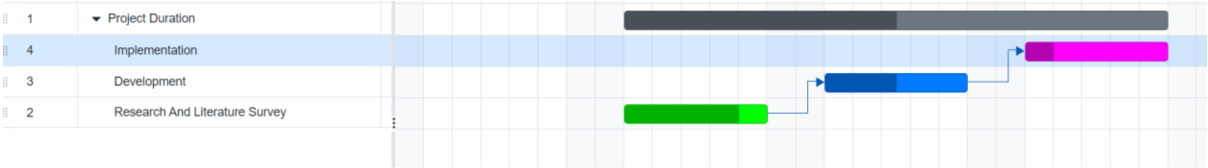
DA2

VAIBHAV RAI

20BCE2651

1.Project Plan



1	Project Duration
4	Implementation
3	Development
2	Research And Literature Survey

2.Introduction

Human language is beautifully complex. Natural language processing attempts to convert human language into machine understandable language. Attempting to understand natural language has become very important in today's world as this forms the basis of many applications like translators, object identification, billboard sign recognition etc.

Text detection is a very important aspect of NLP. It is used in image based translation programs and others. One observation the authors make regarding existing textbox algorithms is that they treat prediction and regression as 2 separate methods and don't take advantage by combining them. Combining both gives the advantage of using regression for learning correctly predicting text boundaries and using prediction to actually draw the text boxes

Natural language finds application in many ways in our lives. We regularly use translation apps which translate words of other languages into the language of our choice. We have the option to use either voice or pictures and videos of the text for translating. Google translate is a good example of a translator. These days, natural language processing is also used to artificially produce voices of people, especially actors and singers.

A lot of research has already been done into improving natural language methods and also making applications that implement these techniques in a new or improved way, including fundamental tasks like text detection where text is detected using boxes. Also, various processing methods have already been proposed not only to process many languages like Bengali and other European languages. Some popular methods include using LSTM, autoencoder among others. The authors of paper used a combination of LSTM and BERT to optimise the processing. In their method, LSTM was used as it could remember complex relationships, which occur in long sentences, and BERT was used to understand the positional awareness of words in the sentence. Autoencoder methods were used to build recommender systems which would give out the next word given the sentence.

However, very less research has been done into the complexity of Indian languages. The unique thing about Indian languages, especially Hindi, is that they are derived from the Devanagari script, whereas the languages on which the above models were trained had Roman script, which would be familiar and easy to manipulate. Hindi needs a separate method of manipulation. As mentioned above, most languages follow Roman script whereas Indian languages follow a totally different script.

The work presented in this paper is aimed at digitising Indian languages. We present a new method of tokenizing words of the Hindi language using LSTM model and studies of paper . We use certain methods already implemented by others, but change them accordingly so that they can recognize Indian languages. We also compare our method with existing methods to get our accuracy.

3.Detailed Literature Survey

Literature Review

Definition: A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. Its purpose is to create familiarity with current thinking and research on a particular topic, and may justify future research into a previously overlooked or understudied area.

1. Introduction:

Text detection is a very important aspect of NLP. It is used in image based translation programs and others. A solution to detect text by accurately drawing detection boxes is given in [1]. This paper emphasizes the importance of scene text detection. The authors consider the problem of irregular text, where text is not in a straight line. The authors propose a method called textfield, a method where intensity of the pixel is learnt to understand the character. Here CNN is used to regress the text field. This paper explores the problem of detecting text by using segmentation methods. The authors argue that existing methods will have difficulty producing boxes of larger aspect ratios.This paper presents what is called Region Proposed networks, which combine both convolutional layers and object detection networks. The MSRA Text Detection 500 Database (MSRA-TD500) contains 500 natural photos acquired with a pocket camera from indoor (office and mall) and outdoor (street) scenarios. The images primarily consist of signs, door plates, and caution plates, guide boards and billboards with elaborate backgrounds of varying resolutions. This is considered as a benchmark to test object detection algorithms.

Recent breakthroughs in neural text-to-speech (TTS) have enabled real-time synthesis of human-like speech. Parallel models may produce mel-spectrograms orders of magnitude quicker than autoregressive models, either by depending on external alignments or by aligning themselves.The authors suggest an unique feed-forward network based on Transformer to create mel-spectrograms in parallel for TTS in paper [2]. The FastPitch model was compared to Tacotron 2 by the authors. The samples were rated using the Crowdsourced Audio Quality Evaluation Toolkit on Amazon Turk. They synthesised speech from the first 30 samples in their LJSpeech-1.1 development subset. At least 250 scores were collected for each model, with a total of 60 distinct speakers taking part in the study. Dataset used in this study was LJSpeech which is a public domain voice collection made up of 13,100 brief audio clips of a single speaker reading portions from seven non-fiction books. Each clip comes with a transcription.

The work in paper [3] offers a deep learning-based approach for multi-criteria recommender systems that uses deep autoencoders to exploit the non-trivial, nonlinear, and hidden relationships between users' multi-criteria preferences and provide more accurate suggestions. Based on the people and things, collaborative filtering creates suggestions. It calculated user-user or item-item similarities and generated customised suggestions based on them. Regardless of their efficiency, CF approaches suffer from well-known issues such as cold start, forecast accuracy, and the inability to capture complicated user-item interactions when the rating matrix is sparse. Deep feedforward neural networks are used in the proposed deep autoencoder-based multi-criteria recommendation algorithm (AEMC). The raw MC user item rating matrix is obtained using the AEMC algorithm. The dataset is created in the first phase in order to train and test the deep autoencoder model. The deep autoencoder creation with all necessary hyper-parameters is the second phase. The deep autoencoder is trained and evaluated in the third step utilising the prepared dataset from phase 1. Only the ratings columns are considered for the calculation of the loss for the reason that it would recommend a particular item to the user based on the predicted user rating to an item.

The paper [5] talks about a nest syllable prediction system using LSTM. The author did the research on the Dzongkha language. The model used in this paper by the author is LSTM. Its parameters are calibrated to obtain a higher accuracy. Long short term memory network was used because it remembers both short and long term textual dependencies. LSTM model is used to predict next 5 syllables in the Dzongkha language. Word embedding and tokenization concept are used in this model. Dzosyll dataset is used which consist of literatures in this language. The metrics to evaluate the model used by the author are the time required to train the model, checking the errors in the model on a dataset and the times the output was right. The proposed system generates sequences from sentences which were fed into 300-dimensional vector size word embedding in order to vectorize the texts. These were fed into the LSTM as the input. Relu activation function and softmax was used for the first dense layer and second one respectively. The syllable with the highest probability was displayed first followed by the remaining syllables. This model was deployed using Django 3.0.7, Bootstrap 4.5.0, and CSS. The best accuracy of 78.33% and lowest loss of 0.7110 was given by a one-layer LSTM model.

In word prediction, we use a language and capture the joint distribution of sequences of words of a natural language. A few characters or words are given to predict the following characters or words. In paper [6] several popular architectures were re-evaluated and found that LSTM when regularized properly, performs better than most of the newer models. In this paper Convolution kernel of CNN was introduced into the language model to fully mine the local feature information and increase the accuracy. A residual connection based MGU unit was conducted in order to solve the problem of vanishing gradient and network degradation. Penn treebank dataset was used by the author. Perplexity was used as evolution metric in the paper.The model was implemented using Tensorflow 1.14.

