For surveys, a comprehensive review of traditional, machine learning, and deep learning-based methods in Bangla natural language processing.

Ayesha Akter
ID-21166014. ayesha.akter2@g.bracu.ac.bd
Brac University,CSE Dept.

ABSTRACT
The most interactive technology between a human and a machine is speech recognition. In the last 70 years, a lot of progress has been made in this crucial field of speech communication. For internet resources, technical knowledge, journals, and documentation, English is the most common language. As a result, many Bangla-speaking people with low English skills find it difficult to access English materials. Many attempts are also being made to make the Bangla language more accessible in the web and technical spheres. Some review papers are available to help you understand historical, current, and future Bangla Natural Language Processing (BNLP) trends. Information Extraction, Machine Translation, Named Entity Recognition, Parsing, Parts of Speech Tagging, Question Answering System, Sentiment Analysis, Spam and Fake Detection, Text Summarization, Word Sense Disambiguation, and Speech Processing and Recognition are among the 11 categories covered in this paper. I looked at articles from 1999 to 2021, and I found that half of the papers were published after 2015. We explore traditional, machine learning, and deep learning methodologies using various datasets, as well as the BNLP's limits, current, and future developments.

KEYWORDS-Support vector machine, artificial neural network, long short-term memory, gated recurrent unit, convolutional neural network, bangla natural language processing, sentiment analysis, speech recognition, support vector machine, artificial neural network, long short-term memory, gated recurrent unit.

I. INTRODUCTION

The computer's invention has a far-reaching impact in our current day, as it is helping to make everything easier for us. Everything is sent and understood by the computer using machine language, which is made up of 0 and 1. NLP is a branch of computer science that deals with computers understanding and interpreting human language (Natural Language Processing). NLP has become a popular topic in computer science in recent years as it solves a variety of real-world problems such as autocorrection, text to speech processing, machine translation, sentiment analysis, and so on. The applications of NLP are growing as computers become more capable of computing power. Because different languages are spoken in different parts of the world, it's become necessary to expand the NLP domain so that other languages can be employed in NLP applications. In the Indian subcontinent, several languages are spoken, and Bangla is one of them. It is frequently used in the Bangladeshi and Indian states of West Bengal.
Bangla Natural Language Processing is the branch of NLP that works with comprehending and processing activities connected to the Bangla language (BNLP). The components of Natural Language Processing are depicted in Figure 1.

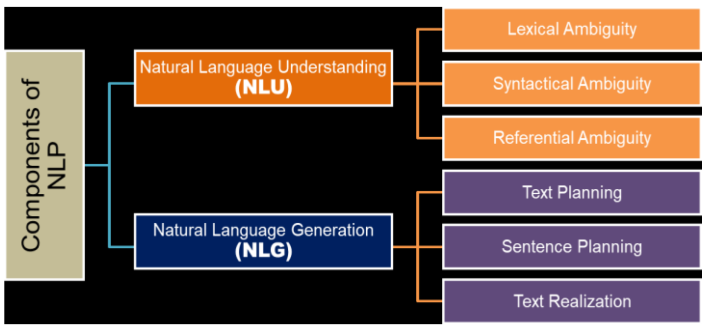

FIGURE 1. Natural Language Processing Components [1].

A. PREPROCESSING TECHNIQUES

The fundamental goal of data preparation is to convert raw data into a usable and analyzable format for the target task, and enable the computer to interpret the collected data in the appropriate format. Data preparation of the obtained raw data is one of the most important elements of any natural language processing task. The preprocessed data speeds up and improves the accuracy of NLP applications for the task at hand[1].

B. CLASSICAL METHODS

The most widely used traditional speech processing and recognition methods are Dynamic Time Warping (DTW) [2], Hidden Markov Model (HMM) [3], Linear Predictive Coding (LPC) [4], Gaussian Mixture Model (GMM) [2], Template Matching [2], Autocorrelation [5], Cepstrum [5], Speech Application Program Interface (SAPI) [6], Factorial Hidden Markov Model [7], Minimum Classification Error [8], Knowledge Based Approaches [9], Template Based Approaches [9], and Perceptual Linear Prediction [10].

