three agglutinative languages : Turkish, Finnish, and Hungarian.	Cosine similarity , Tagset size , Many-to-One Scores, Accuracy Scores Obtained from Different Hyperparameter Values
[10]	Accuracy	1.Strict Dual Dictionary Model(SDDM) 2.Probabilistic Dual Dictionary Model (PDDM)	1. The proposed solution Morfessor 2.0 is evaluated with these three : mixed , highly and uninflected sets and accuracy score are compared with the SDDM and PDDM Stemmer.	Threshold values , Number of EM iterations , Threshold values of the Initialisation
[11]	Document Clustering, Two evaluation matrices, normalized mutual information.	1.Stem-Based Approaches 2.Statistical Approaches 3.Root-Based Approaches	1. The vocabulary size has been significantly decreased when applying root-based stemmers such as Khoja stemmer compared with the other stemming approaches.	Latent Dirichlet allocation (LDA), Dirichlet multinomial mixture (DMM), Term weighting schemes.
[12]	F-measures	1. Root based stemmers 2. Arabic Morphology Information	1. We compare the D1ight stemmer with the three selected Arabic stemmers: Condlight, Light10 and ARLS that reduce.	Arabic Morphology Information Retrieval (AMIR), Condlight, Light10 and ARLS.

		Retrieval (AMIR)		
[13]	Mean Average Precision (MAP), R–Precision (R–P) ,Precision@10, Precision Stemmer (HPS)	1. Corpus-based method	1. The corpus-based stemming methods are broadly classified into two main categories namely Lexicon analysis based approaches and Corpus analysis based approaches.	Corpus-based methods, Yet Another Suffix Stripper (YASS), F-score metric, unsupervised stemming
[14]	Cross-validation	1. “Sastrawi” Algorithm 2. Existing stemming methods	1. The existing Indonesian stemming methods are still oriented towards Indonesian formal sentences by normalizing Indonesian non-formal into formal to be used as a better stemming method.	Accuracy, Classification, Indonesian, Stemming, Text processing.
[15]	Inflectional suffixes	1. Morphological decomposition 2. Inflectional suffixes	1. French inflectional verbal suffixes ,manipulated the surface and cumulative frequencies of verbal inflected forms asked participants to perform a visual lexical decision task.	Morphological operations; Verbal inflection; Decomposition; Word recognition; Psycholinguistics.
[17]	ModSeleCT, Precision, RSeg	1. Light10 2. Segmenter	1. We integrated a semantic space, based on LSA, to enhance the semantic level of topic segmentation. For the pre-processing step, we used Light10 according to the primary results of this work.	Arabic topic segmentation, Arabic stemming algorithms, Arabc99

Evaluation of Proposed Urdu Stemmer, Comparison of proposed approach with a light weight Urdu stemmer, Evaluation of proposed short Urdu text classification approach, (Comparison of proposed short text categorization method with competitors which is compared to Naive Bayes and SVM-based Urdu text categorization algorithms [1] .Next, Optimization of the parameters. According to one illustration the value for NGT, NGTP, NGTS are all n=7. Following that, it demonstrates that the ideal B values for NGT, NGTS, and NGTP are 19, whereas NGTSP is 18. Because the number of distinct grammes with the continuation count 16 is particularly low for a lower order 6-gram, the PER increases at B = 16 [2]. A meta classifier is a classifier that uses all of the predictions as features to generate a final prediction. As a result, it takes the classes predicted by several classifiers and selects the final one as the desired outcome. Morphological Parsing: It is the procedure for determining the morphemes that make up a particular word. It needs to be able to tell the difference between morphological and orthographic rules [4]SAFAR is made up of numerous layers: In first, A set of technical services is included in the utilities layer. In second, the resources layer offers services for accessing language resources such as lexica[5] .This survey paper is also consist of analysis like which are based on the advantages and disadvantages of the proposed solution. So, the table which is given below is consists of proposed method, its advantages and disadvantages.

