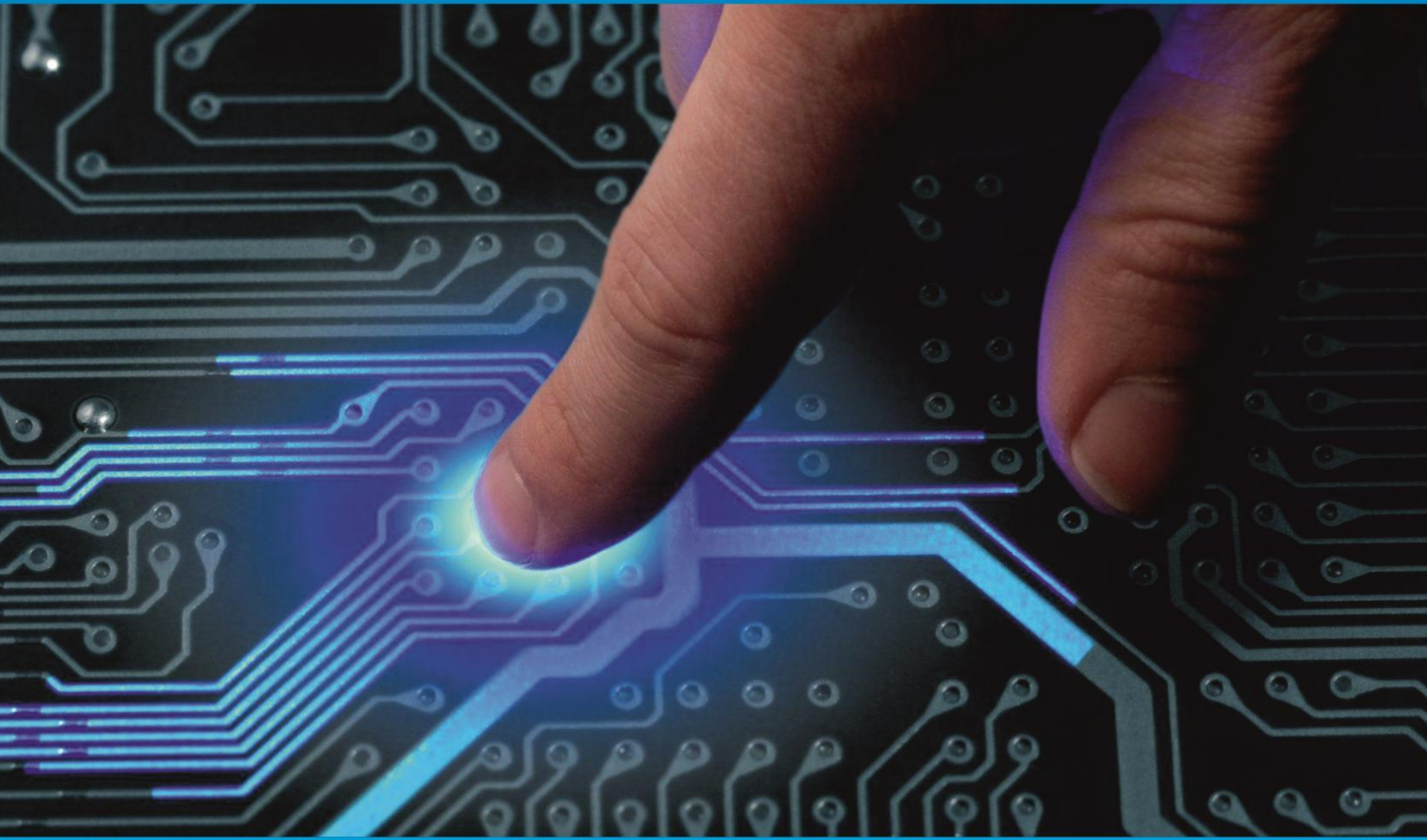
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 10, October 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 10, October 2020 ||

Word Embedding Generation Methods and Tools: A Critical Review

Sajadul Hassan Kumhar[1*], Mudasir M Kirmani[2], Jitendra Sheetlani[3], Mudasir Hassan[4]

Research Scholar, Department of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, (M.P)), India[1*]

Assistant Professor, Department of Computer Science, SKUAST-Kashmir, J&K, India[2]

Professor, Department of Computer Science, Sri Satya Sai University of Technology & Medical Sciences (M.P), India[3]

Research Scholar, Centre of Central Asian Studies University of Kashmir, Srinagar (J&K), India[4]

ABSTRACT: Natural language processing (NLP)and Computer Linguistics are the sub-field of artificial Intelligence (AI)and data science. Humans communicate using natural languages as a means of communication. The natural language processing is used to process these natural languages using computer algorithms and programs for speech recognition, natural language understanding and natural language generations etc. with the advent of technology people used to express and share their ideas and feelings on social media. Social media scientists, computer linguistics and natural language programmers contributed their services to analyses the textavailable on these social media sites. People in large countries like India tend to share this information on these social media sites using the regional language with English as a result the mixed multilingual text. In Indian Sub-Content about 61 million people tend to express their ideas and feelings using mixed Urdu, Roman Urdu and English languages while communication on these social media sites. The research paper is closely related to corpus collection, language detection, embedding generation, translation and transliteration and sentiment analysis of multilingual text which includes Urdu, Roman Urdu and English.In the research paper relevant tools and methods are presented which will investigate into the details of language detection in multilingual dataset and generation of embeddings on it for sentiment analysis.

KEYWORDS: NLP, AI, Social Media, Computer Linguistics, Multilingual, Roman Urdu, Urdu, English, Corpus, Translation, Transliteration, Word Embeddings, Sentiment analysis.

I. OVERVIEW

Natural language processingis becoming more popular trend in computer linguistic research. Yet we know little about it properties and performance. The Natural language processing is a sub field of artificial intelligence and data science.The natural language processing has been carried out on multilingual corpus and different solution have been proposed in different research works, but detection of Urdu language remained an unsung success in the field of natural language processing, so in order to solve the problem of Urdu word detection our research work proposed a solution to the problem of detection of multilingual languages along with the generation of embedding to Urdu word is carried out in the research work. The sentiment on Urdu corpus will be also carried out in our research work.