The classical methods used in question answering system are Cosine similarity Jaccard similarity Anaphora- Cataphora Resolution [12], Vector Space Model [13], Semantic Web Technologies [14], Inductive Rule Learning and Reasoning, Logic Prover [16] and many more. Spam and fake detection systems mainly use the following classical methods: Traditional Linguistic Features [17], Text Mining and Probabilistic Language Model [18], Review Processing Method Time Series and Active Learning.

The majority of sentiment analysis systems used machine learning (ML) methodologies. There are only few traditional ways to sentiment analysis. The rule-based method is one of the traditional ways. The classical approaches on machine translation are: Verb based approach, Rule-based approach using parts of speech tagging, Fuzzy rules, Rule based method and many more methods. Parts of speech tagging use the following classical approaches: Brill's Tagger, Rule-based approach, and Morphological Analysis.

The Heuristic approach is the most commonly used method for text summarization, and the Simple Suffix Stripping Algorithm Score based Clustering algorithm and Feature Unification based Morphological Parsing are the most commonly used methods for parsing. Many statistical approaches have been employed to develop various systems in the context of information extraction for the Bangla languages. Additionally, the usage of traditional approaches to add more feature sets to their datasets is noted in order to create a robust system. To add more features to a dataset for information extraction, techniques such as Hough Transform-based Fuzzy Feature, Gradient Feature, and Haar Wavelet Configuration are utilized.

The majority of research papers in NER (Named Entity Recognition) employ Dictionary-based, Rule-based, and Statistical-based techniques. For NER, many statistical models are used, such as Conditional Random Fields (CRF)[12]. The Ruled-based technique is used in the majority of articles in parsing. For parsing, many Lexical Analysis and Semantic Analysis methods have been applied. The researchers also used several context free grammars to create new parsing rules.
The researchers also built a morphological parser using open-source technologies such as PC-KIMMO[13].
Analyzing various semantic information can be used to develop a parser. Various information retrieval approaches for summarization

with relational graphs are employed in the text summarization task. The Topic-sentiment model and Theme Clustering are used to recognize and aggregate topic sentiment [14]. For text summary, frequency, position value, cue words, and the skeleton of the documents are also utilized. Rule-based procedures are the traditional methodologies employed in word sense disambiguation [15]. The rules for identifying semiotic classes and regular phrases are commonly employed. In addition, several Context-free Grammars are used in word sense disambiguation.

C. MACHINE LEARNING AND DEEP LEARNING METHODS

In speech processing and recognition, the most often used machine learning and deep learning approaches include are Convolutional Neural Network (CNN) Backpropagation Neural Network (BPNN) Transfer Learning Gated Recurrent Unit (GRU)[16] Recurrent Neural Network (RNN) Long Short-Term Memory (LSTM) Artificial Neural Network (ANN) Deep Generative Model and Deep Neural Network (DNN).

In a question answering system, machine learning and deep learning approaches are applied in Naive Bayes algorithm, Support Vector Machine (SVM)[17], Stochastic Gradient Descent (SGD) Decision Tree N-grams Formation and Convolutional Neural Network.

D. CHALLENGES AND FUTURE RESEARCH IN BNLP

➢ Researchers facing problem challenges when they were working with Bangla text and speech data also processing.

➢ Sometimes face lacking of rigorous and comprehensive public corpus availability.

➢ One of the most difficult limitations for academics working with BNLP is the intricacy of Bangla grammar and structure.

➢ When working on the development of a Bangla speech processing and recognition system that can perform accurately in natural, freestyle, noisy, and all feasible contexts, the researchers encountered a common challenge.

2) **Future Research**

Despite being the world's seventh most spoken language, only a modest amount of meaningful research has been done on the Bangla language. Because most individuals converse in romanized Bangla, quality study on the language can be conducted. As improved deep learning models have demonstrated superior performance when working with various languages, more contemporary and appropriate deep learning-based models can be built in the research fields of BNLP. As a result, a large amount of relevant scientific work on the Bangla language is possible.

E. CONTRIBUTIONS

The following are the highlights of this paper:

- This page comprehensively discusses recent efforts on BNLP based on classical, machine learning, and deep learning.
- We divided the papers into 11 categories (textual and visual representation), which are Information Extraction, Machine Translation, Named Entity Recognition, Parsing, Parts of Speech Tagging, Question Answering System, Sentiment Analysis, Spam and Fake Detection, Text Summarization, Word Sense Disambiguation, and Speech Processing and Recognition.
- We highlighted the limitations of the studies, as well as enhancement ideas, diverse datasets, and current and future BNLP research directions.