Table 4 : Describing about the pros and cons of the proposed methodologies

Reference	**Proposed Method**	**Advantage**	**Disadvantage**
[1]	Created infix stripping rules for introduced infix word classes	> Researchers simplified current stemming rules and established general rules that are applicable to every sort of Urdu word in this stemming work. > This proposed Urdu stemmer can create the stems of Urdu words as well as loan words from acquired languages.	> Identification of other stemming rules can be used to improve the proposed set of rules. > The proposed work can be improved in terms of coverage and efficiency, and a stemming technique can be investigated.
[2]	Phonemicization	> Grapheme-to-phoneme conversion (G2P), is a method of translating a word into its pronunciation. > And a more efficient stemmer might be used in the future to improve the NGTSP model.	_
[3]	Language-Independent Stemming	> In almost all of the languages under examination, the proposed method outperforms stemmers based on linguistic knowledge and others. > When compared to no stemming, the suggested technique increased the maximum number of inquiries	> Marathi linguistic stemmer didn't perform well as compare to other languages in the proposed work.
[4]	Multilingual text analytics and text mining	> It includes a wide range of topics, including sentiment analysis, length analysis, language identification, language translation, text categorization, and topic modeling , to name a few.	> Not practical work has been done in the proposed work.
[5]	SAFAR-stemmer (Software Architecture for Arabic language Processing)	> SAFAR-stemmer will be more suited for tasks such as Part-of-Speech tagging, parsing, and so forth. > The SAFAR stemmer increases the accuracy as well as execution time.	> Because of the lexical resource verification, proposed stemmer takes longer than Motaz stemmer and Light10.
[6]	Arabic Morphology Information Retrieval (AMIR) Stemmer	> The proposed method has a hand above due to its new features which can't be seen in the previous designed methods. > AMIR able to distinguish between the core letters of the stems.	_
[7]	Unsupervised Stemmer for Sindhi Language	> This method is giving 87% of accuracy which is better than rule-based stemmer. > Not a time consuming approach like rule-based stemmer.	• The proposed approach have less accuracy as compared to rule-based lemmatizer.
[8]	Stemming effect on text similarity for Arabic language	> This study is first of its kind to talk about the effect of stemming on semantic text similarity at sentence level for Arabic	> Some of different combination of the algorithms are not up to the

		language. > This method has shown improvement on Pearson correlation when using stemming and lemmatization algorithm in measuring similarity between Arabic text.	mark while compared with the existing ones in the values of pearson correlation.
[9]	Unsupervised Joint PoS Tagging and Stemming for Agglutinative Languages	> This joint method performs better when compared with single stemmers or models.	> No such particular disadvantage is found but some ups and downs are their which can be covered in the coming future.
[10]	Unsupervised MDL based Stemmer	> 'Morfessor 2.0' has two dictionaries as compared to 'Morfessor '. > Its complexity is O(M) for stem boundary within a word size of M where it was O(M2) in case of Morfessor model.	> PDDM stemmer was out performing Morfessor in case of highly inflected verbals.
[11]	Light and Khoja stemmers on topic identification using LDA	> The advantage of proposed solution is that DMM has been proved how to handle problem of a severe evaluate CmTLB against the studied weighting schemes by the quality of the learned topics , classification, and clustering tasks.	_
[12]	Light-based Arabic stemmer	1. The advantage of proposed solution is to propose new lists of suffixes and prefixes. 2. AMIR is used to generate/extract stems by applying a set of rules regarding the relationship among Arabic letters to find the root/stem of the respective words used as indexing terms for the text search in Arabic retrieval systems.	1. Among Arabic language the arrangement of its phrases and letters as well as the format of its meanings are more distinct than the other languages due to old list of suffixes and prefixes used in language.
[13]	Unsupervised Stemmer	> An unsupervised stemmer that uses lexical information between word pairs to group morphologically related variants words ,also improve the performance by reducing morphologically variants with same words.	_
[14]	Non-formal Affix Stemming Algorithm	> Stemming using “Incorbiz” methods can be used for various purposes including text clustering, summarizing, detecting hate speech, and other text processing applications in Indonesian with good accuracy rate.	-
[15]	Morphological decomposition and inflectional suffixes	> Tense marking used for largely exploited in Romance languages and appears to influence the processing of other morphemes in the morphological structure in word recognition.	> Sometimes not able to do decompose highly inflected word.

[16]	Text Tiling Segmentation	> It Improves text summarizing and also improve the information retrieval	> There is limitation in terms of choosing the stemming algorithm that was not based on a systematic analysis of Arabic stemming algorithms.