II. MOTIVATION FOR TOPIC NAME

Humans have the ability to communicate which other animals don’t have, being able to communicate humans use natural language as a means of communication for sharing or expressing of opinions, feelings, ideas or information. With the emergence of several social media platforms people tend to share and express their opinions, ideas, feelings or interests through them. In early days’ people tend to share or express their feelings using English only as a medium of communication on these platforms. Today people tend to share their ideas, feelings, opinions using a mixture of languages on these social media platforms. The process of sharing a mixture of multiple language while expressing their feelings is called multilingual language.Urdu is a natural language roughly spoken by 163 million speakers worldwide, Millions of Urdu speaking people sharing information through social media platforms, mostly in Indian sub-continent English and Urdu mixture texts are used to express their opinions and feelings on these social media platforms. Corpus is a collection of large written text mostly written by an author or body of writing about a particular subject. Natural language processing is a technique which is used to process these natural languages by means of computer programs and algorithms for speech recognition, natural language understanding and natural language generations. Natural language processing (NLP) is a very interesting and important topic of Artificial Intelligence where a machine is trained to understand and process the text to make human computer interactions more efficient. Application of natural language processing lies under several fields like machine translation, text processing, and entity extraction and so on. Certainly a need raised in the field of computational linguistics to make the advantage of the computer and computer programs for simplified processing of these languages. The example of mixed multiple language is ***Maray dousth***, if we observe carefully the above text, it will be ambiguous i.e. when a non-Urdu user reads the above word “Maray” he gets confused whether the word is English or Urdu. In this scenario the identification of language used for each word is important for language processing. Lot of work have been done in the field of computer linguistics for detection of the languages in multilingual corpus but no concreate method have been given for detection of Urdu texts in multilingual textual corpus. Furthermore, different researcher has proposed embeddings of text but no evidence supports to the Urdu text. Though sentiment analysis has been performed on the texts available on the social media platforms but no concreate method have been proposed for sentiment analysis on English-Urdu multilingual texts. Thus the research work proposed will lead a beginning to detect the English-Urdu mixture of multilingual text in the corpus, and generation of embeddings of multilingual text for sentiment analysis.

III. RELATED RESEARCH

In literature a number of corpus collection, segmentation and tagging, language identification and detection, translation and transliteration of texts, distributed representation of text into its vector form and sentiment analysis methods and tools has been developed. The efforts have been attempted for research study which directly supports to the study field.To develop large scale freely available standard evaluation resources and to investigate corpus collection from social media sites is a non-trivial task. In the previous literature efforts has been made to develop a bench mark for corpus collection from social media. we will discuss the most relevant and specific studies relevant to my studies. We will present only most prominent and relevant studies. In previous literature efforts have been made to develop corpus for English and other language such as SEMCOR corpus, Google corpus and DutchSemCor corpus. Later in 2001 S. Hussain and M. Afzal [34] released the Urdu software with the phonetic keyboard in the Unicode format also known as “Urdu Zabtie Takhtie”. The “Urdu Zabtie Takhtie”, is a corpus of Urdu text available to carry out research on natural language processing however released urdu software “Urdu Takhtie” which is not fully capable to represent the “Urdu Takhtie” in UNICODE format therefore a standardization need to be expanded to represent Urdu completely in Unicode format including fonts and Urdu letters.**W. J. Teahan (2000) [44]** in their research work proposed a prediction by partial match (PPM) model to predict the next character in the input sequence by using text compression, text classification and text Segmentation techniques. Furthermore, the classification of text is achieved by minimum cross-entropy method and the segmentation of the words is compounded by text correction algorithms.The proposed research model has a weakness to understand the text analyzed for predicting the next character from the previous input sequence of characters and predicts inaccurate syntactic parser from raw text database. **S. Hussain and M. Afzal (2001) [34]** proposed a model for automatically inducing standalone monolingual part of speech tagger, base noun-phrase bracketers, named entity tagger and morphological analyzer for random foreign language such as French, Chines, Czech and Spanish. The proposed model has not used non-English dictionary of roots in the computation of part of speech tagging which has resulted in low performance on the language under observation such as Czech with very low word-alignments. **D. Yarowsky et al (2001) [12]**proposed a model for automatically inducing standalone monolingual part of speech tagger, base noun-phrase bracketers, named entity tagger and morphological analyzer for random foreign language such as French, Chines, Czech and Spanish. The proposed research model has not used non-English dictionary of roots in the computation of part of speech tagging which has resulted in low performance on the language under observation such as Czech with very low word-alignments. **L. Finkelstein et** al**(2001) [16]** proposed Intellizap model based on statistical semantic network with WordNet having three components, one extracting keywords from the captured text and context, second performs high level classification of query to a small set of predefined domains and re-ranking the results obtained from different search engines system for processing queries in their context to open up a new and promising avenue for information retrieval. However, the generic and tailoring approach of Intellizap engine shown has not maximized the context-guided capabilities of individual search engines which has resulted in inefficiency of work on low speed internet connectivity.**D.Becker and K. Riaz (2002) [8]** released freely available Urdu dataset to promote natural language processing research activities in Urdu, they proposed