Next word prediction is a domain of NLP where a sequence of words and characters are treated as input and the output is the most probable. In paper [7] the evidence combination algorithm combines the evidence using Dempster's combination rule. Th REEDS algorithm predicts the next word from a given set of probable words. Relevance and enhanced entropy based Dempster Shafer Algorithm is used by the author in this paper. Brown corpus is used as the test dataset. It is a hybrid of different evidence with EBE and Dempster's combination rule. The proposed model is considered to be better than neural network approach due to less computational burden and lingering learning time of the neural network. In the REEDS algorithm uncertainty measue, credibility of body of evidence(BOE) is calculated , and the difference between average and individual credibility of each BOE is calculated. These are normalized to calculate weight of each BOE and next word is predicted

2. Main Body:

Paper title : Natural Language Processing Based Sentimental Analysis of Hindi (SAH) Script an Optimization Approach

Sentimental analysis is one of the most common applications of Natural Language Processing (NLP). Sentiment analysis, the term itself refers to identify the emotions and opinions of people through written text. This paper presents a sensible dynamic approach on to search out the polarity of any sentence and analyse the opinion of the actual sentence. The projected Sentimental Analysis of Hindi (SAH) script have adopted 2 totally different classifier Naïve Thomas Bayes Classifier and call Tree Classifier is employed for the text extraction. The positive, neutral and negative result validation shows a comparative results of sentimental analysis.Natural Language process (NLP) is employed to use machine learning algorithms to text and speech.NLP are often wont to produce systems like sentimental analysis, speech recognition, document summarization,machine translation, spam detection, named entity recognition, question answering, automotive vehicle complete, prophetic typing, data extraction, and then on. Nowadays, human language technology is used to power search engines, filter spam and to get analytics during a quick and ascendable manner. Sentimental analysis is that the initial NLP task that each information mortal has to perform to grasp the operating mechanism and necessity of information in NLP.One of the major challenge in sentiment analysis is to model the interaction between the specified target entity and its context. Some existing methods of affective sentiment analysis can be divided into three categories which includes knowledge-based techniques, statistical methods, and hybrid methods. Some common sources of sentiment words includes Affective Lexicon, linguistic annotation scheme, WordNet-Affect, SentiWordNet, SenticNet. In the era of digitalization, a large number of people share their views and opinions on the World Wide Web (WWW) such as online reviews. There are a number of other people with around 73–87% who use those online reviews for the choice of purchasing the product. Many researches on sentiment analysis are mainly focused on machine-learning based. However, it is also having a problem, as they need a huge amount of training data to perform well. It is well-known that totally different languages have their own distinctive ways that of expression. the fundamental distinction between English and Hindi language is that the language structure. For example, West Germanic has S-V–O (Subject-Verb- Object) structure, whereas Hindi language follows S–O-V (Subject-Object-Verb) structure. the fundamental structural difference between English and Hindi language has consequences to decide the polarity of text. a similar set of words with slight variations and changes within the order have an effect on the polarity of the words in the text. Therefore, a deeper linguislication analysis is needed while coping with Hindi language to perform Sentiment Analysis.This paper used a dataset of hindi stop words for sentimental analysis. The dataset is created with collective of positive and negative wordlists To test the data collected is tested and proceeded using the NLTK (Natural Language Toolkit) library which is readily available for Python programming language. By importing all required functions and features from NLTK library, the proposed SAH successful validated to get the required analysis which is illustrated in the result section. Moreover, the trained datasets is classified by using NaiveBayesClassifier and DecisionTreeClassifier which is verified for the analysis of positive and negative words to segregate the appropriate mapping.

Paper title : A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context

Natural language generation(NLG) has surfaced from NLP and is now generally employed in colorful operations, including drooling operations. The ideal of this paper is to propose a deep literacy-grounded language generation model that simplifies the process of writing medical recommendations for croakers in an Arabic environment, to ameliorate service satisfaction and case- croaker relations.The proposed model was trained using data attained from Altibbi databases related to medical recommendations, particularly gynecology, dermatology, psychiatric conditions, urology, and internist conditions. Variants of deep literacy models were enforced and optimized for coming word vaticination, grounded . on the unidirectional and bidirectional long short- term memory (LSTM and BiLSTM), the one-dimensional . convolutional neural network (CONV1D), and a combination of LSTM and CONV1D (LSTM-CONV1D). The algorithms were trained using two performances of the datasets and estimated in terms of their training delicacy and loss, confirmation delicacy and loss, and testing

delicacy per their corresponding scores. The proposed models ′performances were similar. CONV1D produced the most promising corresponding score. Assessing NLG tasks is an open exploration direction, and there is no bone standard way to measure the performance.Thus, this paper comparesthe models by consideringthe training and confirmation of delicacy and loss, as well astesting delicacy,
in terms of matching scores. Specifically, matching scores are used to estimate the models by assessi ngwhat chance of matching (1-gram lapping) betweenthe generated textbook (i.e., word) and the ground-verity textbook. Thematching score metric only shows how important the models canfind the exactly matched word with the ground- verity anyhow of whether the generated word is correct and applicable tothe environment. Meanwhile, the end is to induce five prognosticationsthat are all applicable to the separate environment.The employed Arabic-language datasets were obtained from Altibbi. The Arabic language is a Semitic language and the mother tongue of more than 200 million speakers worldwide. It is so common because it is the language of Islam and the means of communication in daily Arabian life. Several variants of deep learning models are implemented to process two different versions of the n-gram datasets of the five specialties.The eight versions of datasets were fed into LSTM, BiLSTM, CONV1D, and a combination of LSTM and CONV1D. They were then evaluated quantitatively and qualitatively. This section introduces the networks used and the details of their implementation.NLG has been adopted for various applications, but it is insufficiently studied in the Arabic context. An NLG-based model is proposed in this paper and applied as a one-time, one-word, predictive text model for Arabic medical recommendations. The objectives are to save doctors ′time, improve service satisfaction, and improve patientdoctor interactions. Variants of deep learning models were utilized to predict the next word of text-based medical recommendations across various specialties.

Paper title : Evaluation of Deep Learning Models for Hostility Detection in Hindi Text