From the above-described table, some of the things can be noted like approximately all the proposed algorithms have some boon or ban. It also shows that where ever disadvantages are listed than there is a need to improve the proposed solutions on the basis of the various comparison methods which are listed above. The comparison which is showed above is based on the experimental behaviour of all the algorithms.

9. Conclusion and future directions

In this paper, we effectively accomplished the objective of creating a multi-purpose stemming calculation that cannot only be utilized for information retrieval tasks but also for non-traditional tasks such as text classification, sentiment analysis, inflection removal, etc. Improvement of the process and efficiency of the stemmer could be improved by including bigrams and trigram words. Our approach finds morphologically related words from the surrounding corpus using dictionary and corpus-based highlights such as lexical similarity, co-occurrence similarity, suffix pair knowledge, and common prefix length. Stemming can impair retrieval performance, but research is mixed. When effects have been discovered, the vast majority have been positive. A different stemmer from the three types of stemmers, Successor Vary, N-grams stemmers, and affix removal stemmers, could be used here. At this point of time, Multilingual environment become important area of natural language processing to proposed further researches. We can utilize different stemmers for understanding the issue to improve the capacity of an IR framework to recover more significant records by solving the issue of lexicon mismatch in queries and reports at the time of indexing and searching. And also, problems like modelling in short and long text document classification can be solved. Using stemmer like LDA, DMM, Stem-Based Approaches this method referred to as Light Stemmers, focus on removing the common prefixes and suffixes of given Arabic words and also improve modelling in short and long text document. This method related to Arabic STS on sentence level have some of the ups and downs but these can be figured out if an algorithm is chosen after having an insight through the problem. In alternatives of this proposed solution, we can use rule-based lemmatizer because we have seen above that lemmatizer was giving better accuracy of around 90% as compared to proposed approach of Unsupervised method. New list of suffixes and prefixes are used in Arabic light-based stemmer. This improved light-based Arabic stemmer focuses on finding and removing the infix patterns under many rules on length words and according to a specific order of the stages of the stemming to remove suffixes and prefixes from words to make easy words. The joint PoS tagging and stemming algorithm is performing better when compared to its fellow algorithms. This algorithm is well defined for mainly agglutinative languages in which highly inflection can be seen which make it difficult to find stem of the given text. An unsupervised stemmer that uses lexical data between word sets to group morphologically related variants words, also improve the performance by reducing morphologically variants with same words. The corpus analysis-based highlights such as cooccurrence recurrence help to conflate all such variations. In future we use unsupervised statistical models that reduce the morphological variants and also more suitable approach for highly inflected languages is unsupervised morphological segmentation. Finally, we can use an unsupervised statistical method for detecting stems given a large set of words. This model has been found to have higher accuracy and less variants in detecting correct stems with less time complexity. For the coming future, there are some sort of limitations i.e., under stemming and over stemming which needs to be covered and in future some other unsupervised algorithm can also be used in order to improve accuracy by applying some lemmatization rules, as we have seen lemmatizer has given better yield as compared to others. Irregular words can be considered as a next target in the coming future. Transformations between the words is our aim to work on, so that it will help us to handle irregular words. High-order NLP task like text categorization can also be explored for an external evaluation. In future, new algorithms can be seen regarding to this topic with better understanding. Our proposed solution can also be upgraded by working on more interesting datasets in order to check and improve its performance across different languages of rich morphology. There is potential to expand the lexicon resource with more words in the future to cover more stems and improve outcomes. Evaluating new stemmers to highlight the limitations and merits of the majority of existing Arabic stemmers so that researchers may choose which ones to utilize in their projects. In future, we will be working on the current study presented an improved light-based Arabic stemmer called Delight by proposing the appropriate list of suffixes and prefixes. This list of suffixes and prefixes is supported by the rules according to ’word’ length (without using a morpheme or patterns on a stem). This improved light-based Arabic stemmer focuses on finding and removing the infix patterns under many rules on length words and according to a specific order of the stages of the stemming to remove suffixes and prefixes from words to make easy words. It is noticeable that even after applying Arabic stemmers, Arabic documents have a large vocabulary size, especially for long documents. We plan to develop a feature selection method that can reduce the feature space of the Arabic corpus without affecting the performance of topic models. We implemented each version with all the studied term weightings using LDA and DMM. And it is noticeable that topics learned by DMM-CmTLB are cleaner and more coherent. We also plan to develop a feature selection method that can reduce the feature space of the Arabic corpus without affecting the performance of topic models.