Urdu linguistic resource such as part of speech tagging and named entity, the research carried out on Urdu text corpus of news articles of BBC Urdu URL. Their proposed model has not carried out much research work on natural language processing on the freely released urdu corpus. **Y**. **Bengio et** al **(2003) [47]**proposed a statistical character-based text compression language model using neural network to learn joint probability function of word sequence along with distribution representation of each word. The proposed research model has not performed well on low speed training and recognizing of a neural network. **T. Pham and D. Tran (2003) [42]**proposed a new approach for written language identification based on machine learning and computational algorithm for machine classification using the method of vector quantization and design codebook for each language which contains a predetermined number of code-vectors. The proposed research model decreased the performance of classification with more use of n-gram frequencies along with the documents of shorter length. **P. Resnik and N. A. Smith (2003) [31]**proposed the parallel corpus or bi-text model for mining the world wide web to extract the parallel text that it contains and layout stranded architecture by training supervised classifier using structural parameters for detection translations. The proposed research model has less impact on the community due to availability of few languages with too little data along with difficulty in dissemination.**Y. Al-Ohali et** al **(2003) [45]** proposed a model to developed a database to recognizing hand written Arabic cheques by using the steps of segmentation binarization, data tagging and validation of the tagging processes. In the proposed research model the research and observations have not been carried out on all the experimental setups.**Y. Qu et al (2003)**[49] proposed a model to automatically generate transliteration from an English lexicon to Japanese Katakana sequence by a set of probability mapping rules and phonetic English dictionary in addition the validation is achieved by monolingual Japanese corpus. In the proposed research model the extracted English-katakana pairs has not increased existing bilingual translation brilliant for applications such as cross language information retrieval and Chinese text.**V. Paola et**al **(2003) [43]** proposed a statistical model to convert the phonemic string into the orthography for transliterating English names using Chinese characters. They further use translating system extrinsically for cross-lingual spoken document retrieval by using English text queries to retrieve Mandarin Chinese audio from the topic Detection and tracking Corpus. In the proposed research model the intrinsic evaluation of transliteration, the system fails to perform better when training rate is increased and the performance is below than the other system and needs further investigation. **T. Korenius et** al **(2004) [40]**proposed amodel that performed normalization throughtwo different processes. First stemming was used based on Porter Stemmer, Next dictionary based lemmatization was used for transformation of words into their basic morphological form. **R. D. Lins and P. Goncalves. (2004) [32]**proposed research model to identifies the language in written texts of English, French, Portuguese and Spanish in time and space-efficient way in addition user's browser is incorporated via proxy to perform English to Portuguese translation with Word classes such as Adverbs, Articles, Conjunctions, Interjections, Numerals, Prepositions and Pronouns for recognition of written texts. The proposed research model doesn't provide good results on the class Article on Spanish language as it translates them into French and also the model showing worse results on Conjunction testing.**B. Martins and M. J. Silva. (2005) [6]** proposed a model to automatically identify the language of web pages by using n-gram model complemented with heuristics and a new similarity measure along with the Good-Turing smoothing approach to smoothing to rare character sequence present in the web page texts. The proposed research model has inability to discriminate Portuguese documents from Brazilian and on small test the system achieves poor precision in addition the model ignored resources from .BR domain. **S. Malikand S. A. Khan(2005) [36]**proposed a model to recognize online handwriting texts to coverts it into urdu text by using analytical approach for feature extraction along with rule based analysis for slant removal and tree based dictionary search for classification. The proposed research model has not improved the slant removal method in addition no hat feature has been used to recognize compound character to make the system complete online handwriting recognition system.**S.-M. Kim and E. Hovy (2006) [38]**proposed a system for sentiment analysis using the maximum-entropy model to train the results to subsequently extracts opinions and developed a mapping of subjective lexicon to other languages by using machine translation system and subjective analysis system for English to other languages. The model proposed don't provide better results for sentiment analysis on debates about political and social agendas for opinion mining and is also not able to predict better results on unannotated data set. **P. Koehn et al (2007)**[30] proposed a phrase based machine translation toolkit Moses model along with confusion network to translate human spoken sentences which consist of all the components needed to preprocess data, train the language models and the translation models. The proposed research quantized language model produces less accurate results for translations. **M. G. A. Malik et al(2008) [23]**Proposed a Hindi Urdu machine transliteration model using Finite State Transducer along with the universal intermediate transcription (UIT) is used for translation to close-surface languages on the basis of their common phonetic repository. The proposed research model has not enhanced cross-scriptural transliteration and Machine translation.**M.Potthast et al(2008) [25]**proposed a Wikipedia based CL-ESA multilingual retrieval model for cross-language similarity along with the document d written in language L to access it similarity on another document D written in that Language $L^{/}$. The system deteriorates retrieval quality at lower dimensions and the results produced based on geography, location or linguistic relations on Wikipedia languages is low on the number shared concept.**Z. Ceskaet al (2008) [50]** proposed MLPlag a novel model for multilingual plagiarism detection. The MLPlag analysis word position and utilizes the Euro WordNet thesaurus which transforms words into independent form and further uses the semantic based word normalization to enhance plagiarism detection and identify the replacement synonyms used by plagiaries to hide the document match. In the proposed research model the EWN thesaurus proposed is still under development and will incorrectly detect texts in cross-lingual plagiarism detection also word sense disambiguation have not been carried out in the model. **L. Lingjun et al (2008) [17]** The research work proposed a model to detect word shift stenographic by using machine classifying along with SVM. The documents are divided into two classes "Soft document" and "hard document" by establishing two level detection. The proposed model hasn't safe word shifts. **A. Malik and L. Besacier et al (2009) [2]**proposed a novel hybrid model for Urdu to Hindi transliteration by developing a finite state machine with statistical word language model based approach to handle in case of omission diacritical marks from the input of Urdu text. The proposed research model has low performance on the transducer-only approach especially when diacritic marks are present**. S. Mukund and R. Srihari (2010) [39]**Developed an NLP infrastructure for Urdu that is customizable and capable of providing basic analysis on which more advanced information extraction tools can be built. This system assimilates resources from various online sources to facilitate improved named entity tagging and Urdu-to-English transliteration. The research model proposed hasn't cross lingual search and machine translation. The accuracy for transliteration for Hindi is very poor while as for English WordNet it doesn't perform well approximate transliteration.**N. Durrani, and S. Hussain (2010) [27]**proposed a model for Urdu word segmentation using orthographic and linguistic feature for triggering Urdu segmentation and tokenization using the hybrid n-gram model along with rule based maximum matching heuristic. In the proposed research the bi-gram statistical natural language processing hasn't been used to merge morphemes for tagging. **S. Kanwal et al (2010) [35]**proposed Named Entity Relationship corpus that contains 926776 tokens and 99718 carefully annotated NEs using the deep learning, NN and RNN for Urdu named entity recognition and generated four word embeddings using two embedding techniques Fast Text and Word2vec on two corpora of Urdu text. The future model has not enhanced the precision by improving the accuracy of tokenization along deep learning with POS Information.**M. Faruqui et al (2011) [22]**proposed a model for unsupervised post processing to replace English non-words in the translation output to determine phonetically similar English words with Soundex algorithm so that Levenshtein distance be chosen to model the correct transliteration. The proposed research model has not included monolingual query expansion in Urdu to improve the non-NE part of the query and training a full Urdu-English transliteration system.**A. Balahur and M. Turchi (2012) [1]**proposed a machine learning model using support vector machine for sentiment analysis on French, Germen and Spanish languages using machine translation engines including google, Bing and Moses. The proposed research model has not performed well for translation which causes increase in the variance in the data and is noisy that the classifier is not able to learn correct information for positive and negative classes. **Y. Haribhakta et al (2012) [48]**proposed a novel model for detection of topics using the unsupervised learning technique furthermore the model detects and extracts the keywords from the corpus by identify the relationship between the words of unstructured data by using bigram and trigram models. The proposed model hasn't built query categorization and document retrieval system to improve the accuracy and performance.