In this paper, they assess distinctive profound learning approaches for the Imperative @AAAI 2021 Hindi Threatening vibe detection dataset. The assignment can be seen as a multi-label classification task mainly for different categories of threatening vibe sort. A combination of diverse profound learning models along side diverse varieties of word embeddings is compared in this paper. The distinctive models compared in this paper straightforward 1D CNN, multichannel CNN, bi-directional LSTM, CNN+LSTM, mBERT, and IndicBERT. The arbitrary and quick content embeddings discharged by IndicNLP and Facebook are the variations of word embeddings utilized. Through our tests, we appear that basic models perform competitively with BERT based models although the last mentioned is marginally way better. The IndicBERT and mBERT perform similarly well. The multi-CNN model when combined with IndicNLP FastText word inserting performs best among the fundamental models.In this work, the execution of diverse profound learning models based on CNN, LSTM, and BERT was assessed on the Hindi Threatening vibe location dataset. The input representation to fundamental models is conveyed word vectors. We compare random initialization of word vectors and pre-trained fast text embeddings prepared on Hindi corpus. The quick content word embeddings are based on subword units and well suited for noisy datasets. They fine-tune the pre-trained quick content word embeddings to adjust the show to the target corpus as the target space is distinctive from the space on which the word vectors were prepared. We utilize two variations of Hindi fast text embeddings discharged by Facebook and IndicNLP.We have utilized the antagonistic vibe location dataset within the Hindi dialect made accessible beneath Constraint@AAAI 2021 Shared Task . The shared assignment basically centered on the hostility detection on three major focuses, i.e. low-resource territorial dialects, discovery in crisis circumstances, and early detection. The dataset contains 8192 online posts collected from different social media stages like Twitter, Facebook, WhatsApp, etc., and are labeled as antagonistic or non-hostile posts. Besides, antagonistic posts are classified as fake, abhor, criticism, and offensive.The stages are broadly utilized by artists to share their craftsmanship and common individuals to share some aspects of their day to day life. It is utilized by individuals to voice their conclusion or pass on valuable data. In any case negative aspects of these social stages are being misused for personal picks up. Extraordinary substance and abhor posts are taking over fun and data. The threatening posts are intentioned created to target a individual, race, ethnicity, sex, or indeed a country. This is incompletely since the fear of uncovering individual personality is no longer an issue. Individuals make a fake or mysterious account to share hostile substance. Such substance and the trolling culture has come about in mental injury and savage clashes between divergent bunches. The discovery and evacuation of unfriendly content from web stages are of foremost significance.

Paper title: DHOT-Repository and Classification of Offensive Tweets in the Hindi Language

To the leading of our information, typically the primary attempt to classify offensive Hindi tweets employing a fastText classifier. Since Hindi is the third most popular language within the world within the social media, it was chosen. Our work recognizes and after that classifies hostile content from non-offensive employing a fastText-based model with palatable comes about. As a future work we are going investigate other calculations and models for making strides the results, in expansion to utilizing comparable strategies for other Indian Dialects.Swearing or reviling is the utilize of befoul words, and is common in day-to-day, casual conversations. The utilize of such dialect is additionally broad in social media discussions. The use of damaging dialect can range from being simply hostile, forceful or being inside and out derisive and rough. differentiate between despise discourse and injurious discourse in their work. The abusive terms collected within the previous step serve as watchwords that were relegated to a information procurement program. This program utilized these keywords as a seed to recover information utilizing Twitter API calls. This shaped the abusive words dataset to be utilized for annotation. To form a more reasonable dataset, tweets were too mined from well known Twitter hashtags of viral topics, well known public figures like lawmakers, sports-personalities, and motion picture performing artists. Both of the datasets were combined to make the essential corpus for the investigate work.The tweets collected had to satisfy certain essential necessities. Tweets with more than 3 Hindi words were collected utilizing Twitter API (Application Programming Interface) to form a physically explained DHOT dataset. The tweets were collected over the course of three months, from September to November, 2018. Tweets were collected with India geolocation limitations on the API nourish to guarantee the dataset had the appropriate setting. The collected dataset was at first composed of 24,596 tweets. These tweets were cleaned and sifted down to expel posts containing as it were URLs, as it were pictures and recordings, or those with less than 3 words, or non-Devanagari scripts.To handle this dataset, both fastText and word2vec modeling were utilized. FastText is an open-source, free, lightweight library that permits clients to memorize content representations and content classifiers. Word2vec is a neural arrange show which treats each word in a corpus like an nuclear substance and produces a vector for each word.A major impediment in this prepare is that physically labeling information with the fitting classes may be a time consuming prepare. There were cases of mockery location, in any case this was most frequently found through manually working with dataset and equivocal classes (like mockery, mind, hidden insuperable, etc), that some instances of damaging content may have been missed At times, this framework isn't sufficient to avoid language misuse as the sheer estimate of the substance online can be scaring, as well as the rate at which it is being posted online. It is exceptionally troublesome to prepare each single post (content, picture, sound, and video) physically, particularly for non-European dialects which need common word references and following conventions. In this work numerous cases of blended English words composed in Hindi-Devanagari script were found. Whereas the lion's share of Marathi and Nepali tweets were caught whereas recognizing Hindi content, there were still some wrong positives and occurences of Nepali and Marathi archives within the Hindi.

5. Next syllables prediction system in Dzongkha using long short-term memory

Dzongkha is the official language of Bhutan, and it's very cumbersome to type. It takes several keypresses to type even a single syllable. This paper proposes a Next Syllables Prediction System for this language in order to make typing easier. It consists of the following stages: data acquisition, text pre-processing, word embedding, model training, and deployment. The "DzoSyll" dataset was used which consists of short stories, news, and essays. Preprocessed text is fed into the word embedding stage in the form of vectors. These sequences of vectorized text are fed into the LSTM model for training. Hyperparameter tuning of word embedding and LSTM helps in getting higher accuracy. Once required accuracy is achieved, the model is saved for deployment. It is deployed with Django, Bootstrap and Javascript. Word embedding is a method of vectorizing text. Words with identical meanings have similar representations. It captures both the semantic and syntactic relationship between words. RNN loses information as the length of the input sequence increases due to vanishing gradients. LSTM and Bi-directional LSTM address this drawback. It has a cell state, hidden state, input gate, forget gate, and output gate. Hidden state is responsible for short term memory and cell state is responsible for long term memory.

The proposed system generates sequences from sentences which were fed into 300-dimensional vector size word embedding in order to vectorize the texts. These were fed into the LSTM as the input. 128 LSTM cells were used. RelU activation function was used for the first dense layer, and softmax for the second one in order to provide a probability distribution. The syllable with the highest probability was displayed first followed by the remaining syllables. This model was deployed using Django 3.0.7, Bootstrap 4.5.0, and CSS. The best accuracy of 78.33% and lowest loss of 0.7110 was given by a one-layer LSTM model with 128 cells.

The limitations of this study are that the model is based on limited text and a small dataset. If the system sees a symbol which is not in the training dataset, the system may not be able to accurately predict the next syllable. The accuracy will benefit from using more data curated from different genres for example, songs and poetry.

6. Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network

This paper proposes a MCNN-ReMGU model based on multi window convolution and residual connected minimal gated unit network to predict natural language words. Different sized convolutional kernels are used to extract local feature information of different graininess between word sequences. These features are fed into the residual-connected MGU network. The results of the prediction are outputted by a SoftMax layer. The residual-connection processing of the MGU network solves the problems of vanishing gradient and network degradation.

The MCNN-ReMGU model consists of the following parts: word embedding layer, CNN local perception layer, residual connection to the MGU network layer and softmax output layer.

Word embedding maps words from high dimensional sparse space to a low dimensional real vector space. The CNN local perception layer extracts feature information of different granularity by using different window sizes for convolution. Multiple (window-size)*(dimension of the embedded vector) convolutional kernels which conduct convolution on the sentence. Batch normalization is performed to prevent covariates transferring within the data. The size of the output of this layer is n*r after r convolutional operations.

As the depth of the network is increased, redundant network layers are generated. This leads to poor network performance due to network degradation. The activation function of the hidden state is the ReLU activation function. It has the ability to avoid vanishing gradients which is caused by the saturation function. This way, deeper networks can be trained.

Output of the ReMGU is taken into the output layer which uses the softmax function to normalize values for the final word prediction.

The model was implemented using Tensorflow 1.14.

The prediction accuracy of the one based on MGU is competitive with LSTM and GRU networks. ReMGU and MCNN-REMGU provide better accuracy than MU and ReMGU respectively.

7. REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model

This paper proposes a solution for predictive text using the Relevance and enhanced entropy based Dempster Shafer (REEDS) approach.

Sometimes the next predicted word is based on different document evidence. Different documents may assign maximum probability to different words. This forbids from taking a decision on the next probable word.

The REEDS algorithm is as follows:
Evidences from different Language Models (LMs) and Body of Evidence (BOE) are taken, Uncertainty measure is calculated, credibility of BOE is calculated and difference between average credibility and credibility of each BOE is calculated. Using these Altered Credibility is calculated. These are normalized to get the weight of each BOE.
Evidence modification is done and the modified BOE is given. Then Evidence fusion takes place and the modified BPAs are fused with Dempster's Rule. This is then used to predict the next word.

The proposed model predicts the next words when given a set of words as input from the mass value of the word from different documents as evidence. It is a hybrid of different evidence with EBE and Dempster's combination rule. REEDS proved to be more feasible than other combination rules such as Dubois rule, Yager's rule, Yager's modification rule as it is able to handle conflicting evidence when different documents assign maximum probability to different words with better convergence and takes the credibility and relevance of the evidence into consideration which other methods aren't able to do. The proposed model is considered to be better than neural network approach due to less computational burden and lingering learning time of the neural network.

8. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading

Two metrics which are used for quantification of word prediction effort are surprisal and next-word entropy. These measures suffer from shortcomings. This paper proposes a new metric called lookahead information gain to overcome the shortcomings of already existing measures.

In order to estimate information theoretic metrics, LSTM recurrent neural network was used. Six LSTMs were used. Each was trained with different initial weights which were drawn randomly from a uniform distribution between +0.1 and -0.1. The biases were initialized as 0 in order to account for the variation.

With more training, lookahead information gain measures became greater on average. This happened regardless of whether or not the probability of the word after the next word was extracted by explicit prediction or base frequencies which was unexpected as better trained models should gain less information with every incoming word.

9. Learning to predict more accurate text instances for scene text detection

Link: https://www.sciencedirect.com/science/article/abs/pii/S0925231221005695

https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S0925231221005695

Scene text detection, especially for multi-oriented text detection or arbitrary shape text detection, is one of the most challenging tasks in computer vision. The progress of deep learning has brought many technical improvements to scene text detection. However, due to the variability of texts, the complexity of background, or the poor shooting circumstance, there is still room for improvement in text detection.

In this paper, a pixel-based text detector is proposed to facilitate the representation and prediction of text instances with arbitrary shapes in a simple manner. Firstly, to alleviate the influence of the target vertex sorting and achieve the direct regression of arbitrary shape text instances, the starting-point independent coordinates regression loss is proposed. Furthermore, to predict more accurate text instances, the text instance accuracy loss is proposed as an assistant task to refine the predicted coordinates under the guidance of IoU. To evaluate the effectiveness of our detector, extensive experiments have been carried on public benchmarks which contain arbitrary shape text instances and multioriented text instances. We obtain 84.8% of F-measure on Total-Text benchmark. The results show that our method can reach state-of-the-art performance.

Their main contributions can be summarized as follows:

The limitations of previous regression-based methods in arbitrary shaped text is discussed and the starting-point independent regression loss is proposed instead of the conventional regression loss to optimize predicted coordinates of text instances. The proposed method solves the problem that the curved text cannot be processed gracefully in the previous pixel-based methods.

The text instance accuracy loss is introduced to obtain text polygons with larger IoU, which further improves performance without increasing the computation of network.

A simple and effective pixel-based method is proposed, which only uses NMS post-processing step. The model is available for arbitrary shape text detection without additional annotation and obtains state-of-the-art performance on Total-Text dataset.

Data set: Total-Text dataset.

10.Scalable and explainable legal prediction

Link: https://link.springer.com/article/10.1007/s10506-020-09273-1

https://sci-hub.se/https://link.springer.com/article/10.1007/s10506-020-09273-1

Recent advances in artificial intelligence (AI) and human language technology (HLT) have created new opportunities to automate routine aspects of case management and adjudication, freeing human experts to focus on aspects of these tasks that most require human judgment and knowledge. An important application of this technology is decision support for routine administrative decision-making and adjudication.

This paper describes two approaches to an important form of legal decision support—explainable outcome prediction—that obviate both annotation of an entire decision corpus and manual processing of new cases.

This paper describes two approaches to explainable legal decision prediction that operate on textual inputs. Each system was prototyped on a collection of 16,024 World Intellectual Property Organization (WIPO) domain name dispute cases.

Data set: WIPO dataset

The first system, ANP (Attention Network-based Prediction), used an Attention Network for prediction and highlights salient case text based on attention weights for decision support. ANP was evaluated as a decision aide for attorney and non-attorney subjects to predict the case decisions. While the case decisions themselves were found to be predictable using this approach, attention-weightbased text highlighting was not shown to improve decision speed or accuracy.

This negative result motivated a second approach, termed semi-supervised case annotation for legal explanations (SCALE), in which the justifcation portions of a representative set of cases were annotated, and tags for factual and legal fndings were propagated to sentences in unannotated cases that shared a high degree of similarity to the annotated sentences in a semantic embedding space. This approach exploits the structural and semantic regularities in case corpora to identify text patterns with predictable relationships to case decisions.

11. The value of text for small business default prediction: A Deep Learning approach

Link: https://www.sciencedirect.com/science/article/abs/pii/S0377221721001983

https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S0377221721001983

Proposed method can speed up the interpretation process by extracting an automated risk score from the text that we show to be predictive.

It is capable of combining these two separate sources of information (i.e. text and structured data) and suggesting a single score, and a corresponding accept/reject decision.

Two types of pre-processing are applied before fitting the three types of models, i.e. Logistic Regression (LR), Random Forests (RF) and Deep Learning (DL).

First, for the DL model, which utilises Google’s pre-trained BERT model (Devlin et al., 2018), the word pieces are mapped to a pre-existing vocabulary in the pre-trained model.

Second, for the LR and RF models, unlike with the DL approach where the text representation learning forms an active part of the model training, such an approach is not directly compatible. Instead, a representation of the text statements must be precalculated using a statistical method to produce a 1-dimensional vector per text statement.

Data set: text (the text alone), structured (standard credit scoring features alone) and combined (both the structured and text inputs).

We seek to assess both the predictive power of the text and of the modelling techniques. Accordingly, we consider three types of model (i.e. LR, RF and DL).

In addition to these benchmark models, three Deep Learning models are trained – one on the structured data only, another on the (pre-processed) text data, and a third on the combined data. However, there are notable differences between the DL architectures chosen for the structured and the text data.

Firstly, at the core of the structured DL model is a simple Multi-layer Perceptron (MLP); however, in the MLP, we implement the relatively new concept of categorical embeddings .

Secondly, for the text-only DL model, we use BERT, a state-ofthe-art language model produced by researchers at Google AI Language .

Thirdly, the Deep Learning model for the combined data extends the model for the structured data with a fixed-length representation vector from the text-only BERT model.

