RESEARCH OF NEXT WORD PREDICTION MODEL

Mrs. T. Sai Kumari[1]., K. Sathwik[2]., A. Anish reddy[3]., V. Aravind[4]., S . Ajay reddy.[5]

[1] Assistant Professor, [2,3,4,5]Students B.Tech -IT, (20S11A1235, 20S11A1207, 20S11A1208,20S11A1202

Malla Reddy Institute of Technology and Science, Maisammaguda.,Medchal.,Telangana,India

[1]saikumari,thakur@mrits.ac.in, [2]sathwikkodimyala@gmail.com, [3]anishamballa317 @gmail.com

[4]Velpulaaravind100@gmail.com,[5]ajayreddysurakanti0494 @gmail.com

ABSTRACT

We present a language model based framework for instant messaging, that can predict probable next word given a set of current words. Our goal is to facilitate the task of instant messaging by suggesting relevant words to the user. Generally, at the time of sending personal messages, a user follows a specific style of communication with a specific group of people. This phenomenon is much more evident in other languages apart from English. For example, in Bengali language, there are three counterparts of you, that is used to address the second person in English. Considering this fact, there are at least three styles of writing texts in Bengali language: informal, semi-formal and formal. Therefore, it is quite necessary to generate next words based on the linguistic style adopted by a user, when sending messages to a specific set of people. In this paper, we develop a solution to this issue by adopting different language models when exchanging messages with different groups of people. Our method clusters language models based on user interactions, and we show the effectiveness of our method using a popular metric hit ratio. This model can be widely adapted for predicting next words in smart-phone devices and expedite the communication between users, specifically at the time of phonetic typing.

1. INTRODUCTION

In this world embraced with social media, people have conversation with each other almost every day as informal dialogue exchange, i. e. chatting or Instant Messaging (IM) and it has become one of the mostly used paradigms for communication. People from different countries mostly administer phonetic typing while having casual conversation through messaging or chatting. In order to facilitate and expedite the task of phonetic typing, most 1M software includes a component, that predicts and suggests a set of words, given the current input words of the message sender. Thus the next word prediction component would be actually beneficial for everyday use by reducing the time consumed for typing. In reality, human communication is predominantly personalized i.e. a person internally uses and manages a specific dictionary for finding appropriate words to chat with another specific person. Languages like Bengali is exposed to more personalized level of communication. For example, there are three different words for denoting a person: tumi, tui and apni, in Bengali against a single word you in English. Based on these three types of addressing, for a single sentence in English, there can be three possible sentences in Bengali with the same meaning the personalized prediction system would be a contribution to gear up the typing speed while having informal computer-based communication; especially for the case of languages like Bengali. Thus the need for constructing personalized language based statistical models can never be obviated and undervalued.

2. LITERATURE SURVEY

Michael I. Jordan and Tom M. Mitchell, "Machine learning: Trends perspectives and prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015. [1] states that Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science.

Abhaya Kumar Sahoo, Chittaranjan Pradhan and Himansu Das, "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making", Nature inspired computing for data science, pp. 201212, 2020. [2] states that "There is a growing demand for detailed and accurate landslide maps and inventories around the globe, but particularly in hazard-prone regions such as the Himalayas. Most standard mapping methods require expert knowledge, supervision and fieldwork. In this study, we use optical data from the Rapid Eye satellite and topographic factors to analyze the potential of machine learning methods, i.e., artificial neural network (ANN), support vector machines (SVM) and random forest (RF), and different deep-learning convolution neural networks (CNNs) for landslide detection.

N Keerthana, S Harikrishnan, M Konsaha Buji and J B. Jona, "Next Word

Prediction", International Journal of Creative Research Thoughts (IJCRT), vol. 9, no. 12, December 2021. [3] states Text generation, in particular, next-word prediction, is convenient for users because it helps to type without errors and faster. Therefore, a personalized text prediction system is a vital analysis topic for all languages, primarily for Ukrainian, because of limited support for the Ukrainian

language tools. LSTM and Markov chains and their hybrid were chosen for next-word prediction. Their sequential nature (current output depends on previous) helps to successfully cope with the next-word prediction task. The Markov chains presented the fastest and adequate results. The hybrid model presents adequate results but it works slowly. Using the model, user can generate not only one word but also a few or a sentence or several sentences, unlike T9.