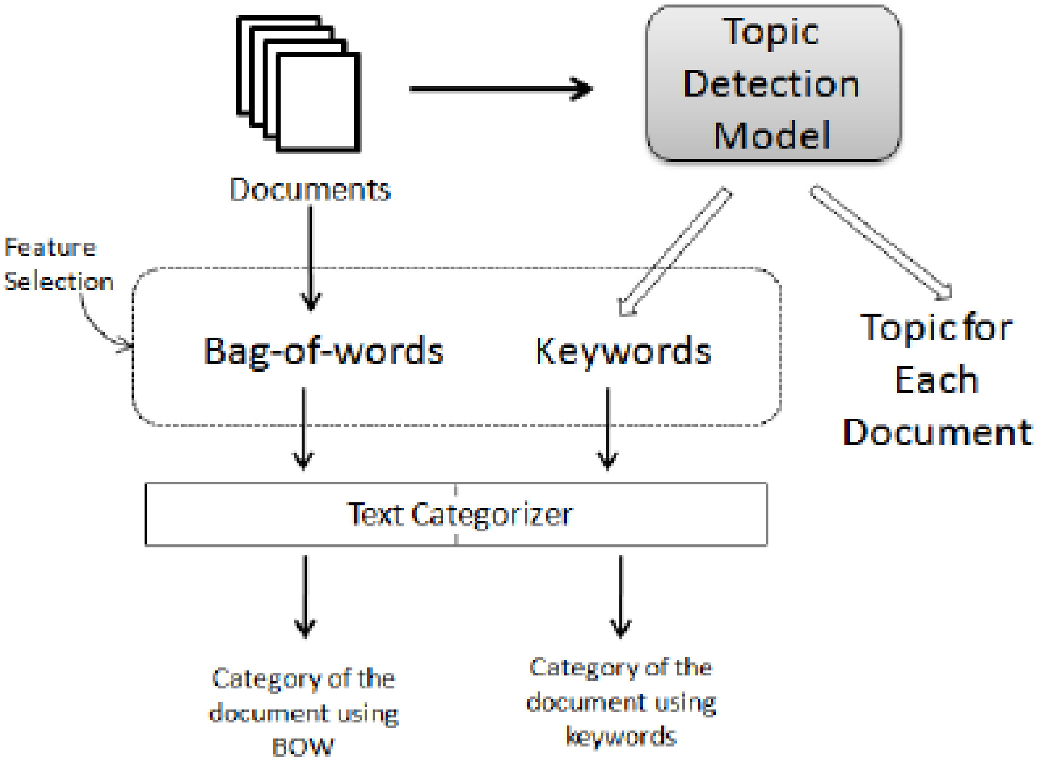


Fig. 1: Text categorization using keywords from Topic Detection Model

D. Nguyen et al (2013) [9]proposed a model for Dutch Turkish language identification at root level in multilingual corpus using the tag and dictionaries and introduces context for achieving 98% accuracy.In the proposed research model the system don't perform well with multi word in multilingual conversation data and languages that are typologically less distinct from each other or dialects. **Y. Bengio, A. Courville, and P. Vincent (2013) [46]** proposed representation learningmodel a review and new perspective to motivate long-term unanswered questions about the appropriate objectives for learning good representation, for computing representations and the geometrical connection between representation learning, density estimation and manifold learning. The proposed research model hasn't optimization on very large dataset and no clear cut idea supports the success and failure of training deep architectures both in supervised and unsupervised case.**T. Mikolov et al (2013) [41]**proposed a skip gram model to efficiently learn distributed vector representations to capture large number of syntactic and semantic word relationships along with a hierarchal softmax have been introduced to speed up and learn more word representations. The proposed research model has low performance using different choice of architectures, the size of the vector, the sub sampling rate and the size of training. **P. D. Turney (2013) [28]**Proposed the SuperSim model which is a combination functions based on supervised learning model with a unified approach to analogy relational similarity and paraphrase compositional similarity in addition the supervised learning is achieved by support vector machine. The proposed research model doesn't work beyond noun-modifier paraphrase which results in dissimilarity of sentence pairs.**G. Grigonyte et al (2014) [13]**proposed a model to identify particular documents on a web namely multilingual dictionaries moreover the model performs the mining of parallel corpora, automatic construction of bilingual dictionaries and thesaurus, automatic detection of multilingual documents along with classification, further the research is carried on eight target languages for pre-identification and Wikipedia dump. The research model proposed hasn't fully detect domain-specific dictionaries and low density languages, low dictionaries and Wikipedia articles. **M. Lui et al (2014) [24]**proposed a model for language identification in multilingual documents using supervised learning generative mixture model inspired by probabilistic model. The research model proposed has not accurately identify the off-the-shelf language with cross domain effect, and hasn't ability to identify languages of generative mixture models. **P. Gupta et al (2014)**