F. PAPER SUMMARY

Various deep learning approaches were discussed, as well as the ongoing challenges and advances in Bangla natural language processing. It has demonstrated various information extraction, machine translation, named entity recognition, parsing, parts of speech tagging, question answering system, sentiment analysis, spam and fake detection, text summarization, word sense disambiguation, and speech processing and recognition methods and techniques. Datasets are one of the most important variables in the growth of BNLP research, so we addressed numerous text datasets and speech datasets in the final part. Finally, this research demonstrated numerous developments, problems, and strategies applied in BNLP.

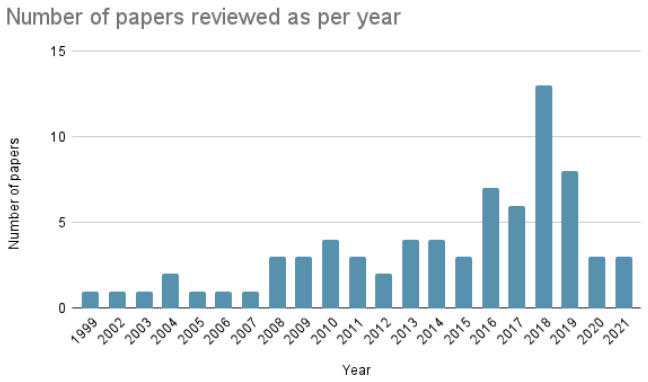

FIGURE 2. Number of Papers Reviewed as per Year.

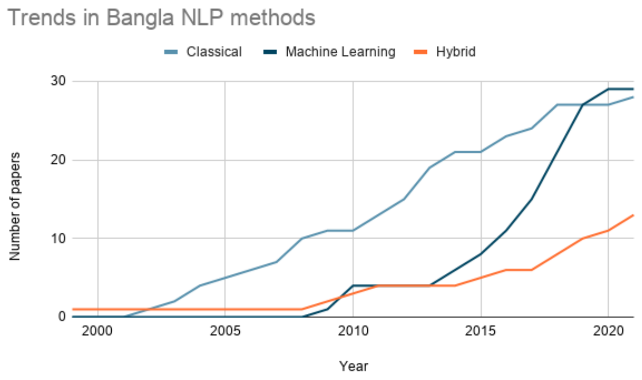

FIGURE 3. Trends in BNLP Methods.

The remainder of this paper is organized as follows.

❖ Bangla Language: The features and complexity of the Bangla language are briefly presented.

❖ Bangla Natural Language Processing: The necessity and needs, as well as the various categories and methods employed in BNLP, are briefly presented. This section discusses articles published in both text and speech formats on various areas of BNLP up to the present.

❖ BNLP's Difficulties: This section describes the obstacles encountered in BNLP study fields on both text and speech forms.

❖ BNLP Datasets: This section describes the datasets used in BNLP research articles published to date on both text and speech formats.

❖ Conclusion: A summation of our work as a whole.

II. BANGLA LANGUAGE

A. CHARACTERISTICS

Bangla's phonology is comparable to that of other Indo-Aryan languages. The basic Bangla alphabet has 11 vowels and 39 consonants. There are 10 numerical and compound characters as well. These compound characters are made up of a consonant and a vowel or a consonant and a vowel [18][19][20].

B. COMPLEXITY

There is also less collaborative activity among the scholars in this sector. Furthermore, the corpus of the research is frequently not professionally labeled, which has been a major worry in Bangla NLP activities. The lack of an uniform corpus and labeled data makes it difficult for scholars to do research in this area.

III. BANGLA NATURAL LANGUAGE PROCESSING

It is a method of comprehending the world around us. Every living thing has a way of communicating. Humans, on the other hand, developed languages in order to communicate and send intelligible facts to one another [21]. Throughout the years, numerous languages have emerged as a result of regional culture and environment. Some languages evolved, while others became extinct. However, the basic purpose of the language remained the same: talking with one another.

A. NECESSITY AND NEEDS

In BNLP, information extraction enables machines to decode and extract the fundamental knowledge of words from Bangla documents. Machine translation is utilized in BNLP to translate Bangla text data to another language. The goal of named entity recognition in BNLP is to detect and classify every word in a Bangla document into specified named categories.