10. Reference

[1] Ali, M., Khalid, S., & Aslam, M. H. (2017). Pattern based comprehensive Urdu stemmer and short text classification. IEEE Access, 6, 7374-7389.

[2] Suyanto, S., Sunyoto, A., Ismail, R. N., Rachmawati, E., & Maharani, W. (2021). Stemmer and phonotactic rules to improve n-gram tagger-based indonesian phonemicization. Journal of King Saud University-Computer and Information Sciences.

[3] Alotaibi, F. S., & Gupta, V. (2018). A cognitive inspired unsupervised language-independent text stemmer for Information retrieval. Cognitive Systems Research, 52, 291-300.

[4] Khan, A. M., & Afreen, K. R. (2021). An approach to text analytics and text mining in multilingual natural language processing. Materials Today: Proceedings.

[5] Jaafar, Y., Namly, D., Bouzoubaa, K., & Yousfi, A. (2017). Enhancing Arabic stemming process using resources and benchmarking tools. Journal of King Saud University-Computer and Information Sciences, 29(2), 164-170.

[6] Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. Egyptian Informatics Journal, 21(4), 209-217.

[7] Nathani, B., Joshi, N., & Purohit, G. N. (2020). Design and development of unsupervised Stemmer for Sindhi language. Procedia Computer Science, 167, 1920-1927.

[8] Alhawarat, M. O., Abdeljaber, H., & Hilal, A. (2021). Effect of stemming on text similarity for Arabic language at sentence level. PeerJ Computer Science, 7, e530.

[9] Bölücü, N., & Can, B. (2019). Unsupervised joint PoS tagging and stemming for agglutinative languages. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3), 1-21.

[10] Özbey, C., & Karcili, A. (2021, October). Unsupervised Lexicon-Based Stemming by Dual Dictionary Models. In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.

[11] Ma, T., Al-Sabri, R., Zhang, L., Marah, B., & Al-Nabhan, N. (2020). The impact of weighting schemes and stemming process on topic modeling of arabic long and short texts. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(6), 1-23.

[12] Alshalabi, H., Tiun, S., Omar, N., AL-Aswadi, F. N., & Alezabi, K. A. (2021). Arabic light-based stemmer using new rules. Journal of King Saud University-Computer and Information Sciences.

[13] Singh, J., & Gupta, V. (2019). A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. Knowledge-Based Systems, 180, 147-162.

[14] Rianto, R., Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the Accuracy of Text Classification using Stemming Method, A Case of Non-formal Indonesian Conversation.

[15] Estivalet, G. L., & Meunier, F. (2020). Morphological operations in French verbal inflection: Automatic, atomic, and obligatory. Lingua, 240, 102839.

[16] Saharia, N., Sharma, U., & Kalita, J. (2014). Stemming resource-poor Indian languages. ACM Transactions on Asian Language Information Processing (TALIP), 13(3), 1-26.

[17] Abuaiadah, D. (2016). Using bisect k-means clustering technique in the analysis of Arabic documents. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 15(3), 1-13.

[18] Angelelli, P., Marinelli, C. V., & Burani, C. (2014). The effect of morphology on spelling and reading accuracy: a study on Italian children. Frontiers in psychology, 5, 1373.

[19] Baker, M. (1985). The mirror principle and morphosyntactic explanation. Linguistic inquiry, 16(3), 373-415.

[20] Granlund, S., Kolak, J., Vihman, V., Engelmann, F., Lieven, E. V., Pine, J. M., ... & Ambridge, B. (2019). Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: A crosslinguistic elicited-production study of Polish, Finnish and Estonian. Journal of Memory and Language, 107, 169-194.

[21] Porter, M. F. (1980). An algorithm for suffix stripping. Program.

[22] Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013, April). Twitter news classification using SVM. In 2013 8th International Conference on Computer Science & Education (pp. 287-291). IEEE.

[23] Krishnakumar, A. (2006). TEXT CATEGORIZATION Building a KNN classifier for the Reuters-21578 collection. Department of Computer Science.