11. Cyberattack detection model using community detection and text analysis on social media

Link: https://www.sciencedirect.com/science/article/pii/S2405959521001685

https://reader.elsevier.com/reader/sd/pii/S2405959521001685?token=77D3F89FF2D5CC19BF4B85372742C7A3121DB691943823093F12D10D40D426F9952C7DE00E759D6205E1CCF0AD491708&originRegion=eu-west-1&originCreation=20220204132050

They aim to quickly detect cyber attacks by tracking the trend of tweets written by cyber attack-related users.

Our methodology has a distinguishing feature from the existing studies in that we incorporate the semantics in Tweets to evaluate the relevance with cyber attacks and adapt community detection to identify the most relevant group to the cyber attacks.

Data sets : (1) the user list related to the cyber security intelligence (CSI User List),

(2) the phrase list related to cyber security intelligence (CSIPhrase List),

(3) Tweets written by the users in CSI UserList (Tweet Data Set),

(4) relationship between the users inCSI User List (Relationship Data Set), and

(5) news articles

The entire process consists of (1) data collection,

(2) data analysis, and

(3) detection.

First, we collect five kinds of data sets. Then, data analysis consists of graph analysis and text analysis. In graph analysis, we classify the entire cyber attack-related users into groups according to community detection. In text analysis, we analyze the similarity between tweets and the cyber attack-related keywords. Finally, we construct a cyber attack detection model by connecting the analyzed results with attack information extracted from CSI News Data Set.

12. Multi-orientation scene text detection with scale-guided regression

Link: https://sci-hub.hkvisa.net/https://doi.org/10.1016/j.neucom.2021.07.026

This paper solves a simple but key problem most text detection algorithms suffer from: accurate drawing of the text boundary. According to the paper, existing algorithms can perform well on pictures of short words but an entire sentence causes them to draw text boxes inaccurately. The existing algorithms struggle to properly define boxes where text length is more than

1400 pixels. Also most of these papers have either classification based prediction or regression based prediction.

The authors of this paper propose a model called Scale-Guided Regression Module (SRM), which uses both classification and regression methods for best results. SRM contains both height-guided and width-guided kernels, which are used to get much higher accuracy of prediction boxes(here classification is used). Upto 10 feature maps are captured which are combined using regression to obtain the bounding box.

Two types of losses are used here.

1) IoU loss: Intersection-over-Union loss is a metric used to get the accuracy of the algorithm on the dataset. It is more formally defined as Area of overlap between ground-truth box and predicted box(intersection)/area of union between ground-truth and predicted box

2) Smooth L1-loss can be thought of as a hybrid of L1 and L2-loss. When the absolute value of the argument is high, it behaves as L1-loss, and when the absolute value of the argument is close to zero, it behaves as L2-loss. The equation is as follows: $L_{1;smooth} = \begin{cases} |x| & \text{if } |x| > \alpha; \\ \frac{1}{|\alpha|}x^2 & \text{if } |x| \leq \alpha \end{cases}$

Dataset used:MSRA-TD500

The MSRA Text Detection 500 Database (MSRA-TD500) contains 500 natural photos acquired with a pocket camera from indoor (office and mall) and outdoor (street) scenarios. The images primarily consist of signs, door plates, and caution plates, guide boards and billboards with elaborate backgrounds of varying resolutions. This is considered as a benchmark to test object detection algorithms

13) FASTPITCH: PARALLEL TEXT-TO-SPEECH WITH PITCH PREDICTION:

Link: https://sci-hub.hkvisa.net/https://doi.org/10.1109/ICASSP39728.2021.9413889

The authors have proposed a model called Fastpitch which is based on FastSpeech. FastSpeech is a feed-forward neural network based algorithm that can generate natural speech and to fine tune the generated speech, an encoder-decoder based model is also used. FastPitch improves upon fastspeech by making the pitch of the generated speech sound more human-like. Frequency of the words is analysed 900 times using transformer architecture.

Traditional TTS based models suffer from slow inference speed, and the synthesised speech is usually not robust (i.e., some words are skipped or repeated) and lack of controllability (voice speed or prosody control). FastPitch improves the quality of the produced speech by estimating the pitch contours for every word and fine tuning it automatically. This basically helps the model to produce speech in various languages and dialects.

Dataset: LJSpeech
This is a public domain voice collection made up of 13,100 brief audio clips of a single speaker reading portions from seven non-fiction books. Each clip comes with a transcription. The clips range in length from 1 to 10 seconds and have a cumulative duration of about 24 hours.The texts were published between 1884 and 1964 and are now in the public domain. The audio was recorded in 2016-17 by the LibriVox project and is also in the public domain.

14. A deep learning based algorithm for multi-criteria recommender systems

Link: A deep learning based algorithm for multi-criteria recommender systems - ScienceDirect

The authors of this paper propose a multi-criteria recommender system. Most recommender systems , according to the paper, rely only on the search history of the user to provide personalised recommendations. The proposed recommender system uses autoencoders and multiple criteria to deeply analyse the type of preferences the user has and also understand how the user interacts with the website.

The most common technique used in recommender systems is collaborative filtering. However this suffers from known problems like inability to capture complex user interaction with the website, and even accuracy of the predictions. Deep learning models can take various aspects of user interaction and understand how the user interacts with the website in a more detailed manner. This paper proposes using deep learning using various user interaction metrics using autoencoder functions.

Datasets used:

1) TripAdvisor dataset
2) Yahoo! Movies
3) Netflix

15. Research on sentiment classification of futures predictive texts based on BERT

Link: Research on sentiment classification of futures predictive texts based on BERT | SpringerLink

The authors of this paper propose a BERT based sentiment analysis model on a stock market dataset to predict sentiment based on the market trends.

This model des not use a traditional dataset, instead uses web crawlers to crawl to the websites of various futures companies and capture customer sentiment.

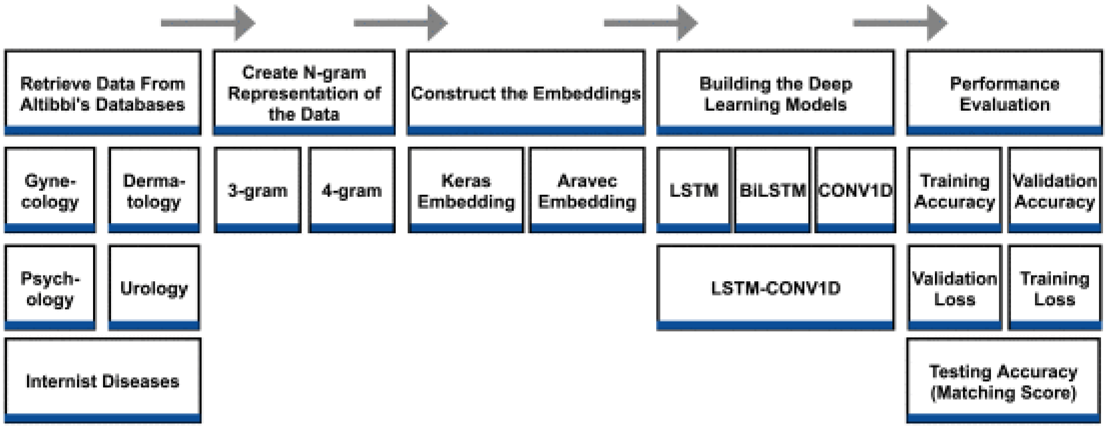
Irregular expression is used to clean the data, leaving only the words behind . The words are then translated to Chinese to obtain the corpus.

Then the BERT model is used to obtain the sentiment analysis and predict future market trends. This is also compared to traditional classification algorithms like random forest.

16)A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context

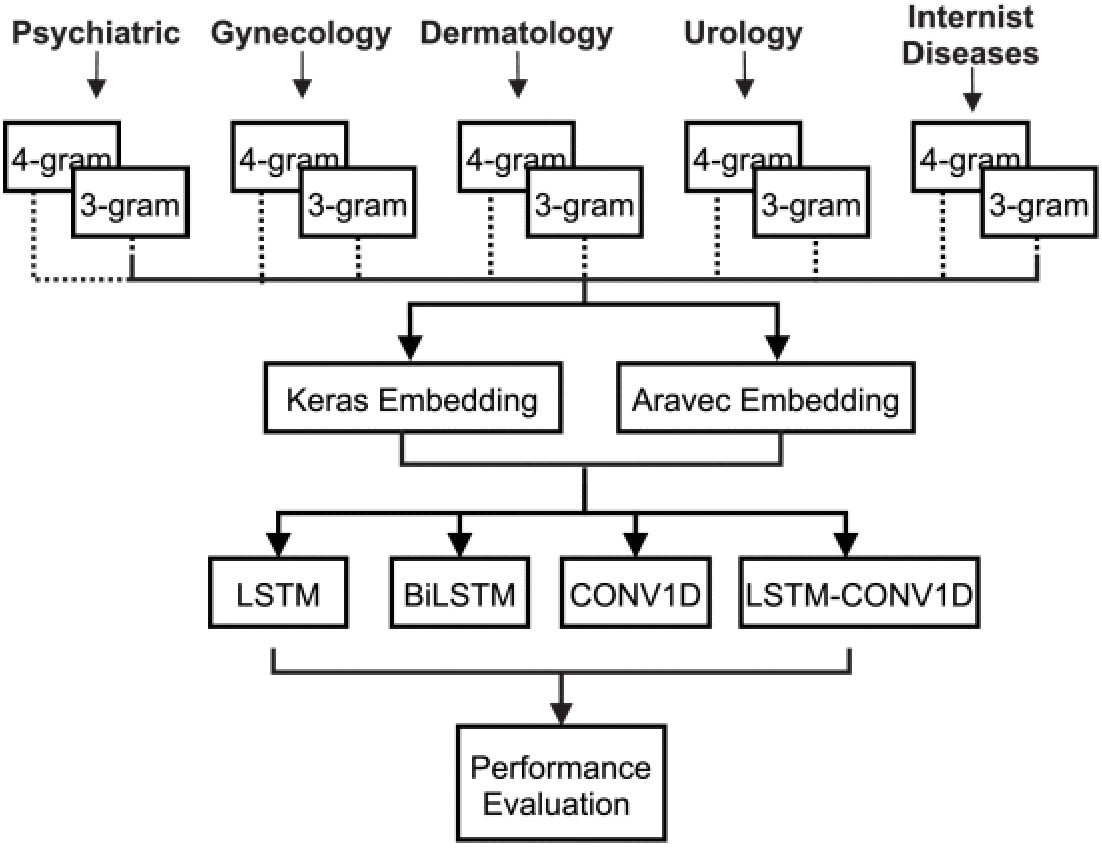
Link:https://ieeexplore.ieee.org/document/9448234

In this paper, Altibbi which is a digital health platform providing telemedicine services accessible to the people by connecting them to doctors. It helps doctors by suggesting the next possible word when they type. So this paper discusses the problem of predicting and generating next word in arabic for doctors so that it saves time and be efficient for both doctors and patients and help improve their interaction. Previously there was a deep reinforcement approach for paraphrase generation which consist of neural generator and evaluator. There was a variational autoencoder framework to generate text at character level. There was a neural text generation model which integrates GAN,RNN and reinforcement learning.There was a model for textual patent generation. BART and T5 were the other models that was used in previous papers.



This figure shows the proposed idea where it has three parts data preparation, system architecture and evaluation criteria.

The Arabic language dataset is obtained from altibbi. 3 gram and 4 gram dataset is used so that it considers other word from history and predict the next word.



These are the methodologies used in the base paper.

Embeddings are low dimensional representation of text represented as real valued vector.

The proposed solution is the mixture of different existing algorithm and has modification to existing proposals.

LSTM solves problem of vanishing gradients experienced by RNN.It consist of forget, input and output gates.BiLSTM includes two distinct LSTM networks. One of them considers historical word while the other considers the future word hence improving understanding of text.

The system is evaluated by testing its accuracy.Each testing sample has 5 possibilties of highest probability.

The future work can be to predict phrases and long sentences rather than just predicting the next word.

17. Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network

LINK:
Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network | IEEE Journals & Magazine | IEEE Xplore

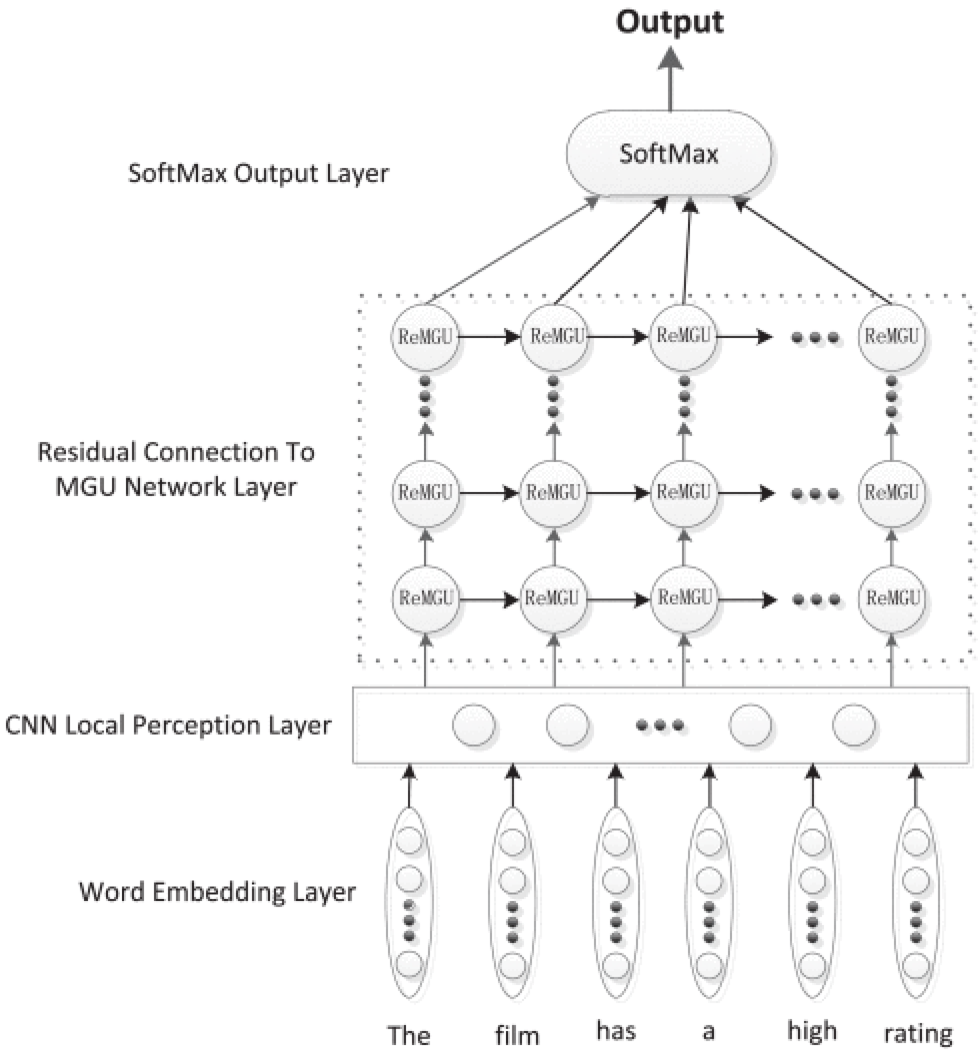
In this paper the problem discussed is regarding vanishing gradient and network degradation which resist us from fully making use of the existing historical information. It also discusses the problem of dependence on the word sequence for next word generation.