B. PREPROCESSING

Noise Removal-emphasis, Hamming Window, Segmentation, Sampling, Phoneme Mapping, Speech Coding, stemming without Noise Removal, Voice Activity Detection, Framing, and many other techniques are commonly used in Natural Language Processing on speech data [22][23].

C. INFORMATION EXTRACTION

1) Classical Approaches

For recognizing English words in Benglish and Hinglish languages, Chandra et al. [34] proposed statistical methods paired with a rule-based approach. Some fundamental patterns can be noticed in the code-mixed scenario with English, such as English words written within roman letters. After the English words, certain suffixes may be added.

2) Machine Learning Approaches

The photos were normalized by converting them to grayscale and resizing them to 320x240 pixels. The images were then subjected to a 3x3 median filter and edge-based detection based on the Sobel edge detection filter, which was used for text detection and extraction. The writers transformed the edited photos into binary images after detecting the text in the image. Text segmentation was required in order to recognize the characters in the binary images. To determine individual characters and transform them into a feature vector, line segmentation and character segmentation algorithms are used. The characters were then recognized using a multilayer perceptron model.

3) Combination of Classical and ML Approaches

Image augmentation for ISI datasets includes 19,392 training photos and 3,986 testing images However, by applying augmentation to the photos in the ISI training dataset, the authors were able to build a dataset of 58,176 to train the model on. They utilized this dataset to train the model, and CMATERDB3.1.1 had 6,000 photos for testing the model. The suggested model's main characteristic is that it achieved good accuracy while utilizing minimal computation power. In only 55 epochs, their model achieved a 99.02 percent accuracy. Another paper by Sazal et al. [121] employed DBN to recognize Bangla handwritten characters.

E. NAMED ENTITY RECOGNITION

The rule for named entity might be positive or negative, and the authors proposed 12 possible rules for named entity based on this fact. If the term follows specific rules, a Weight is assigned to it. A specified threshold value is set to determine whether a word is a name entity, and the weight of the word must be greater than the threshold weight to be accepted as a name entity. If a word is not a name entity, a different threshold weight is applied. If the weight of the test word is less than this threshold weight, it must also be less than the threshold value, indicating that it is not a name entity.

F. PARSING

To ensure a single parse tree, the ambiguity of two parse trees was overcome by adding two additional features and applying feature unification. Five types of nominal and pronominal inflections were identified. The features were utilized to make changes to the lexicon. Following that, compound words with inflectional suffixes were divided into four types. PC-Kimmo format was used to generate the final grammar. The morphological analyzer was implemented in PC-KIMMO version 2. The implementation yielded a 100 percent right result on complex words present in Bangla grammar texts.

G. PARTS OF SPEECH TAGGING

The bespoke stemming and rules were sufficient to detect the tags as precisely as feasible. Concentrating on all the subcategories of each base tag with punctuation will allow you to develop the tagger even more. Accuracy can be improved by using a limited dataset and combining a probabilistic technique.

H. QUESTION ANSWERING SYSTEM

Figure shows the basic system architecture of question answering System.

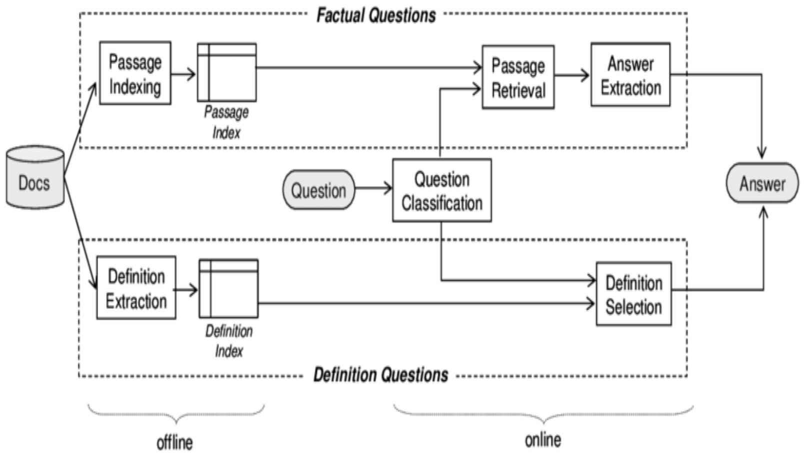

FIGURE 9. Basic System Architecture of Question Answering System

I. SENTIMENT ANALYSIS

The dataset was too small and too noisy. The author intended to utilize unlabeled data during the training process of DBN and CNN-based models. A recursive neural network proposal was also presented. A larger dataset can increase the model's overall performance. Adding more layers to the model is another option.