[24] Khamar, K. (2013). Short text classification using kNN based on distance function. International Journal of Advanced Research in Computer and Communication Engineering, 2(4), 1916-1919.

[25] Achanta, S., Pandey, A., & Gangashetty, S. V. (2016, July). Analysis of sequence-to-sequence neural networks on grapheme to phoneme conversion task. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 2798-2804). IEEE.

[26] Adriana, M., Asian, J., Nazief, B., Tahaghoghi, S. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. ACM Transactions on Asian Language Information Processing (TALIP), 6(4), 1-33.

[27] Emiru, E. D., Li, Y., Xiong, S., & Fesseha, A. (2019, November). Speech recognition system based on deep neural network acoustic modelling for low resourced language-Amharic. In Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering (pp. 141-145).

[28] Hlaing, A. M., & Pa, W. P. (2019, October). Sequence-to-Sequence Models for Grapheme to Phoneme Conversion on Large Myanmar Pronunciation Dictionary. In 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) (pp. 1-5). IEEE.

[29] Liu, L., Finch, A., Utiyama, M., & Sumita, E. (2020). Agreement on target-bidirectional recurrent neural networks for sequence-to-sequence learning. Journal of Artificial Intelligence Research, 67, 581-606.

[30] Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4), 357-389.

[31] Bacchin, M., Ferro, N., & Melucci, M. (2002, December). The effiectiveness of a graph-based algorithm for stemming. In International Conference on Asian Digital Libraries (pp. 117-128). Springer, Berlin, Heidelberg.

[32] Chavula, C., & Suleman, H. (2017). Morphological cluster induction of Bantu words using a weighted similarity measure. In Proceedings of the South African Institute of Computer Scientists and Information Technologists (pp. 1-9).

[33] Dolamic, L., & Savoy, J. (2010). Comparative study of indexing and search strategies for the Hindi, Marathi, and Bengali languages. ACM Transactions on Asian Language Information Processing (TALIP), 9(3), 1-24.

[34] Singh, J., & Gupta, V. (2017). An efficient corpus-based stemmer. Cognitive Computation, 9(5), 671-688.

[35] Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

[36] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

[37] Bikel, D., & Zitouni, I. (2012). Multilingual natural language processing applications: from theory to practice. IBM Press.

[38] Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 28(1), 56-75

[39] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analysing unstructured data. Cambridge university press.

[40] Aljlayl, M., & Frieder, O. (2002, November). On Arabic search: improving the retrieval effectiveness via a light stemming approach. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 340-347).

[41] Al-Kabi, M., & Al-Mustafa, R. (2006). Arabic root-based stemmer. In proceedings of the international Arab conference on information technology, Jordan.

[42] Chen, A., & Gey, F. C. (2002, November). Building an Arabic Stemmer for Information Retrieval. In TREC (Vol. 2002, pp. 631-639).

[43] Dukes, K., & Habash, N. (2010, May). Morphological annotation of quranic Arabic. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

[44] Jaafar, Y., & Bouzoubaa, K. (2015, April). Arabic natural language processing from software engineering to complex pipeline. In 2015 First International Conference on Arabic Computational Linguistics (ACLing) (pp. 29-36). IEEE.

[45] Khoja, S., & Garside, R. (1999). Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University.

[46] Mustafa, M., Aldeen, A. S., Zidan, M. E., Ahmed, R. E., & Eltigani, Y. (2019). Developing two different novel techniques for Arabic text stemming. Intelligent Information Management, 11(01), 1.

[47] Azman, B. (2019). Root identification tool for Arabic verbs. IEEE Access, 7, 45866-45871.

[48] Kanaan, G., Al-Shalabi, R., Ababneh, M., & Al-Nobani, A. (2008, December). Building an effective rule-based light stemmer for Arabic language to inprove search effectiveness. In 2008 International Conference on Innovations in Information Technology (pp. 312-316). IEEE.

[49] Majgaonker, M. M., & Siddiqui, T. J. Discovering suffixes: A Case Study for Marathi.

[50] Krishn, A., Guha, R. S., & Mukherjee, A. (2012). Unsupervised Morphological Analysis of Hindi.

[51] Bhat, S. (2012, December). Morpheme segmentation for kannada standing on the shoulder of giants. In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (pp. 79-94).

[52] Motlani, R., Tyers, F., & Sharma, D. M. (2016, May). A finite-state morphological analyser for Sindhi. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 2572-2577).