[29]proposed a concept of mixed script information retrieval to handle the mixed script term matching and spelling variation. The script is modelled jointly in a deep-learning architecture for handling spelling variation and transliteration mining. The proposed research model has not studies associated research avenue such as code-mixing in queries and documents and in more general setup of MSIR such as Mixed-Script Multilingual Information Retrieval (MS-MLIR). **D. Vilares et al (2015) [11]**The research work proposed polarity classification on twitter over different languages using three machine learning model monolingual model for opinion mining, monolingual pipeline with language detection and multilingual model that join the two monolingual models. The research model proposed has low performance due to the smaller presence of Spanish words in the corpus. The annotators also noticed that Spanish terms presents a larger frequency of grammatical errors than the English ones.**A. Tripathy et al (2015) [5]**proposed a model for sentiment analysis on movie reviews in English language using Machine Learning Techniques, furthermore the preprocessing of data has been carried out by removing top words, punctuation characters and numerical characters in addition a numerical matrix, TF-IDF (Term Frequency-Inverse Document Frequency) were generated using labeled polarity movie dataset where rows represent reviews and columns represent features by using Machine learning algorithms (NB, SVM) in order to train the model. The proposed research model has weaker performance for Nave Bayes and for TF-IDF on sentiment analysis than support Vector Machine. **C. K. Raghavi et al (2015) [7]**The research work proposed a question answering model on code-Mixed multilingual data based on question classification system, the question asked by the user is analyzed with question classification and infers the expected answer type, furthermore the support vector machine of learning algorithm have been used for translation, transliteration and the word identification on the multilingual data. The model proposed in the research work has not improved the accuracy in fine-grained classification.**M. Ahmed et al (2015) [19]**proposed Acoustic Modelling Using Deep Belief Network for Bangla Speech Recognition to combined with Hidden Markov Model. Mel frequency Cepstral coefficient is used to extract features from the speech data and then DBN is trained with these features vectors to calculate phoneme states after that Viterbi decoder is used to determine the resulting hidden state sequence to generate word. The DBN feature Vector is performed in two steps; generative pre-training step is used to train to network layer by layer and in second step enhance gradient is used to adjust the parameters in order to make it more accurate. The proposed research model hasn’t increased the Hopfield Neural Network information about the past to carry out the interpretations for the future.**A. Mogadala and A. Rettinger (2016) [3]**proposed BRAVE (Bilingual Paragraph Vector) learning model to learn bilingual embeddings of words from the document without word alignment either from label aligned non-parallel in text document or sentence-aligned parallel document corpora to support cross-language text classification. In the proposed research model the embedding model don’t perform well for machine translation for asymmetry languages as it couldn’t find specific equivalent English and do not learn multilingual semantic space having more label classes. **D. Nguyen et al (2016) [10]**proposed a model for automatic language identification and analysis of code-switching patterns within words by using computational methods, furthermore the code-switching text is automatically identified into more manageable smaller units called chunks using the Morfessore tool on both the majority language (Dutch) with minority language (Limburgish). The proposed research model has failed on spelling mistakes, occurrence of English words, moreover concatenated words were incorrectly identified as words containing code-switching. **M. Zampieri (2016) [26]**proposed automatic language identification by using n-gram language model, the methods designed to work on written text data in addition a couple of tools have been proposed such as TextCat and VarClass for text processing applications. In the proposed research model the short words don’tprovide information to distinguish between similar languages and the Uni-gram model don’t take context into account and lastly the author discusses nothing about fine line between languages.**M. Abdalla and G. Hirst (2017) [18]**proposed matrix translation model to infer and predict cross-lingual sentiment and computes a matrix to convert vector space of one language to another, in addition they observed that sentiment is preserved accurately even sub-para translation and also maintains fine-grained sentiment information between languages. The research model proposed hasn’t improved the accuracy of sentiment regression and isn’t able to fine-grain of cross lingual document in sentiment of words in over time of single language and prediction of next word is too poor. **M. Alam and S. U. Hussain (2017) [20]**proposed neural machine translation model that is based on neural network composed of two types, one that is responsible for input language and other part that handles the desired output language sentences. The model proposed encoder-decoder architecture. They transform Roman-Urdu to Urdu transliteration into sequence to sequence learning problem. The proposed research model hasn’t the capability to correctly predict sentences beyond word length 10.**L.-C. Yu et al (2017) [15]**proposed a word vector refinement to refine existing pre-trained word vectors using real-valued sentiment analysis intensity scores provided by sentiment lexicon furthermore the model is applied for refinement to pre training word vector word2vec and Glove. The proposed research model hasn't evaluated the method on other than SST dataset.**P.A. Chandra et al (2017)** proposed metaheuristic and optimal clusters-heads based on K-means and cuckoo search model for sentiment analysis on twitter dataset. The proposed research model hasn't improved accuracy furthermore the model hasn't improved the sarcasm and irony tweets. **Z. Sharf et al (2018) [51]**proposeda model for *Security* Neural Network Based sentiment analysis for large set of corpus in Roman-Urdu from social media sites. The corpus is cleaned, lexically normalized for standard representation of words and performs part of speech tagging for better identification. The research model proposed produces low values on standardization of words and are consequently missed by the tagger.**K.S. Nawaz et al (2018)** proposed a machine learning model based approach for Urdu Segmentation by adopting conditional random fields (CRF) to achieve Urdu word segmentation. The proposed research model hasn't provided a concreate word segmentation boundaries i.e. space insertions, omissions, and reduplication of foreign words and isn't able to produce good precision.**M.Ali etal (2018) [21]**proposed a compressive rule-based stemming model for Urdu text which has capability to generate stem to Urdu words and to lone words by simply removing the prefixes, infixes and suffixes. The model hasn't studied the efficiency by using n-gram model along with deep learning methods.**S. A. B. Andrabi and Abdul Wahid (2019) [33]**proposed sentence alignment algorithms based on character length, word length and lexical matching based two techniques and approaches statistical approach and lexical approach. The proposed model hasn't performed well for the languages that are syntactically different like English and Urdu. **A. Rafique et al (2019) [4]**proposed a model that focus on sentiment analysis of comments and opinions in Roman Urdu by using three supervised learning algorithms namely NB (Naive Bayes), LRSGD (Logistic Regression with Stochastic Gradient Descent) and SVM (Support Vector Machine). The research model proposed don't extend the dataset to more domains and no deep learning approach have been used for sentiment analysis.**S. M. Hassan et al (2019) [37]**proposed web scrapping tool for acquisition of data from different websites and uses Unigram and Bigram language models for identify long distance dependencies between sentences in a corpus. In the proposed research model the random forest hasn't yield much better results in identifying distances for similar sentences. **H. Mukherjee et al (2019) [14]** proposed a model based on deep learning for language identification for seven Indic spoken languages to visualize sound texture by using spectrogram. The proposed research model hasn't enhanced the architecture for real time audio processing for active learning. In addition, the model hasn't active voice detection to improve the performance.

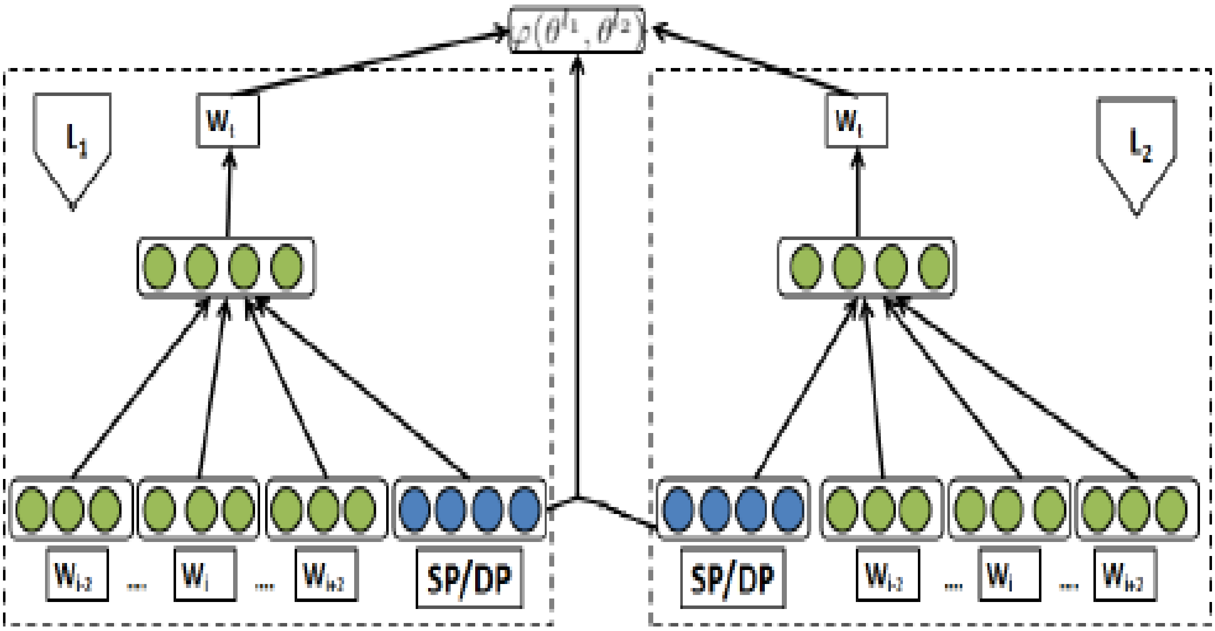


Fig. 2: Bilingual Word embeddings learned using sentence or document paragraph vectors along with word vectors