J. SPAM AND FAKE DETECTION

When linguistic characteristics were combined with SVM, the system achieved an F1-score of 91 percent. For news embedding, RF produced a 55% F1-score, while SVM and LR produced 46% and 53% F1-scores, respectively. In CNN, average pooling and the global max approach yielded 59 percent and 54 percent F1-scores, respectively. F1-Score was 68 percent in the BERT model. Only about 8.5K news items were manually annotated by the writers in this case. More data can be annotated to improve this article.

K. TEXT SUMMARIZATION

The extractive approach employs a text summary methodology based on the K-means clustering algorithm. In this paper, sentence scoring was applied in the extractive technique to improve summarization. The proposed method is applicable to both single and multiple Bangla documents.

L. WORD SENSE DISAMBIGUATION

The authors used the Indian Language Corpora Initiative's Bangla POS tagged corpus and the Indian Statistical Institute's Bangla WordNet. The authors made the text more readable by deleting superfluous spaces, punctuation, and delimiters. They turned the entire text to Unicode. The non-functional words were then eliminated from the sentences.

M. SPEECH PROCESSING AND RECOGNITION

The authors constructed this virtual assistant by combining a computer with home appliance peripheral devices. Because most virtual assistants are in English, this virtual assistant will benefit visually challenged people. The suggested virtual assistant was a user-independent system that could conduct numerous computer operations in the home setting using Bangla human voice commands.

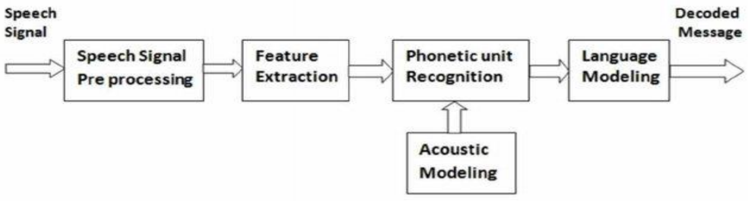

FIGURE: Basic system architecture of Speech Processing and Recognition System

This paper can be enhanced by increasing the size of the datasets used for training and testing, as well as generating a suitable and upgraded version of deep learning models for the recognition job.

N. METHODS USED IN BNLP

1) Classical provides a brief explanation of the classical approaches employed in BNLP up to this point.
2) M.L. and D.L. provides a brief explanation of the machine learning and deep learning methods that have been applied in Bangla natural language processing up to this point.

IV. CURRENT CHALLENGES AND FUTURE TRENDS IN BNLP

The most common challenges in preprocessing Bangla speech data are: preprocessing of incorrectly interpreted Bangla speeches, preprocessing of very noisy or loud Bangla speeches, preprocessing of reverberations and overlapping Bangla speeches, preprocessing of a Bangla speech signal with an unnecessary delay at the end of speech frames, preprocessing of sparsely spoken speech, and preprocessing of ambiguous speech.

V. BNLP DATASETS

The authors dealt with 1,000 high-frequency terms in this study. They chose 52 sentences at random from three issues of daily newspapers, including these high-frequency words, for recording. Twenty-six thousand audio files from 50 male and 50 female people were kept as training corpus, and 15,600 audio files from another 25 male and 25 female people were kept as testing corpus. The speakers were between the ages of 18 and 25. Fifty-two sentences were chosen from 1,000 high-frequency terms for a total of 62 hours of recording. Noisy files were cleaned with various filters, and the amplitudes of short speech level data were enlarged to a reasonable level. The unedited original audio files were altered to complete a sentence. We present a summary of the datasets used in the Bangla voice processing and recognition system.

VI. CONCLUSION

Natural language processing is a developing area in the realms of modern machine learning and deep learning. Natural language processing is becoming more popular by the day. The significance of BNLP is clear, as Bangla is one of the world's most widely spoken languages. The importance and demand for BNLP in the present world were also emphasized. In the corresponding sections, we provide several methods and approaches from many categories, providing a quick overview of implemented methodologies. We also discussed the difficulties encountered during the creation of BNLP systems. In addition, we have summarized a number of regularly used datasets. Finally, we offered a thorough analysis of BNLP approaches, including datasets and outcomes, while addressing limits, providing enhancement ideas, and evaluating current and future trends.

ACKNOWLEDGMENT

Acknowledgement: The authors would like to thank Taif University Researchers Supporting Project number (TURSP-2020/239), Taif University, Taif, Saudi Arabia, for their assistance.

REFERENCES

[1] “Natural language processing (nlp) simplified : A step-by-step guide.” Accessed: 2021-03-21.

[2] M. A. Ali, M. Hossain, M. N. Bhuiyan, et al., “Automatic speech recognition technique for bangla words,” International Journal of Advanced Science and Technology, vol. 50, 2013.

[3] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech communication, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] A. K. Paul, D. Das, and M. M. Kamal, “Bangla speech recognition system using lpc and ann,” in 2009 Seventh International Conference on Advances in Pattern Recognition, pp. 171–174, IEEE, 2009.

[5] M. N. A. Aadit, S. G. Kirtania, and M. T. Mahin, “Pitch and formant estimation of bangla speech signal using autocorrelation, cepstrum and lpc algorithm,” in 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 371–376, IEEE, 2016.

[6] S. Sultana, M. Akhand, P. K. Das, and M. H. Rahman, “Bangla speechto-text conversion using sapi,” in 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 385–390, IEEE, 2012.

[7] T. Virtanen, “Speech recognition using factorial hidden markov models for separation in the feature space,” in Ninth International Conference on Spoken Language Processing, 2006.

[8] B.-H. Juang, W. Hou, and C.-H. Lee, “Minimum classification error rate methods for speech recognition,” IEEE Transactions on Speech and Audio processing, vol. 5, no. 3, pp. 257–265, 1997.

[9] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A review on speech recognition technique,” International Journal of Computer Applications, vol. 10, no. 3, pp. 16–24, 2010.

[10] N. Dave, “Feature extraction methods lpc, plp and mfcc in speech recognition,” International journal for advance research in engineering and technology, vol. 1, no. 6, pp. 1–4, 2013.

[11] M. Kowsher, M. M. Rahman, S. S. Ahmed, and N. J. Prottasha, “Bangla intelligence question answering system based on mathematics and statistics,” in 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1–6, IEEE, 2019.

[12] S. Khan, K. T. Kubra, and M. M. H. Nahid, “Improving answer extraction for bangali q/a system using anaphora-cataphora resolution,” in 2018 International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–6, IEEE, 2018.

[13] S. Sarker, S. T. A. Monisha, and M. M. H. Nahid, “Bengali question answering system for factoid questions: A statistical approach,” in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5, IEEE, 2019.

[14] A. B. Abacha and P. Zweigenbaum, “Means: A medical questionanswering system combining nlp techniques and semantic web technologies,” Information processing & management, vol. 51, pp. 570–594, 2015.

[15] A. Mitra and C. Baral, “Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[16] D. Moldovan, C. Clark, et al., “Natural language question answering system and method utilizing a logic prover,” Nov. 17 2005. US Patent App. 10/843,178.

[17] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, “Banfakenews: A dataset for detecting fake news in bangla,” arXiv preprint arXiv:2004.08789, 2020.

[18] R. Y. Lau, S. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, “Text mining and probabilistic language modeling for online review spam detection,” ACM Transactions on Management Information Systems (TMIS), vol. 2, no. 4, pp. 1–30, 2012.

[19] R. Ghai, S. Kumar, and A. C. Pandey, “Spam detection using rating and review processing method,” in Smart Innovations in Communication and Computational Sciences, pp. 189–198, Springer, 2019.

[20] A. Heydari, M. Tavakoli, and N. Salim, “Detection of fake opinions using time series,” Expert Systems with Applications, vol. 58, pp. 83–92, 2016.

[21] M. I. Ahsan, T. Nahian, A. A. Kafi, M. I. Hossain, and F. M. Shah, “Review spam detection using active learning,” in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–7, IEEE, 2016.

[22] S. Mandal, S. K. Mahata, and D. Das, “Preparing bengali-english codemixed corpus for sentiment analysis of indian languages,” arXiv preprint arXiv:1803.04000, 2018.

[23] M. Rabbani, K. M. R. Alam, and M. Islam, “A new verb based approach for english to bangla machine translation,” in 2014 International Conference