[53] Nathani, B., Joshi, N., & Purohit, G. N. (2019). Design and development of lemmatizer for Sindhi language in devanagri script. Journal of Statistics and Management Systems, 22(4), 635- 641.

[54] Alhaj, Y. A., Xiang, J., Zhao, D., Al-Qaness, M. A., Abd Elaziz, M., & Dahou, A. (2019). A study of the effects of stemming strategies on arabic document classification. IEEE Access, 7, 32664-32671.

[55] Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. Egyptian Informatics Journal, 21(4), 209-217.

[56] Al-Ramahi, M. A., & Mustafa, S. H. (2012). N-gram-based techniques for arabic text document matching; case study: courses accreditation. Abhath Al-Yarmouk. Basic Sciences and Engineering, 21(1), 85-105.

[57] Meng, F., Lu, W., Zhang, Y., Cheng, J., Du, Y., & Han, S. (2017, August). Qlut at semeval-2017 task 1: semantic textual similarity based on word embeddings. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 150-153).

[58] Zeroual, I., & Lakhouaja, A. (2017, April). Arabic information retrieval: Stemming or lemmatization?. In 2017 Intelligent Systems and Computer Vision (ISCV) (pp. 1-6). IEEE.

[59] Merialdo, B. (1994). Tagging English text with a probabilistic model. Computational linguistics, 20(2), 155-171.

[60] Paik, J. H., Mitra, M., Parui, S. K., & Järvelin, K. (2011). GRAS: An effective and efficient stemming algorithm for information retrieval. ACM Transactions on Information Systems (TOIS), 29(4), 1-24.

[61] Peng, F., Ahmed, N., Li, X., & Lu, Y. (2007, July). Context sensitive stemming for web search. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 639-646).

[62] Melucci, M., & Orio, N. (2003, November). A novel method for stemmer generation based on hidden Markov models. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 131-138).

[63] Goldwater, S., & Griffiths, T. (2007, June). A fully Bayesian approach to unsupervised part-ofspeech tagging. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 744-751).

[64] Sever, H., & Bitirim, Y. (2003, October). FindStem: Analysis and evaluation of a Turkish stemming algorithm. In International Symposium on String Processing and Information Retrieval (pp. 238-251). Springer, Berlin, Heidelberg

[65] Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. Artificial Intelligence Review, 48(2), 157-217.

[66] Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing (TSLP), 4(1), 1- 34.

[67] Beesley, K. R. (1996). Arabic finite-state morphological analysis and generation. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

[68] Desai, N., & Dalwadi, B. (2016, March). An affix removal stemmer for Gujarati text. In 2016 3rd international conference on computing for sustainable global development (INDIACom) (pp. 2296-2299). IEEE.

[69] Yin, J., & Wang, J. (2014, August). A dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 233-242).

[70] Wilson, A., & Chew, P. A. (2010, June). Term weighting schemes for latent dirichlet allocation. In human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 465-473).

[71] Abuaiadah, D., El Sana, J., & Abusalah, W. (2014). On the impact of dataset characteristics on arabic document classification. International Journal of Computer Applications, 101(7).

[72] Li, X., Zhang, J., & Ouyang, J. (2019, July). Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7884-7891).

[73] Wang, N., Wang, P., & Zhang, B. (2010, June). An improved TF-IDF weights function based on information theory. In 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (Vol. 3, pp. 439-441). IEEE.

[74] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002, August). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 275-282).

[75] Hajjar, M., & Zreik, K. (2010, May). A system for evaluation of Arabic root extraction methods. In 2010 Fifth International Conference on Internet and Web Applications and Services (pp. 506-512). IEEE

[76] Saad, M. K., & Ashour, W. M. (2010). Arabic morphological tools for text mining. In Corpora, 6th ArchEng International Symposiums, EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science (Vol. 18).

[77] Alhaj, Y. A., Al-qaness, M. A., Dahou, A., Abd Elaziz, M., Zhao, D., & Xiang, J. (2020). Effects of light stemming on feature extraction and selection for arabic documents classification. In Recent Advances in NLP: The Case of Arabic Language (pp. 59-79). Springer, Cham.