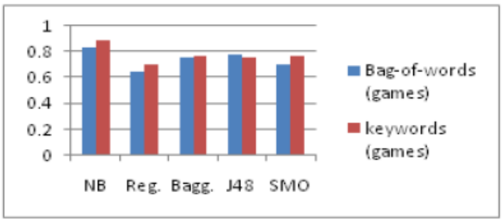


Figure 2: F-measure after stop words removal (games corpus)

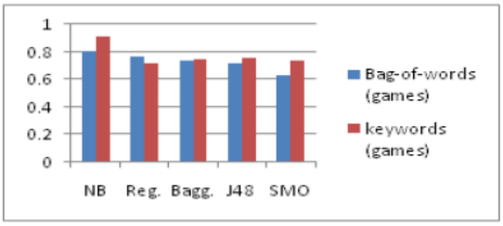


Figure 3: F-measure after stop words removal and stemming (games corpus)

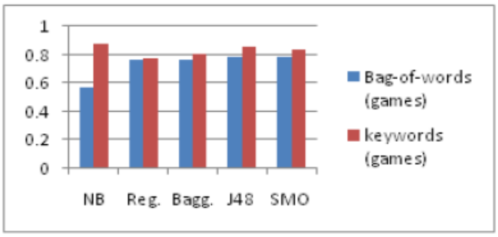


Figure 4: F-measure after stopword removal, stemming and word merging (games corpus)

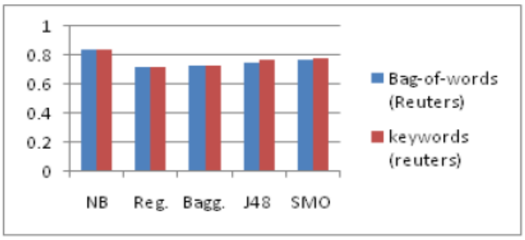


Figure 5: F-measure after stop words removal (reuters corpus)

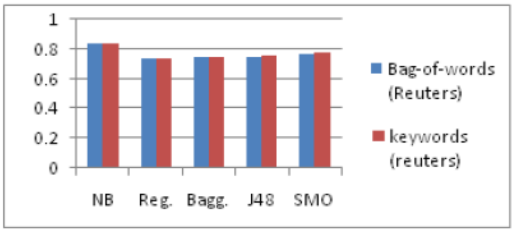


Figure 6 : F-measure after stop words removal and stemming (reuters corpus)

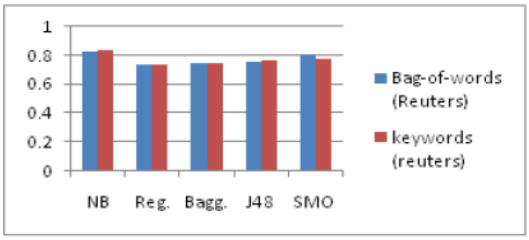
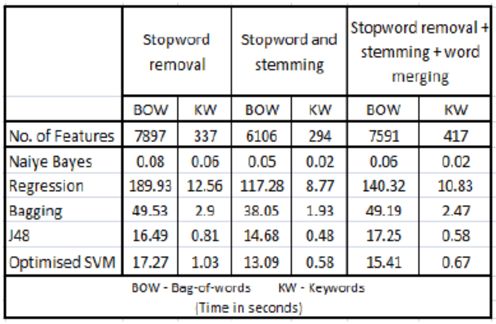


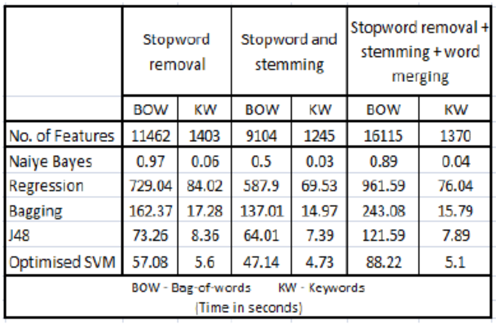
Figure 7: F-measure after stopword removal, stemming and word merging (reuters corpus)



	Stopword removal		Stopword and stemming		Stopword removal + stemming + word merging	
	BOW	KW	BOW	KW	BOW	KW
No. of Features	7897	337	6106	294	7591	417
Naiye Bayes	0.08	0.06	0.05	0.02	0.06	0.02
Regression	189.93	12.56	117.28	8.77	140.32	10.83
Bagging	49.53	2.9	38.05	1.93	49.19	2.47
J48	16.49	0.81	14.68	0.48	17.25	0.58
Optimised SVM	17.27	1.03	13.09	0.58	15.41	0.67

BOW - Bag-of-words KW - Keywords
(Time in seconds)

Figure 8: Time taken for text categorization for games corpus



	Stopword removal		Stopword and stemming		Stopword removal + stemming + word merging	
	BOW	KW	BOW	KW	BOW	KW
No. of Features	11462	1403	9104	1245	16115	1370
Naiye Bayes	0.97	0.06	0.5	0.03	0.89	0.04
Regression	729.04	84.02	587.9	69.53	961.59	76.04
Bagging	162.37	17.28	137.01	14.97	243.08	15.79
J48	73.26	8.36	64.01	7.39	121.59	7.89
Optimised SVM	57.08	5.6	47.14	4.73	88.22	5.1

BOW - Bag-of-words KW - Keywords
(Time in seconds)

Figure 9: Time taken for text categorization for reuters corpus

Comparison Table of some relevant model papers

Shumaila Malik, Shoab A. Khan	Malik, M. G. Abbas, Christian Boitet, and PushpakBhattacharyya	Abbas Malik Laurent Besacier, Christian Boitet, Pushpak Bhattacharyya	Mukund and Rohini Srihari
The research work proposed a model to recognize online handwriting texts to coverts it into Urdu text by using analytical approach for feature extraction along with rule based analysis for slant removal and tree based dictionary search for classification however The proposed research model has not improved the slant removal method in addition no hat feature have been used to recognize compound character to make the system complete online handwriting recognition system.	The research work Proposed a Hindi Urdu machine transliteration model using Finite State Transducer along with the universal intermediate transcription (UIT) is used for translation to close-surface languages on the basis of their common phonetic repository but the model has not enhanced cross-scriptural transliteration and Machine translation.	The research work proposed a novel hybrid model for Urdu to Hindi transliteration by developing a finite state machine with statistical word language model based approach to handle in case of omission diacritical marks from the input of Urdu text however The proposed research model has low performance on the transducer-only approach especially when diacritic marks are present	Developed an NLP infrastructure for Urdu that is customizable and capable of providing basic analysis on which more advanced information extraction tools can be built. This system assimilates resources from various online sources to facilitate improved named entity tagging and Urdu-to-English transliteration however The research model proposed hasn't cross lingual search and machine translation. The accuracy for transliteration for Hindi is very poor while as for English WordNet it doesn't perform well approximate transliteration.