Gend Lal Prajapati and Rekha Saha, "REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model", Journal of

Computational Science, vol. 35, pp. 1-11, 2019. [4] states that Next word prediction (NWP) is an acute problem in the arena of natural language processing. The recent approaches are solely based on the probability distribution of the Language Model. Word prediction is the problem of guessing which word is likely to continue a given initial text fragment. A Language Model consists of a number of documents and each document consists of a set of words. Each document will generate a group of evidence based on the information contained in it. The main task is to fuse the evidence to get a reasonable result. In this paper, a novel relevance and enhanced Deng entropy based Dempster's combination rule is proposed where various documents act as evidence source namely relevance and enhanced entropy based Dempster Shafer approach (REEDS) to predict the next probable word from the Language Model.

Qu Xiaoyun, Kang Xiaoning, Zhang Chao, Jiang Shuai and Ma Xiuda, "Short-term prediction of wind power based on deep long short-term memory", 2016 IEEE PES AsiaPacific Power and Energy Engineering Conference (APPEEC), pp. 1148-1152, 2016. [5] states that This paper proposes a wind power prediction model based on the Long Short-Term Memory model, one of the deep learning method. Deep learning conforms to the trend of big data and has powerful capability of learning and generalization for mass data. Principal component analysis (PCA) is used to choose input samples and reduce the dimensions of the input variables of the LSTM prediction model based on numerical weather prediction(NWP) data.

Eloisa Vargiu and Mirko Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising", Artif. Intell. Res., vol. 2, no. 1, pp. 44-54, 2013. [6] states that Web scraping is the set of techniques used to automatically get some information from a website instead of manually copying it. The goal of a Web scraper is to look for certain kinds of information, extract, and aggregate it into new Web pages. In particular, scrapers are focused on transforming unstructured data and save them in structured databases. In this paper, among others kind of scraping, we focus on those techniques that extract the content of a Web page. In particular, we adopt scraping techniques in the Web advertising field.

S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term

Memory", 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1-4, Oct. 2020. [7] states that The sentiment analysis is an emerging research area where vast amount of data are being analyzed, to generate useful insights in regards to a specific topic. It is an effective tool which can serve governments, corporations and even consumers. Text emotion recognizing lays a key role in this framework.

Y. Kim, J. An, M. Lee and Y. Lee, "An Activity-Embedding Approach for Next-Activity Prediction in a Multi-User Smart Space", 2017 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 1-6, May 2017. [8] states that Since the advent of the IoT era, various IoT devices have proliferated, transforming ordinary spaces into smart spaces such as smart home, smart office, and smart building. To provide user-friendly service to people, the majority of previous studies have focused on activity recognition and prediction in singleuser environments such as ambient assisted living (AAL) and activities of daily living (ADL).

J. Pennington, R. Socher and C. Manning, "Glove: Global Vectors for Word Representation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014. [9] states that Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global logbilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods.

R. Sharma, N. Goel, N. Aggarwal, P. Kaur and C. Prakash, "Next Word Prediction in Hindi Using Deep Learning Techniques", 2019 International Conference on Data Science and Engineering (ICDSE), pp. 55-60, 2019. [10] states that Natural Language Generation (NLG) focuses on the generation of natural, human-interpretable language. This study proposes a novel methodology to predict the next word in a Hindi sentence. By predicting the next word in a sequence, the number of keystrokes of the user can be reduced.

S. Siami-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series", 2019 IEEE International Conference on Big Data (Big Data, pp. 3285-3292, 2019. [11] states that Machine and deep learning-based algorithms are the emerging approaches in addressing prediction problems in time series. These techniques have been shown to produce more accurate results than conventional regression-based modelling.

3. DATASET

There are several data sets you can check out for next word prediction model. Some Popular ones include the penn treebank dataset, the WikiText dataset, and the CommonCrawl dataset. Each of these datasets contain a laege corpus of text that you can use for training your model.

4. METHODOLOGY

Existing System

The reviewed methodologies for personalized word prediction emphasizes mostly on estimation based upon the built-in statistical language models, which are consisted of using the dictionary of a particular language. However, for phonetic typing the task is not that simple as the spelling of a word may differ from user to user and deviate from the standard form, if one is considered as standard. Apart from this, several other issues compound the task and we look forward to developing a personalized next word predictor as a remedy to this situation. Based on the problems mentioned above, we would like to address the following research questions in this project: How personalized statistical language models can aid the next word prediction component of IM? Will statistical language models be more useful for other languages apart from English when performing phonetic typing? Can we actually help users in Bengali phonetic typing using statistical language models? How can we quantify user satisfaction?.

Proposed System

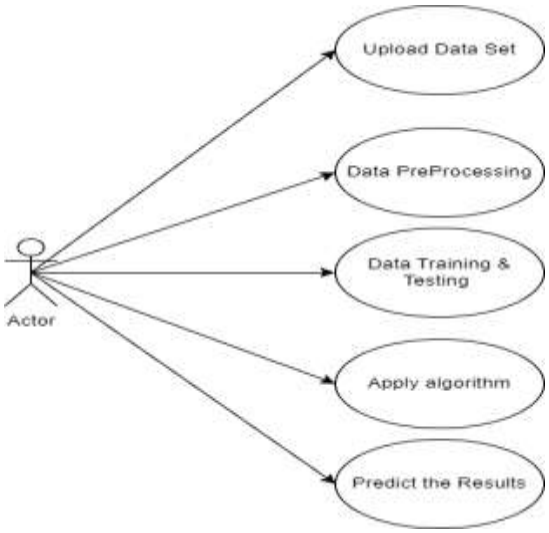
In order to answer these questions we would like to develop a smart-phone based typing solution which would take assistance from person specif
ic statistical language models and the system would learn those models from the existing chatting data of the user. The prediction system could use the existing system's dictionary without possessing an embedded word list and it could also learn instantly from previous messaging history, personal documents and chat logs to be convenient from the initiation. In case of one user or sender, the concept of making the system more user-friendly and personalized is to provide different suggestions for different receivers.

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standardis managed and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object- oriented computer software. In its current form, UML iscomprised of two major components: a Meta-model anda notation. In the future, some form of method or process may also be addedto; or associated with, UML. The Unified Modeling Language is a standard languagefor specifying, Visualization, Constructing and documenting the artefacts of software systems, as well as for business modelling andother non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems. The UML is a very important part of developing object- oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM

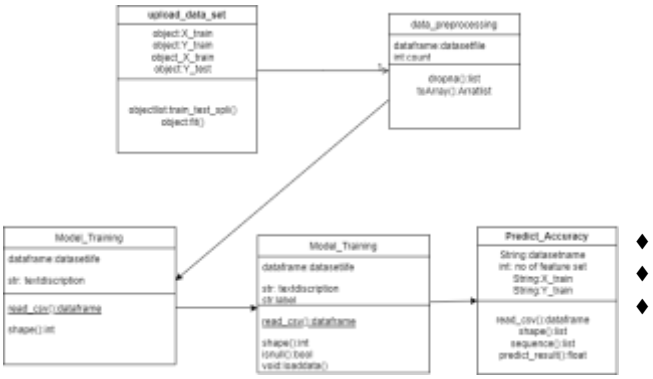
A use case diagram in the Unified Modeling Language(UML) is a type of behavioral diagram defined by andcreated from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of theactors in the system can be depicted.



USE CASE DIAGRAM

CLASS DIAGRAM

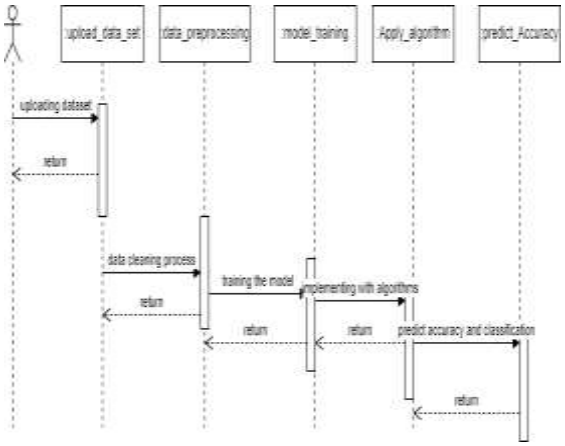
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of staticstructure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



CLASS DIAGRAM

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



SEQUENCE DIAGRAM

SYSTEM STUDYFEASIBILITY

STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY
TECHNICAL FEASIBILITY
SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placedon the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail unacceptably. There are various types of tests.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expectedresults.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields.

Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems thatarise from the combination of components

Functional testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration- oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most otherkinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under testis treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Integration testing

Software integration testing is the incremental integration testing of two or more integratedsoftware components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that

components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participationby the end user. It also ensures that the system meets the functional requirements.

5. ACKNOWLEDGEMENT

The team members of the research project want to sincerely thank our guide Assistant Professor Mrs. T. Sai Kumari and the Department of Information Technology, Malla Reddy Institute of Technology and Science, India for their encouragement and support forthe completion of this work.

6. CONCLUSION AND FUTURE SCOPECONCLUSION

CONCLUSION

The research could lead to the development of a user oriented and very own flavoured word predictor for instant messaging, by which people who use computer-based communication can be significantly assisted. The ever-growing field of social media and instant messaging have created the necessity to design a system that could come to support for fast, comfortable, and smooth typing. Moreover, we would like to argue that this system would be much more effective for our language English, as it is inherently complicated with different salutation styles and thus our approach can really bolster 1M communication and increase number of words typed per minute.

FUTURE SCOPE

The future of next-word prediction in natural language processing is poised to become significantly more sophisticated, with a host of potential enhancements that could revolutionize the field. As we move forward, we expect to see advancements in contextual understanding that go beyond word and sentence levels, enabling systems to comprehend the entirety of documents and complex conversations. The integration of multimodal data, such as images and videos, will likely play a crucial role in providing richer context for more accurate predictions.

7. REFERENCES

[1] Bengio, Y., Simard, P., Frasconi, P., 1994.Learning long dependencies with gradient descent is troublesome. IEEE transactions on neural networks five, 157– 166.

[2] R. Kneser and H. Ney, "Improved backing-off for n-gram language modeling", Conference on Acoustics speech and Signal Process, pp. 181-184, 1995.

[3] Mohd. Majid and Piyush Kumar, Language Modelling: Next word Prediction, 2019.

[4] Bengio, Y., Simard, P., Frasconi, P., 1994., Learning NLP with gradient descent is troublesome. IEEE transactions on neural networks five, 157–166.

[5] Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue mistreatments generative stratified neural network models. In Proceedings of the 30th Conference on Artificial Intelligence. AAAI.

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

[7] M. K. Sharma, S. Sarcar, P. K. Saha and D. Samanta, "Visual clue: Associate approach to predict and highlight next character," 2012 fourth International Conference on Intelligent Human pc Interaction (IHCI), Kharagpur, 2012, pp. 1-7, doi:10.1109 /IHCI.2012.6481820.

[8] Sukhbaatar, S., Weston, J., Fergus, R., et al., 2015. End-to-end memory networks, in: Advances in neural information processing systems, pp. 2440–2448.

[9] Zhou, C., Sun, C., Liu, Z., Lau, F., 2015. A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630.

[10] Joel Stremmel, Arjun Singh. (2020). Pretraining Federated Text Models for Next Word Prediction using GPT2.

[11] S. M. Sarwar and Abdullah-Al-Mamun, "Next word prediction for phonetic typing by grouping language models," 2016 2nd International Conference on Information Management (ICIM), London, 2016, pp. 73-76, doi: 10.1109/ INFOMAN.7477536.

[12] J. Yang, H. Wang and K. Guo, "Natural Language Word Prediction Model Based on MultiWindow Convolution and Residual Network," in IEEE Access, vol. 8.pp.188036-188043, 2020, doi: 10.1109 / ACCESS.20202.3031200.