[78] Creutz, M., & Lagus, K. (2005, June). Inducing the morphological lexicon of a natural language from unannotated text. In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05) (Vol. 1, No. 106-113, pp. 51- 59).

[79] Creutz, M., & Lagus, K. (2004, July). Induction of a simple morphology for highly-inflecting languages. In Proceedings of the 7th meeting of the acl special interest group in computational phonology: Current themes in computational phonology and morphology (pp. 43-51).

[80] Baroni, M., Matiasek, J., & Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. arXiv preprint cs/0205006

[81] Dawson, J. (1974). Suffix removal and word conflation. ALLC bulletin, 2(3), 33-46

[82] Paik, J. H., Pal, D., & Parui, S. K. (2011, July). A novel corpus-based stemming algorithm using cooccurrence statistics. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 863-872).

[83] Utami, E., Hartanto, A. D., Adi, S., Putra, R. B. S., & Raharjo, S. (2019, August). Formal and non-formal Indonesian word usage frequency in twitter profile using non-formal affix rule. In 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS) (Vol. 1, pp. 173-176). IEEE.

[84] Putra, R. B. S., & Utami, E. (2018, March). Non-formal affixed word stemming in Indonesian language. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 531-536). IEEE.

[85] Putra, R. B. S., Utami, E., & Raharjo, S. (2019, July). Accuracy measurement on Indonesian non-formal affixed word stemming with Levenhstein. In 2019 International Conference on Information and Communications Technology (ICOIACT) (pp. 486-490). IEEE.

[86] Yuwana, R. S., Suryawati, E., & Pardede, H. F. (2018, November). On Empirical Evaluation of Deep Architectures for Indonesian POS Tagging Problem. In 2018 International Conference on Computer, Control, Informatics and Its Applications (IC3INA) (pp. 204-208). IEEE.

[87] Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. Journal of Big Data, 8(1), 1-16.

[88] Saharia N, et al (2013) An improved stemming approach using HMM for a highly inflectional language. In: Gelbukh A (eds) Computational linguistics and intelligent text processing, CICLing, 2013, Lecture Notes in Computer Science.

[89] Rahman M, Sarma SK (2016) Analysing morphology of Assamese words using finite state transducer. Int J Innov Res Comput Commun Eng 4(12):21801–21807

[90] Tabassum T, et al (2016) A corpus based unsupervised Bangla word stemming using N-gram language model. In: International Conference on informatics, electronics and vision (ICIEV).

[91] A. El-Shayeb, S. R. El-Beltagy, and A. Rafea. 2007. Comparative analysis of different text segmentation algorithms on arabic news stories. In Proceedings of the IEEE International Conference on Information Reuse and Integration. 441–446.

[92] Brants, F. Chen, and A. Farahat. 2002. Arabic document topic analysis. In Proceedings of theWorkshop on Arabic Language Resources and Evaluation (LREC'02).

[93] Abu-Rabia, S., Share, D., & Mansour, M. S. (2003). Word recognition and basic cognitive processes among reading-disabled and normal readers in Arabic. Reading and writing, 16(5), 423-442.

[94] Saiegh-Haddad, E., & Joshi, R. M. (Eds.). (2014). Handbook of Arabic literacy: Insights and perspectives (Vol. 9). Springer Science & Business Media.

[95] Culbertson, J., & Adger, D. (2014). Language learners’ privilege structured meaning over surface frequency. Proceedings of the National Academy of Sciences, 111(16), 5842- 5847.

[96] Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., ... & Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. First Language, 31(4), 461-479.

[97] Jabbar, A., ul Islam, S., Hussain, S., Akhunzada, A., & Ilahi, M. (2019). A comparative review of Urdu stemmers: Approaches and challenges. *Computer Science Review*, *34*, 100195.

[98] Steinberger, R. (2012). A survey of methods to ease the development of highly multilingual text mining applications. Language resources and evaluation, 46(2), 155-176.

[99] Jabbar, A., Iqbal, S., Khan, M. U. G., & Hussain, S. (2018). A survey on Urdu and Urdu like language stemmers and stemming techniques. Artificial Intelligence Review, 49(3), 339-373.

[100] Lakshmi, R. V., & Kumar, S. B. R. (2014). Literature review: stemming algorithms for Indian and Non-Indian languages. Int J Adv Res Comput Sci Technol, 2(3), 349-352.table.