IV. SUMMARY

The detailed overview, motivation and efforts which has been done in the previous literature relevant to corpus collection, word segmentation and tagging, word detection and identification, translation and transliteration, word embeddings generation along with sentiment analysis has been presented. The relevant tools and techniques which exists in the literature of research studies for corpus construction, development and collection from various domains have been presented. The word segmentation and tagging tools and techniques has been presented in order to have a look on the previous efforts that has been made in computer linguistic for word level segmentation along with language tagging. The most prominent, specific and relevant tools for word level language detection and identification from the corpora which has been developed in the previous research work has been presented. The relevant and most prominent tools and techniques which leads to the translation and transliteration from one language into another language and from one script of language to another script of the same language has also been investigated. The details survey on the literature available for word embeddings and sentiment analysis has been studied to further investigate the efforts that has lead in the field of research study.

V. CONCLUSION

Natural language processing is the sub-field of artificial intelligence and computer linguistics. The natural language processing is used to process natural languages that humans use for communication. With the advent of technology people tend to social media and expressed their ideas and feeling through them. However, on these social media sites people used mixed multiple languages for expressing their sentiments which resulted into the mixed multilingual corpus. In Indian sub-contents about 61 million people tend to express their ideas and feeling using Urdu, Roman Urdu and English in mixed form. This mixed multilingual information available on these social media sites posed certain challenges. In this research paper we have taken the relevant literature for collection of mixed multilingual Urdu-English corpus. then in the next step the relevant material for detection of languages in the available dataset have been taken into account. Next the review of the literature available for translation and transliteration has been investigated into details. In the next step the materials available for Urdu word embeddings generation has been investigated to further improvement in the system. In the last step relevant research literature for sentiment analysis has been done. In future the collection of mixed multilingual data of Urdu, English and Roman Urdu will be carried out on social media sitesand the data so collected will be transliterated and translated into monolingual Urdu text. Finally, the monolingual Urdu text obtained will be distributed into word vector for sentiment analysis.

REFERENCES

1. A. Balahur, and M. Turchi, “Multilingual Sentiment Analysis using Machine Translation”, In Proceedings of the 3rd Workshop Association for Computational Linguistics pages 52–60, Jeju, Republic of Korea, 12 July, 2012.
2. A. Malik, L. Besacier, C. Boitet, and P. Bhattacharyya, “A Hybrid Model for Urdu Hindi Transliteration”, In Proceedings of the 2009 Named Entities Workshop. ACL-IJCNLP 2009, pages 177–185, Suntec, Singapore, 7 August, 2009.
3. A. Mogadala, and A. Rettinger, “Bilingual Word Embeddings from Parallel and Non-Parallel Corpora for Cross-Language Text Classification”, In Proceedings of NAACL-HLT, pages 692–702, San Diego, California, June 12-17, 2017.
4. A. Rafique, M. K. Malik, Z. Nawaz, F. Bukhari, and A. H. Jalbani, “Sentiment Analysis for Roman Urdu”,Mehran University Research Journal of Engineering & Technology, Vol. 38, No. 2, Page 463-470, April 2019.
5. A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of Sentimental Reviews Using Machine Learning Techniques”, In Proceedings of 3 International Conference on Recent Trends in Computing, pp. 821 - 829, Elsevier B.V. 2015.
6. B. Martins, and M. J. Silva, “Language identification in web pages”, In Proceedings of the 2005 ACM symposium on Applied computing, pages 764–768, Santa Fe, USA. 2005
7. C. K. Raghavi, M. Chinnakotla, and M. Shrivastava, “Answer ka type kya he?” Learning to Classify Questions in Code-Mixed Language”, In Proceedings of the 24th International Conference on World Wide Web, pages 853–858. ACM, 2015.
8. D. Becker, and K. Riaz, “A study in Urdu corpus construction”, In Proceedings of the 3rd workshop on Asian language resources and international Standardization-Volume 12, 2002, pp. 1-5. 2002.
9. D. Nguyen, and A. S. Dogruoz, “Word Level Language Identification in Online Multilingual Communication”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 857–862, Seattle, Washington, USA, Association for Computational Linguistics.18-21 October, 2013.
10. D. Nguyen, and L. Cornips, “Automatic Detection of Intra-Word Code Switching”, In Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 82–86, Berlin, Germany, August 11. 2016.
11. D. Vilares, M. A. Alonso, and C. Gomez-Rodriguez, “Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora”, In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), pages 2–8, Lisboa, Portugal, 17 September, 2015.