RNN finds the connection between the sequence of words and gives a semantic meaning to the sentence. But due to gradient dispersion there is a problem in processing long time information. LSTM is thus used to fully use the historical information. In the previous paper independent RNN was proposed to overcome network degradation.

This paper used the high-dimensional feature extraction capability of convolution kernels to fully obtain the depth feature information between word sequences, Also the residual connection processing to the MGU network to make the network more sensitive to gradient changes, and solved the network degradation problem caused by deep network depth and finally proposed the MCNN-ReMGU model

This solution proposes a new model with major modifications in the existing ones to solve the above mentioned problems.

Dataset: Penn Treebank and WikiText-2 datasets



This is MCNN-ReMGUmodel proposed in this paper.

Word embedding layer marks words from high dimensional to low dimensional vector space.

To solve problem of network degradation this model uses residual connection to MGU network layer.

The softmax finally provides normalized value for net word prediction.

This model further helps to predict the next word by making best use of the existing historical information and is better as compared to existing ones.

For the future it should also make full use of existing sentences to predict next ones.

18. An Approach for a Next-Word Prediction for Ukrainian Language

Link:

https://doi.org/10.1155/2021/5886119

This paper discusses the problem related to the errors in the next word prediction in Ukrainian Language. It also discusses the problem of predicting only one word and not the sentences.

The previous papers asserted the importance of next word prediction and also federated learning to predict the next word successfully.

Also sequence to sequence generation is proposed in previous paper for banglalanguage.In another paper Markov model was used for polish language.

This paper compares two models LSTM and Markov models in predicting next word. After comparing it also checks a hybrid model including both LSTM and Markov chains.

In this model if the word exist in the LSTM list it chooses the word predicted by it otherwise it choses the word predicted by Markov chains.

This model has slight modification from existing ones and is combination of multiple models.

DATASET:UKRANIAN POEMS because they consist of everything and is good for this study.

After checking everything it finds out that Markov chain is the fastest model.For future work this could be checked on other languages as well.

19. Research on sentiment classification of futures predictive texts based on BERT

Link:https://link.springer.com/article/10.1007/s00607-021-00989-9

The paper considers the investing and stock market dataset and discusses the problem of predicting the sentiment in the market by the trend it follows.

This model uses BERT to find the sentimental analysis .
Previous paper consist of the Hybrid algorithm framework for sentiment classification of Chinese based on semantic comprehension and machine learning.

The proposed solution has slight modification only from the previous papers proposals.

The dataset used is particular companies dataset.

The BERT model is better than other models .

Author	Title	Problem addressed	Findings and conclusions	Limitations or weaknesses	How your research can the gap
Lahijan Branch (2016)	Evaluation of Deep Learning Models for Hostility Detection in Hindi Text	It is critical to recognise hostile posts in order to preserve social media hygiene. The problem is particularly acute in languages with limited resources, such as Hindi. In this paper, we offer algorithms for detecting hostile writing in Hindi.	They take a hierarchical approach, separating aggressive and non-hostile posts first. Individual hostile models are then used to assign multi-labels to the post that has been classified as hostile. They don't use any external features for classification, instead relying solely on the post's textual content.We compare Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and Bidirectional Encoder	Random initialization has no evident advantage over pre-trained embeddings. This could be because the target dataset for social media posts differs significantly from the Wikipedia and News datasets used to train these embeddings.	

			Representations from Transformers (BERT) as deep learning methods for text classification. Simple 1D CNN, multi-channel CNN, bi-directional LSTM, CNN+LSTM, mBERT, and IndicBERT are the models employed. We also test Facebook and IndicNLP's Hindi quick text word embeddings in conjunction with CNN and LSTM model		
Amandeep Kaur, Rudra Bohra, Agarwal	DHOT-Repository and Classification of Offensive Tweets in the Hindi Language	While social media provides people with an online forum to communicate their opinions, expertise, experiences, and feelings, a huge issue arises when these interactions become a platform for abusive remarks, comments, and dialogues. Slurs	Using a fastText-based approach, this work proposes a methodology for distinguishing and then classifying offensive and non-offensive text. A Devanagari Hindi Offensive Tweets (DHOT) data corpus was used to train the algorithm. During fastText model	Manually tagging data with the relevant classes is a time-consuming operation, which is a key constraint in this method. There were instances of sarcasm detection, however this was most	Davidson in their research define an abusive text a offensive spe with a vague target and mi intention to h the sentimen of the receive A similar grid search metho originally devised by Soner was us in experimentat

		are used to show scorn, difference of opinion, and in some circumstances amusement, in addition to being offensive in talks. Abusive language can be used to upset someone, to promote a point of view, and in certain situations, to be amusing.	runs, a grid-search strategy was used to modify hyperparameters, which revealed intriguing insights into model accuracy and precision. Our fastText model obtained 92.2 percent accuracy when processing on a desktop class system.	commonly discovered by manually working with the dataset and unclear classes (such as sarcasm, wit, veiled insults, and so on), which means that some instances of abusive material may have been overlooked. This method is not always sufficient to avoid language misuse, as the sheer volume of online content, as well as the rate at which it is posted, can be daunting.	with the processed da
	A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context	The goal of this work is to offer a deep learning-based language generation model that streamlines the process of producing medical recommendations for doctors in an Arabic setting,	The suggested model was trained utilising data from Altibbi databases pertaining to medical recommendations, specifically in the fields of gynaecology, dermatology,	Different datasets had performed best in algorithms. Like When the embedding is Aravec, the BiLSTM performed the best in terms of validation	

		with the goal of improving service satisfaction and patient-doctor interactions.	psychiatry, urology, and internist illnesses. Deep learning models based on unidirectional and bidirectional long short-term memory (LSTM and BiLSTM), one-dimensional convolutional neural network (CONV1D), and a combination of LSTM and CONV1D were constructed and optimised for next word prediction (LSTM-CONV1D)	accuracy.The CONV1D, on the other hand, outperformed the other five specialisations in the Keras embedding.	
	Natural Language Processing Based Sentimental Analysis of Hindi	This paper presents a practical dynamic approach on to find the polarity of any sentence and analyse the opinion of the particular sentence.	It is common knowledge that different languages have their own distinct ways of expressing themselves. The primary distinction between English and Hindi is the structure of the languages. The English language, for example, has an S-V–O (Subject-Verb-Object) structure,	More research has been done in sentimental analysis to pinpoint the restructuring of words in various ways therefore this analysis was insufficient.	

			but Hindi has a S–O–V (Subject-Object-Verb) structure. The fundamental structural difference between English and Hindi has ramifications in determining text polarity. The polarity of the words in the text is affected by the same set of words with minor modifications and changes in word order. As a result, when working with Hindi, a more in-depth linguistic examination is required to undertake Sentiment Analysis.Trained Dataset validation of positive and negative SVO is observed on the basis of which SAH procedures are marked and tested to conform the necessity and applicability under various field of research.		

3. Conclusion

We explore how natural language processing can be applied to predict the next text in Hindi Language. Using LSTM [2] to predict the most favourable word to be predicted in Hindi Language. In this paper, we attempted to implement a next-word prediction algorithm for the Hindi language. We used LSTM so that it would be able to remember the last couple of words before predicting the next word. We had done prediction of text in hindi language and our model Using LSTM 10 epochs and 64 batch size and by applying cross entropy loss function, we got the loss function was 0.4264. As there is not much research done on Text prediction in Hindi Language, We can not compare our model as a whole to past methods/models in this field of text prediction in Hindi language as it doesn't follow roman script but the devnagri script. But our model is successful in predicting the next text in hindi language.

4. References

[1] Min Liang, Jie-Bo Hou, Xiaobin Zhu, Chun Yang, Jingyan Qin, and Xu-Cheng Yin. 2021. Multi-orientation scene text detection with scale-guided regression. Neurocomput. 461, C (Oct 2021), 310–318. DOI:https://doi.org/10.1016/j.neucom.2021.07.026

[2] Lancucki, A. (2021). *Fastpitch: Parallel Text-to-Speech with Pitch Prediction. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* doi:10.1109/icassp39728.2021.941

[3] Shambour, Qusai (2020). A deep learning based algorithm for multi-criteria recommender systems. Knowledge-Based Systems, (), 106545–. doi:10.1016/j.knosys.2020.106545

[4] Xiaofeng, W., Jinghua, Z., Chenxi, J., & Yiying, J. (2021). *Research on sentiment classification of futures predictive texts based on BERT. Computing.* doi:10.1007/s00607-021-00989-9

[5] Wangchuk, K., Riyamongkol, P., & Waranusast, R. (2021). Next syllables prediction system in dzongkha using long short-term memory. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.01.001

[6] J. Yang, H. Wang and K. Guo, "Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network," in IEEE Access, vol. 8, pp. 188036-188043, 2020, doi: 10.1109/ACCESS.2020.3031200.

[7] Prajapati, Gend & Saha, Rekha. (2019). REEDS: Relevance and Enhanced Entropy based Dempster Shafer Approach for Next Word Prediction using Language Model. Journal of Computational Science. 35. 10.1016/j.jocs.2019.05.001

[8] Li, X., Liu, J., Zhang, G., Huang, Y., Zheng, Y., & Zhang, S. (2021). Learning to predict more accurate text instances for scene text detection. Neurocomputing, 449, 455-463.

[9]Branting, L. K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., ... & Liao, B. (2021). Scalable and explainable legal prediction. Artificial Intelligence and Law, 29(2), 213-238.

[10]Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research, 295(2), 758-771.

[11] Cyberattack detection model using community detection and text analysis on social media

[12] Vikas Kumar Jha, Hrudya P, Vinu P N, Vishnu Vijayan, Prabaharan P, DHOT-Repository and Classification of Offensive Tweets in the Hindi Language,

[13] M. Habib, M. Faris, R. Qaddoura, A. Alomari and H. Faris, "A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context," in *IEEE Access*, vol. 9, pp. 85690-85708, 2021, doi: 10.1109/ACCESS.2021.3087593.

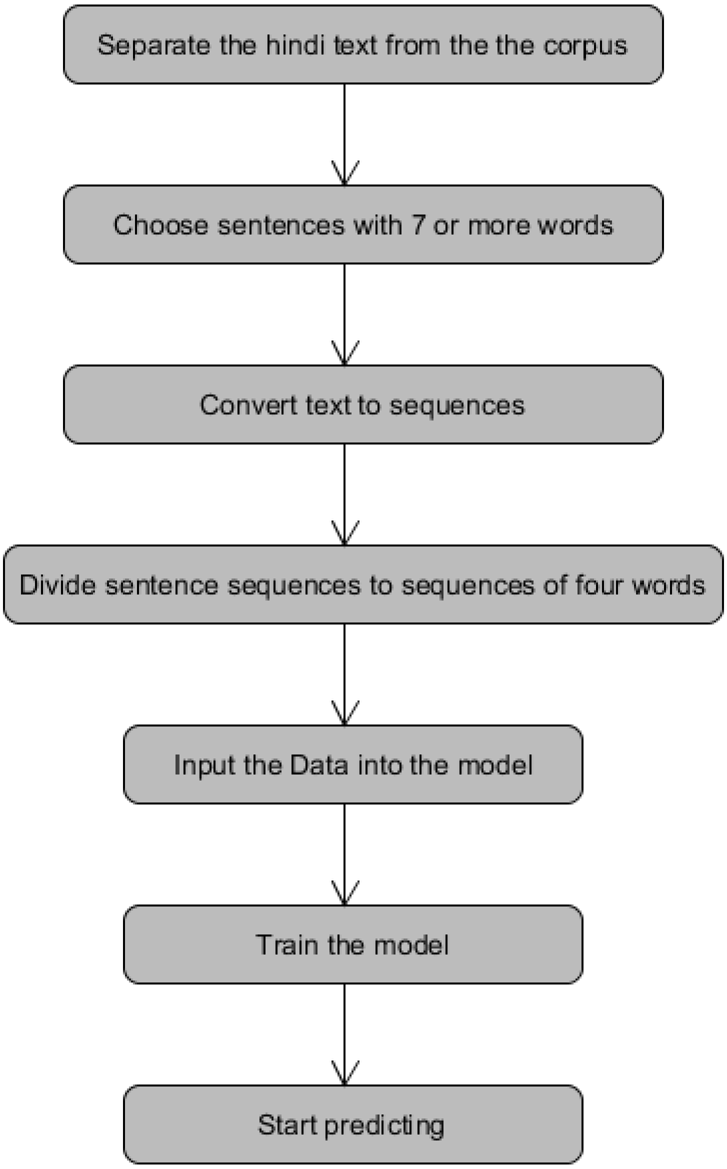
[14] R. Joshi, R. Karnavat, K. Jirapure and R. Joshi, "Evaluation of Deep Learning Models for Hostility Detection in Hindi Text," *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, pp. 1-5, doi: 10.1109/I2CT51068.2021.9418073.

[15] Shrestha, Hewan; Dhasarathan, Chandramohan; Munisamy, Shanmugam; Jayavel, Amudhavel (2020). *Natural Language Processing Based Sentimental Analysis of Hindi (SAH) Script an Optimization Approach. International Journal of Speech Technology, (), –.* doi:10.1007/s10772-020-09730-x

[16]An Approach for a Next-Word Prediction for Ukrainian Language https://doi.org/10.1155/2021/5886119

[17] A Predictive Text System for Medical Recommendations in Telemedicine: A Deep Learning Approach in the Arabic Context https://ieeexplore.ieee.org/document/9448234

4.Methodology



Separate the hindi text from the the corpus
Choose sentences with 7 or more words
Convert text to sequences
Divide sentence sequences to sequences of four words
Input the Data into the model
Train the model
Start predicting