12. D. Yarowsky, G. Ngai, and R. Wicentowski, “Inducing multilingual text analysis tools via robust projection across aligned corpora”, In Proceedings of the first international conference on Human language technology research, pages 1–8. Association for Computational Linguistics. 2001.
13. G. Grigonyte, and T. Baldwin, “Automatic Detection of Multilingual Dictionaries on the Web”, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 93–98, Baltimore, Maryland, USA, June, 2014.
14. H. Mukherjee, S. Ghosh, S. Sen, O. S. Md, K. C. Santosh, S. Phadikar, and K. Roy, “Deep learning for spoken language identification: Can we visualize speech signal patterns?”, In Neural Computing and Applications Springer-Verlag London, 5, September 2019.
15. L.-C. Yu, J., Wang, K. R. Lai, and X. Zhang, “Refining Word Embeddings for Sentiment Analysis”,In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pages 534–539 Copenhagen, Denmark, September 7–11, 2017.
16. L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, “Placing search in context: the concept revisited”, In Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, pages 406–414. 2001.
17. L.Lingjun, H. Liusheng, Y. Wei, Z. Xinxin, Y. Zhenshan, andC. Zhili, “Detection of Word Shift Steganography in PDF Document”, In Secure Communication,22-25, 2008, Istanbul, Turkey, 2008 ACM. September, 2008.
18. M. Abdalla, and G. Hirst, “Cross-Lingual Sentiment Analysis Without (Good) Translation”, In Proceedings of the 8th International Joint Conference on Natural Language Processing, pages 506–515, Taipei, Taiwan, November 27 – December 1.
19. M. Ahmed, P. C. Shill, K. Islam, M. A. S. Mollah, and M.A.H.Akhand, “Acoustic Modelling using Deep Belief Network for Bangla Speech Recognition”, In Proceedings of the 2004 ACM Symposium on Applied Computing (SAC 2004), pages 1128–1133, Nicosia, Cyprus.
20. M. Alam, and S. U. Hussain, “Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration”, In20th International Multitopic Conference (INMIC’ 17).
21. M. Ali, S. Khalid, and M. H. Aslam, “Pattern Based Comprehensive Urdu Stemmer and Short Text Classification”, In IEEEAccess, VOLUME 6,2018.
22. M. Faruqui, P. Majumder, and S. Pado, “Soundex-based Translation Correction in Urdu–English Cross-Language Information Retrieval”, In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 25–29, Chiang Mai, Thailand, November 8-12.2011.
23. M. G. A. Malik, C. Boitet, and P. Bhattacharyya, “Hindi Urdu machine transliteration using finite-state transducers”, In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 537–544, Manchester, UK, August. 2008.
24. M. Lui, J. H. Lau, and T. Baldwin, “Automatic Detection and Language Identification of Multilingual Documents”, In Transactions of the Association for Computational Linguistics, 27–40, February, 2014.
25. M. Potthast, B. Stein, and M. Anderka, “A Wikipedia-Based Multilingual Retrieval Model”, In Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, ECIR'08, pages 522-530, Berlin, Heidelberg. Springer-Verlag. 2008.
26. M. Zampieri, “Automatic Language Identification. In Working with Text: Tools, Techniques and Approaches for Text Mining”, chapter 8, pages 189–205. Elsevier, 2016.
27. N. Durrani, and S. Hussain, “Human Language Technologies”, In The 2010 Annual Conference of the North American Chapter of the ACL, pages 528–536, Los Angeles, California, Association for Computational Linguistics. June 2010.
28. P. D. Turney, “Distributional Semantics Beyond Words: Supervised Learning of Analogy and Paraphrase”, In Transactions of the Association for Computational Linguistics, 1 (2013), Page 353–366. 2013.
29. P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, andP. Rosso, “Query Expansion for Mixed-Script Information Retrieval”, In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 677–686. ACM, 2014.
30. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Open source toolkit for statistical machine translation”, In Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session, pages 177–180. Columbus, Oh, USA. 2007.
31. P. Resnik, and N. A. Smith, “The web as a parallel corpus”, In Computational Linguistics, 29(3):349–380. 2003.
32. R. D. Lins, and Jr. P. Gonçalves, “Automatic language identification of written texts”, In Proceedings of the 2004 ACM Symposium on Applied Computing, (SAC 2004), pages 1128–1133, Nicosia, Cyprus. 2004
33. S. A. B. Andrabi, and A.Wahid, “Sentence Alignment for English Urdu Language Pair”, In International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-1, May 2019.
34. S. Hussain, and M. Afzal, “Urdu computing standards: Urdu zabta takhti (uzt) 1.01”, In Multi Topic Conference, IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, 2001, pp. 223-228. 2001

35. S. Kanwal, K. Malik, K. Shahzad, F Aslam, and Z. Nawaz, “Urdu Named Entity Recognition: Corpus Generation and Deep Learning Applications”, In ACM Transactions on the Web, Vol. 9, No. 4, Article 39. March 2010.
36. S. Malik, and S. A. Khan, “Urdu Online Handwriting Recognition”, In International Conference on Emerging Technologies IEEE, Page 17-18, Islamabad. September, 2005.
37. S. M. Hassan, F. Ali, S. Wasi, S. Javeed, I. Hussain, and S. N. Ashraf, “Roman-Urdu News Headline Classification with IR Models using Machine Learning Algorithms”, In Indian Journal of Science and Technology, Vol 12(35), September, 2019.
38. S.-M. Kim, and E. Hovy, “Automatic identification of pro and con reasons in online reviews”, In Proceedings of the COLING/ACL Main Conference Poster Sessions, pages 483-490. 2006.
39. S. Mukund, and R. Srihari, “An Information-Extraction System for Urdu—A Resource-Poor Language”, InACM Transactions on Asian Language Information Processing, Vol. 9, No. 4, Article 15, Pub. date: December 2010.
40. T. Korenius, J. Laurikkala, K. Järvelin, M. Juhola, “Stemming and lemmatization in the clustering of Finnish text documents”, In Proceedings of the thirteenth ACM international conference on Information and knowledge management, page625-633. ACM. 2004.
41. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, In Advances in neural information processing systems, pp. 3111–3119. 2013.
42. T. Pham, and D. Tran, “VQ-based Written Language Identification”, In Proceedings of the Seventh International Symposium on Signal Processing and Its Applications (ISSPA 2003), volume 1, pages 513–516, Paris, France. 2003.
43. V. Paola, and K. Sanjeev, “Transliteration of Proper Names in Cross-Language Applications”, In proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR’03, Toronto, Canada. July 28–August 1, 2003.
44. W. J. Teahan, “Text Classification and Segmentation Using Minimum Cross-Entropy”, In Proceedings the 6th International Conference “Recherched ’Information Assistee par Ordinateur” (RIAO’00), pages 943–961, Paris, France. 2000.
45. Y. Al-Ohali, M. Cheriet, and C. Suen, “Databases for recognition of handwritten Arabic cheques”, In Pattern Recognition 36, pages 111–121, January, 2003.
46. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives”, Pattern Analysis and Machine Intelligence, IEEE Transactions, 35(8):1798-1828, 2013.
47. Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model”, In Journal of Machine Learning Research, 3:1137–1155. 2003.
48. Y. Haribhakta, A. Malgaonkar, and D. P. Kulkarni, “Unsupervised Topic Detection Model and Its Application in Text Categorization”, In Proceedings of the CUBE International Information Technology Conference ACM, September 2012.
49. Y. Qu, G. Grefenstette, and D. A. Evans, “Automatic Transliteration for Japanese-to-English Text Retrieval”, In proceedings of the 26 annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp: 353–360. 2003.
50. Z. Ceska, M. Toman, and K. Jezek, “Multilingual Plagiarism Detection”, In Proceedings of the 13th International Conference on Artificial Intelligence (ICAI 2008), pages 83-92, Varna, Bulgaria. Springer-Verlag. 2008.
51. Z. Sharf, and D. S. U. Rahman, “Performing Natural Language Processing on Roman Urdu Datasets”, In IJCSNS International Journal of Computer Science and Network Security VOL.18 No.1, January, 2018.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING