An Anatomization Of Language Detection And Translation Using NLP Techniques

Jagadeesh sai D[1] and Krishna Raj P M[2]

[1]Depatment of Information Science and Engineering

[2]Ramaiah Institute of Technology, Bengaluru, India

[1]djsai@msrit.eduand [2]krishnarajpm@msrit.edu

ABSTRACT

The issue with identifying language relates to process of determiningnatural language in which specific text is written. This is one of the big difficulties in the processing of natural languages. Still, they also pose a problem in improving multiclass classification in this area. Language detection and translation a significant Language Identification task are required. The language analysis method may be carried out according to tools available in a particular language if the source language is known. A successful language detection algorithm determines the achievement of the sentiment analysis task and other identification tasks. Processing natural language and machine learning techniques involve knowledge that is annotated with its language. Algorithms for natural language processing must be updated according to language's grammar. This paper proposes a secure language detection and translation technique to solve the security in natural language processing problems. Language detection algorithm based on char n-gram based statistical detector and translation Yandex API is used. While translating, there should be encryption and decryption for that we are using AESAlgorithm

1. INTRODUCTION

Language Recognition (LR) is a method in which the substance of the document is written in a natural language. Language identification is a wide-ranging research field because it is mostly considered in applications for natural language (NL) identification, like machine translation, informationretrieval, summary as well as answering questions, etc. They need prior identification of the language before processing. Identification of language falls into 2 methods: 1) non-computational,and 2) computational. Non- computational methods require authors to havingenough knowledge of language to be recognized, diacritics& symbols, the most common words utilized characterscombination, & so on.

In contrast, computational methodsdepend on statistical techniques to solve related problems rather than linguistic knowledge. Fast growth of less well-known languages on Internet has generated requirements for LR for applications such asmachine translation, spell checking, multi-lingual information retrieval, etc.Three factors complicate this task: several sizes of character set utilized to encode diverse languages,use various character sets for single language, also more than one language sharing the same script[1][2]. Automatic treatments of these texts, for any purpose requiring Natural Languageprocessing.

Such as WWW indexing and interrogation or providing reading aids necessitates a preliminary identification of the language used. For example, morphological based stemming has proven essential in improving information retrieval, and applying language-specific algorithms implies knowing the language used.In developing new digital goods and services, data is now a sort of capital, on a par with financial and human capital. Everyone is overloaded with information overload due to the proliferation of information innews, medical records, corporate files, court hearings, government papers,as well as social media. The bulk of this data is unstructured, i.e., free text, making it hard to have reasoning and interpretation.

A. Sentiment Analysis (SA) for LanguageIdentification

Sentiment Analysis (SA) was one of the fields of quantitative study in the production of natural languages [3].SA usually performs the processing of knowledge relevant to emotions or beliefs. From a community for a given subject. Furthermore, opinions havebeen obtained at the document level from specific applications. SA has gained prominence in many fields, including politics [4], business, and marketing. It has believed that the records will contain views while undertaking SA. However, for so many cases, only factual information and evidence are set out in such papers (the news document is one such example). Even materials that are supposed to contain feelings (opinions) that often include descriptive sentences as a part of them. Hence, the most crucial aspect of SA is the recognition of the form and essence of sentences. Thus, the sentences have extracted, classified, and included in the given analysis, depending on the subjective or objective. Classification of subjectivity is the critical activity at SA, which provides for the classification of sentences as factual or subjective. Generally speaking, SA requires several complicated processes. The analysis has accompanied by a set of activities, including the designation of emotions, individualinterpretation, the extraction of opinion holders, and the extraction of aspects or objects[5]. The subjective research includes evaluating the same as subjective or objective as a document or a word to mark. Those documents or phrases classified as objective are automatically discarded after this stage because they are not very useful for theSA[6][7].

B. LanguageProcessing

Artificial intelligence is now widely debated as a buzzword and is rapidly evolving. AI is computer programs that can do something smart like a person. It's merely a machine that mimics human beings to execute tasks in his absence and, generally speaking, often in improved and productive way. Machine learning is an AI subgroup of AI.

Using machine learning, machine intelligence is enhanced by learning algorithms and analyzing various types of data. Machine learning is a subset of Deep Learning and Neural Networks. According to the performance obtained, deep learning algorithms repeatedly analyze various data sets through algorithms and enhance machine intelligence. In the field of computer science, the analysis of natural languages is an important part of machine learning and computational linguistics. Natural Language Processing (NLP) field includes the development of computer systems with natural and human language to perform meaningful tasks. In the future, NLP is so important because it allows one to construct models and procedures, which embrace information as an input or as a voice or a word or both and exploit it on an algorithm in the machine.Input can then be voice, text, or picture where both speech and written output of an NLP [16][17] device can be processed. Different algorithms produced to improve the effectiveness of the text type processing of the language that we will address hereare:

- Sequence 2 Sequenceframework
- The long short termmemory
- User preference graphframework
- Named Entity Recognitionframework
- Feature-based sentence extraction by fuzzy inferencerules.
- A template-based approachbyan automatic text summarization
- Word Embeddingmodel [8].

C. Language Identification

Identification of a language typicalmeans process that tries to categorizetext intoa predefined set of available languages in a language. It is a crucial technique for NLP, mainly in the operating text dependsupon language and classification. Excellent results have been obtained by several researchers[9][10][11][12] on language Recognition rely ona few European languages picked. But, most Asian &African languages also staynot tested. This highlights that the search engines have even less supporting for most African & Asian languages in their language-specific searchingcapability.For correct language categorization, all LSE properties are essential. Also, to defineaccurate tools for text processing atthe last level, LSE identification is necessary. To choose a good translator to translate source text into an additional language, a computer translation tool must first learn the script.

D. N-gram

An n-gram mayseem from longer sequences as a sub- sequence of N objects. A word, letter, syllable, or some logical data form specified by the application can be referred to as the item described. As it is simple to apply and determine next possible succession from the known sequence with great accuracy, the n-gram probability model is one of the very successful NLP methods for statistical results. The main principle of just using n-gram is that each language has its specific n-grams and that these n-grams are often used much more than other languages, offering a vocabulary hint. A monogram is named as n-gram order 1 (that is n=1); n-gram of order 2 is called bigram, n-gram of order 3 is called trigram. N-gram number 2 is sometimes calledbigram.The"rest"isusuallycalled"n-gram.""Use"

via space) for the character-level trigrams and bigrams will be as follows:

Bigram: No o- -4 45 56

Trigram: No- o-4 -45 456

Numerous authors [13][14] announced that the best language recognition result was obtained using the trigram model on preferred European languages. But, several African & Asian languages have not dependedupon Latin script which usesseveral European languages [15]. Therefore, the analysis tests the efficacy of the n-gram orders (n=1, 2,3,..., 6) &unique n-gram mix framework for language orders. Recognition in preferred languages. The paper is structuredinthe following sections: Relevant history and related research are given in Section II. Section III addresses our problem statement, proposed system model, and algorithm. Section IV discusses the effects of the simulation and the analysis. The article is finalized in Section Vand Section VI, which determines the futurescope.

PROPOSEDMETHODOLOGY

A. ProblemStatement

The critical challenge is the overload of information, which presents a significant problem with accessing a particular, relevant piece of data from vast datasets. Due to consistency and usability problems, semantic and meaning comprehension is essential and challenging for summary systems. It is also crucial to identify the context of interaction b/wobjects&entities, mainly with high- dimensional, heterogeneous, complex,as well as poor- quality data.Because there is no aspect to encrypt and decrypt data in a secure format. To find a relationship between objects&entities, semantics are essential. Extraction of text and visual data by persons and objects could not offer reliable information unless the interaction's meaning & semantics were known. Often, instead of keyword-based search, search engines currently available will search for items (objects or entities). Semantic search engines are needed since user queries traditionally written in natural language are betterunderstood.

B. Methodology

Methodologies used throughout comprise data extraction, NLP, and ml techniques that play an essential role in deciding language detection. Language detection algorithm based on char n-gram based statistical detector and translation Yandex API is used.The Yandex online machine translation tool can be accessed from this API. It can translate separate words or full texts in over 90 languages. The API allows Yandex to be implemented.While translating, there should be encryption and decryption for that we are using AES Algorithm.The main aim to get a secure detection and translation environment.

C. Material andMethod

In recent years, NLI has focused the attention of many authors and researchers. With the influx of new researchers, the most substantive research in this area has contributed to the joint mission. The role focuses on the recognition of a writer's native language based on his writing in another language. The second language, in this case, was English. The task was to predict a writer's native language from the provided text / XML file containing English-language Facebook comments.Four languages were proposed to include for this task. They were English, Hindi, Urdu, Gujarati.

a) Dataset

The task's training dataset was XML files, which contain FB comments in English by different native language speakers. XML files were annotated as EG HI, UR, GJ for English, Hindi, Urdu, Gujarati language, respectively.

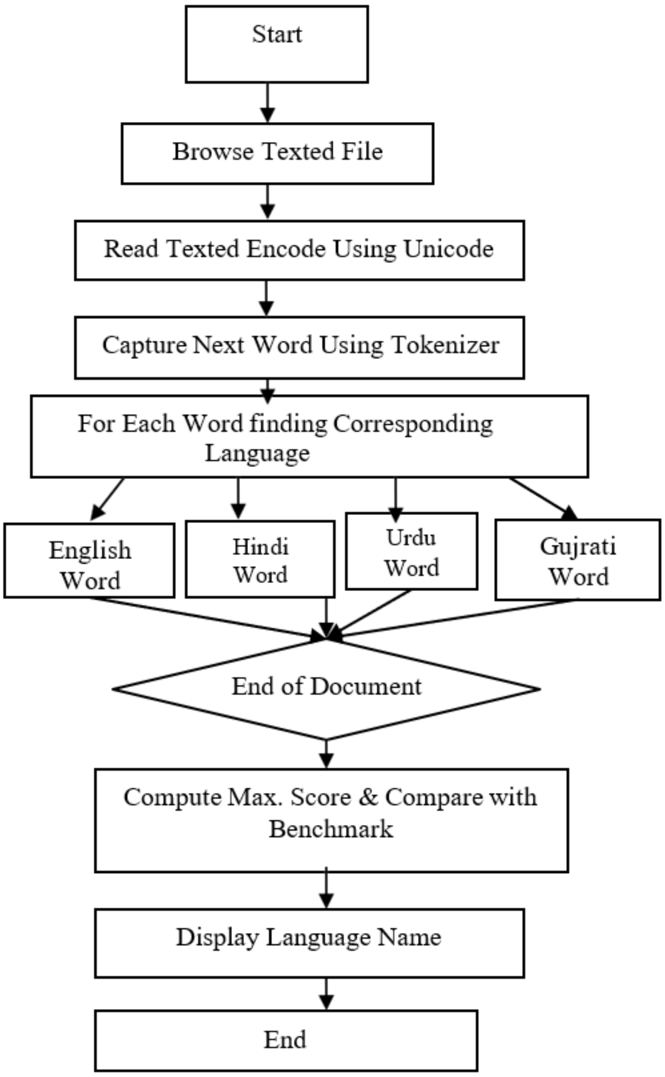
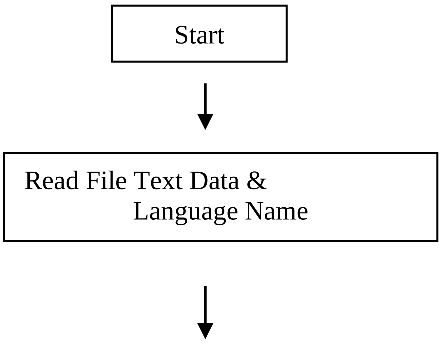

Start
Browse Texted File
Read Texted Encode Using Unicode
Capture Next Word Using Tokenizer
For Each Word finding Corresponding Language
English Word
Hindi Word
Urdu Word
Gujrati Word
End of Document
Compute Max. Score & Compare with Benchmark
Display Language Name
End

Fig 1: Flow Chart for Language Detection

The above Fig.1 describes the proper step by step approach for the Language Detection process. A language detection algorithm is used the determine the language of a given text. Some languages can be determined reliably from their script alone. A widely used approach is supervised machine learning algorithms based on character n-gram based statistical detector.



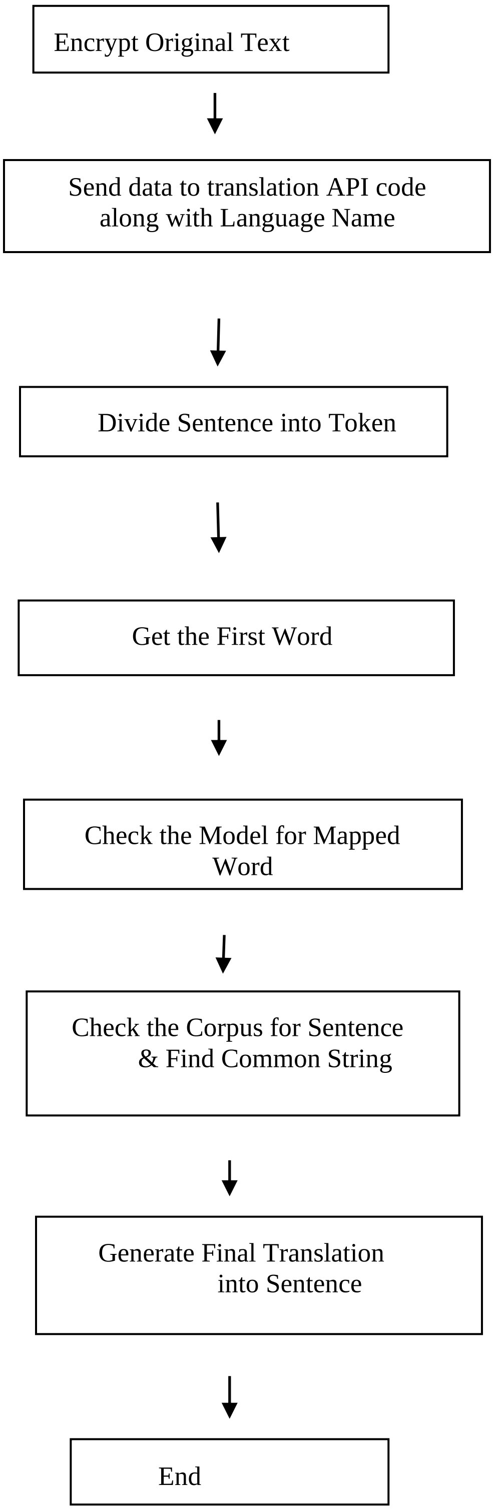

Encrypt Original Text
Send data to translation API code along with Language Name
Divide Sentence into Token
Get the First Word
Check the Model for Mapped Word
Check the Corpus for Sentence & Find Common String
Generate Final Translation into Sentence
End

Fig 2: Flow Chart for Language Translation

The above Fig.2 describes the proper step by step approach tothe Language Translation process.For several NLP functions, such as text summarization, speech recognition, DNA sequence modeling, among others, Sequence-to- Sequence (seq2seq) models are used. Our goal is to translate those phrases from one language to another.

AES

AES is a symmetric block cipher selected to safeguard sensitive information by the U.S. government.To encrypt confidential data, AES is executed in hardware &softwarearound the world. Cybersecurity and electronic data protection are essentialgovernment information security.AES was first developed by NISTin 1997 to substitute the Data Encryption Standard (DES)at that time when it realized the needtohappen tosusceptible to the brute-force attacks (BFAs). ES contains 3 block ciphers: AES-128, AES-192 & AES-256. To decrypt &encrypt messagesblock, AES-128, AES-192 and AES-256utilizes128-bit, 192-bit &256-bit key size to encrypt & decrypt messages, respectively. Every cipher encrypts & decrypts data into 128-bit blocks with 128, 192, and 256-bit encryption keys. Ciphers use the same encryption and decryption key, which is also known as the secret key, and both the sender & the receiver must know and also use the similar private key. Knowledge is categorized into three groups by the government: confidential, hidden, or top secret. To safeguard the Confidential and Secret stage, all key lengths can be used.

D. Tokenization

Extracting words from the text may appear to be a simple task. The top-down method breaks book on whitespace characters such as space, Tab, or a punctuation character. Nonwhite spacenames are concatenated to form a word or token. The bottom-up method builds tokens one character from a text stream until a nontoken character is encountered. The simplest definition of a token is any consecutive string of alphanumeric characters. Between tokens, we find one or more nontokencharacters.

E. LanguageTranslation

Once the Language Identification task is completed, then the next task is to translate the document. Machine translation

It is the translation method into the target language from the source language. The following is a list of problems when attempting to do machine translation that one has to face. Not all words have corresponding words in onelanguage.

- In different tongues inspecific examples, a word in one language must be represented in another by a group of words. 2 provided languages may havedifferent.
- Structures. English has an SVO structure, for instance, whilstTeluguorKannadahavean SOV structure. There is also a shortage of one-to-onecommunication.
- Speech sections for two languages. Kannada / Telugu color words, for instance, are nouns,

although they are adjectives in English. The forms in which sentences are placed together vary between languages, too. Words may have more than one meaning and sometimes ameaning.

- In a language, a group of words / an entire sentence can have above one meaning. Ambiguity is called this problem. Not all problems with translation can be explained by applying.
- Grammar'svalues. It's too tricky for software programs to forecast
- Meaning. Translation includes not only grammar &vocabulary however also information collected by previousexperience.

Algorithm:

1: Start

2: Browse Text File

3: If detect language=yes, then goto step 5 4: Ifsee language=No, then goto step 1

5: Detect Language API

6: EncryptOriginal Text Using AES

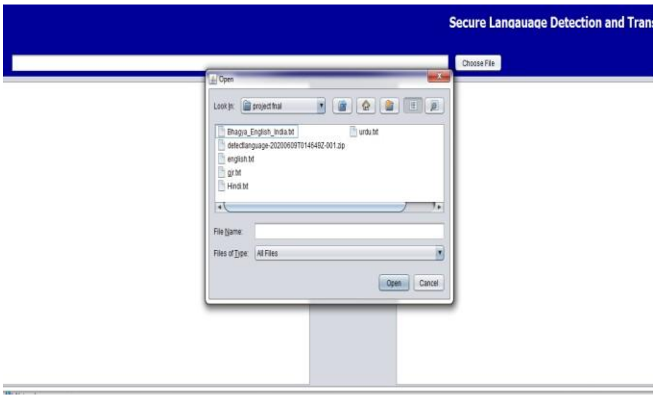
7: If translationlanguage=yes then goto step 9 8: Otherwise, goto step 1

9: Translation language API 9: Decrypt

10: Display the Result

RESULT ANDDISCUSSION

It shows the results of the outcome analysis utilising Java and the eclipse-implemented methods. First, the results of the orientation categorization assignment are reported. We submitted the system's output to a shared task workshop that offered test data. A single run of each approach was provided for four different languages, as well as the native language classification results for all of the languages. We need to figure out how to detect and translate languages in a secure environment.

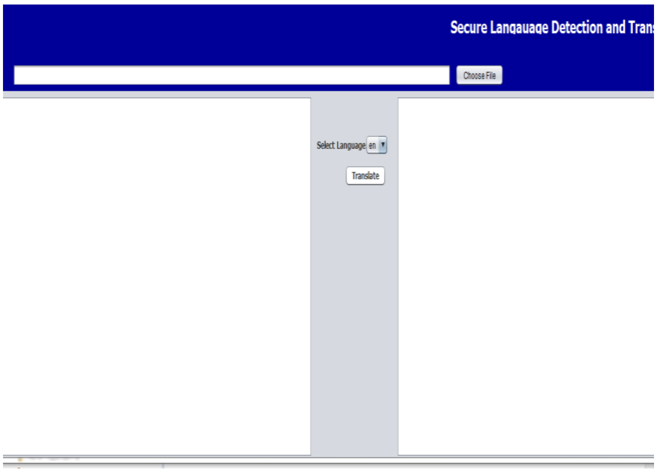


Look In: project final

Files of Type: All Files

Fig 3: Browse texted file

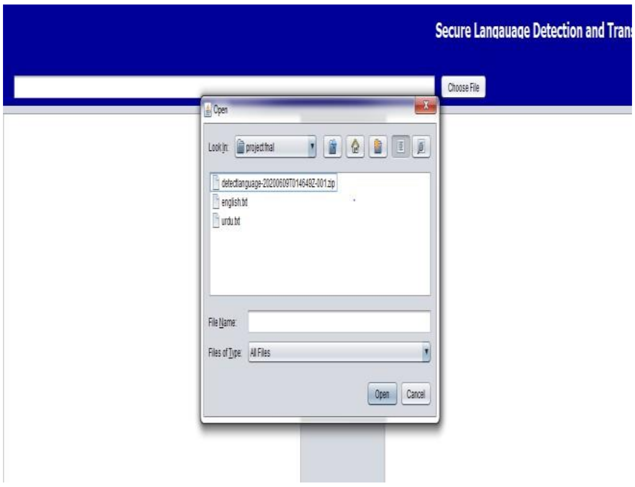
The browsing textual file is shown in fig.3 above. Because the Language Detector Model is employed, no training is necessary.



Select Language: en

Fig 4: Selection of Language for Translation

The selection of a language for translation is depicted in Fig.4. The basic goal is to translate supplied sentences into another language. Both the input and output are sentences in this case.



Look In: projectfinal

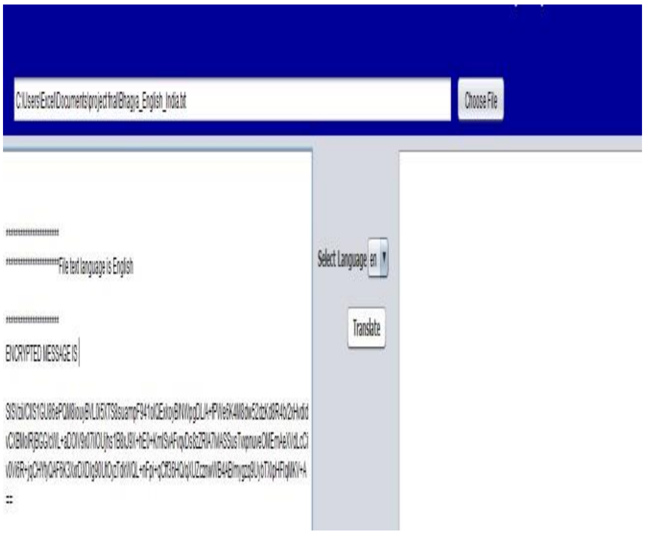
detectlanguage-20200609T014649Z-001.zip
english.txt
urdu.txt

File Name:

Files of Type: All Files

Fig 5: Browse File for Translation

The fig.5 above shows the File for Translation. It is the task to convert one natural language automatically into another, maintain the essence of the input text, and produce fluent text in the output language.



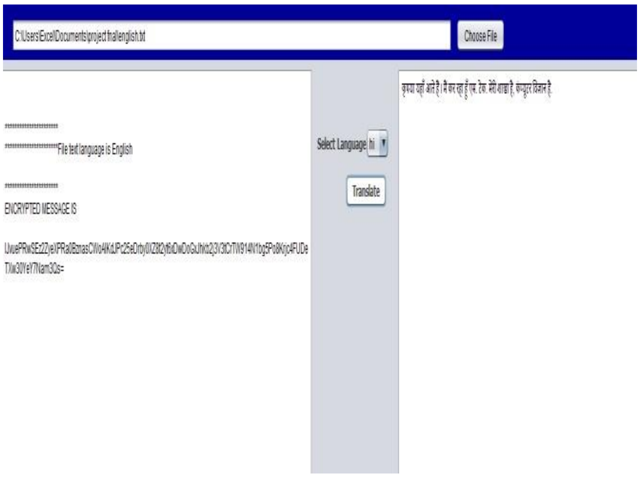
C:\Users\Excel\Documents\project final\Bhagya_English_India.txt

```
###################
###################File text language is English

###################
ENCRYPTED MESSAGE IS

SISVzbICNS1GU86ePQM8iouyBVL0i5XTS9suampF941nICEtxoyBNWpgDLI4+fPWe6K4M8dwE2dzKd8R4b2rHvdId
vCXBMoIRjBGGIdWL+aDON9x07IOUjhs1B8uJ9X+hE0+KmlSvAFvqvDs8zZRIA7hMSSusTwpnuveOMEm4eXWdLcCi
vWi6R+jqCHYbQAFBK3XxrDXDIg90UtOyzTdkWQL+nFpi+qCff36HQ/qXUZczmwWB4ABmygzq9UybTXlpHFiqMKV+A
==
```

Select Language: en

Fig 6: Encryption of Original Text of File

The results have given in the above fig.6 based on the AES technique. Itis used to encrypt the original text of the selected file.



C:\Users\Excel\Documents\project final\english.txt

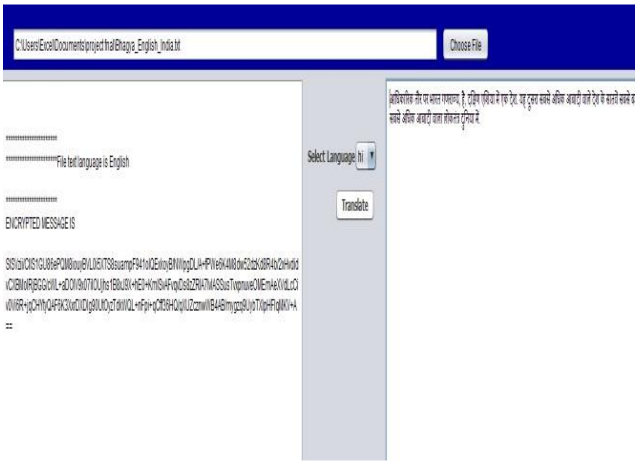
```
###################
###################File text language is English

###################
ENCRYPTED MESSAGE IS

UvuePRwSEz2ZyeXPRa0BznasCWo4lKdJPc25eDrby0IZ8t2yt6dDwDoGuJhkb2j3V3tCrTW914N1bg5Po8Krjc4FUDe
TXw30YeY7Nam3Qs=
```

Select Language: hi

कृपया यहाँ आते है। मैं कर रहा हूं एम. टेक. मेरी शाखा है, कंप्यूटर विज्ञान है.

Fig 7 Translation the Encrypted Text for Single Sentence

The results given in the above fig.7have based on the translation methodology, which provides secure detection and translation environment

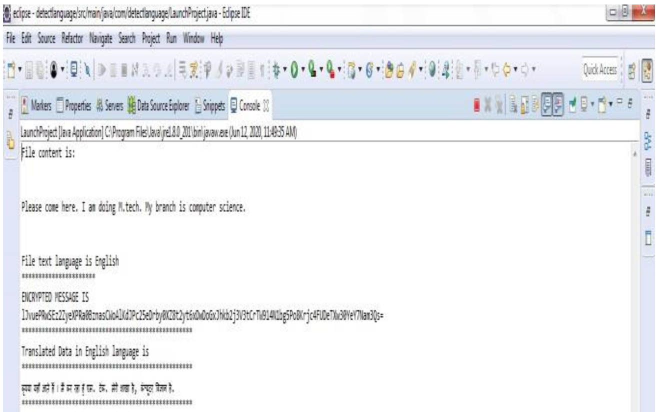

C:\Users\Excel\Documents\project final\Bhagya_English_India.txt
Choose File
Select Language hi
Translate

```
******************
******************File text language is English

******************
ENCRYPTED MESSAGE IS

SlSVzaiCXlS1GU86ePQM8iouyBVL0i5XTS8suampF941olQExkoyBNWpgDLiA+fPWe6K4M8dw52dzKd8R4b2tHvdid
vCXBMoIRjBGGIcWL+aDOW9x07lCUjhs1B8uJ9X+hE0+KmlSi4FvqxDs8zZRIA7hMASSusTwpnuveOMEmAeXVdLcCi
vW6R+jqCHYtyQ4F6K3XwDXDIg90UtOyzTdkWQL+nFpi+qCff36HQlqXUZczmwWB44BmygzqzJ9UybTXlpHFlqMKV+A
==
```

आधिकारिक तौर पर भारत गणराज्य, है, दक्षिण एशिया में एक देश. यह दूसरा सबसे अधिक आबादी वाले देश के सातवें सबसे ब
सबसे अधिक आबादी वाला लोकतंत्र दुनिया में.

Fig: 8 Translation the Encrypted Text for Multiple Sentence

The results given in the above fig.8 have based on the translation methodology, which provides secure detection and translation environment.The work which is discussed in this research is verified and validated through various testprocesses [18] [19] and test techniques [20] [21].


eclipse - detectlanguage/src/main/java/com/detectlanguage/LaunchProject.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Console

```
LaunchProject [Java Application] C:\Program Files\Java\jre1.8.0_201\bin\javaw.exe (Jun 12, 2020, 11:49:35 AM)
File content is:


Please come here. I am doing M.tech. My branch is computer science.


File text language is English
**********************
ENCRYPTED MESSAGE IS
lJvuePRw5Ez2ZyeXPRa0BznasCWo4lKdJPc25eDrby0XZ8t2yt6xDwDoGxJhkb2j3V3tCrTW914N1bg5Po8Krjc4FUDeTXw30YeY7Nam3Qs=
**************************************************
Translated Data in English language is
**************************************************
कृपया यहाँ आते हैं। मैं कर रहा हूं एम. टेक. मेरी शाखा है, कंप्यूटर विज्ञान है.
**************************************************
```

Fig 9 Display the Results (Console Screen)

CONCLUSION

For applications such as the fast growth of less-known languages on Internet have generated requirements for language recognition through machine translation, spell checking, multi-language information retrieval, etc. Three factors complicate this role: different sizes of character sets utilized to encode further languages, use of various character sets for single language, as well as more than one language sharing the same script. In this paper, particularly for English, Hindi, Urdu, and Gujrati, we present modules to identify and translate English into Indian languages or Indian Interlingual Languages. A language detection algorithm based on a statistical detector based on char n- gram and Yandex API is used for translation. There should be encryption and decryption when translating, so we use the AES algorithm to obtain a protected language detection and translation environment.One of the essential fields of natural language processing is languagetranslation.

FUTURE SCOPE

We will also use this language identifier module for translation in the future. For bilingual computer translation from English into the Urdu / Hindi language, this will be very helpful.

One of the fundamental difficulties is that English has the structure of the Subject Verb Object (SVO), whereas Urdu has the design of the Subject Object Verb (SOV) in Machine Translation.This study's proposal is a contribution and breakthrough.Our future work also lies in improving the NLI system's performance by considering features, which can classify native languages in a better way, and Language recognition task n-grams may be used for multi-lingual sentiment analysis for fine-grainedclassification.

References

1. R. Prabowo and M. Thelwall." Sentiment analysis: A combined approach." Journal of Informetrics, 3(2): 143-157, b2009.
2. J. Bollen, H. Mao, and X. Zeng "Twitter mood predicts the stock market." Journal of Computational Science, 2(1): 1-8, 2011.
3. Y. Kano, M. Miwa, K. B. Cohen, L. E. Hunter, S. Ananiadou and J. Tsujii, "U-Compare: A modular NLP workflow construction and evaluation system," in IBM Journal of Research and Development, vol. 55, no. 3, pp. 11:1-11:10, May-June 2011, doi: 10.1147/JRD.2011.2105691.
4. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research [review article],” IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48–57, May 2014.
5. N. Nayak and D. P. Mohapatra, “Automatic test data generation for data flow testing using particle swarm optimization,” in Proc. Int. Conf. Contemporary Comput., 2010, pp. 1–12.
6. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. Mcclosky, “The Stanford Corenlp natural language processing toolkit,” in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations, 2014, pp. 55–60.
7. Z. Liu, C. Lu, H. Huang, S. Lyu and Z. Tao, "Hierarchical Multi-Granularity Attention- Based Hybrid Neural Network for Text Classification," in IEEE Access, vol. 8, pp. 149362-149371, 2020, doi: 10.1109/ACCESS.2020.3016727.

8. Z. Hu, J. Luo, C. Zhang and W. Li, "A Natural Language Process-Based Framework for Automatic Association Word Extraction," in IEEE Access, vol. 8, pp. 1986-1997, 2020, doi: 10.1109/ACCESS.2019.2962154.

9. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, arXiv:1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473

10. Chen, F. Tang, P. Tino, A. G. Cohn, and X. Yao, ''Model metric colearning for time series classification,'' in Proc. 24th Int. Joint Conf. Artif. Intell., 2015, pp. 3387–3394

11. R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, ''Semi-supervised recursive autoencoders for predicting sentiment distributions,'' in Proc. Conf. Empirical Methods Natural Lang. Process. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 151–161.

12. T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' 2013, arXiv:1301.3781. [Online]. Available: https://arxiv.org/abs/1301.3781

13. J. Pennington, R. Socher, and C. Manning, ''Glove: Global vectors for word representation,'' in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1532–1543.

14. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''Bert: Pre-training of deep bidirectional transformers for language understanding,'' in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2019.

15. Y. Bao, S. Chang, M. Yu, and R. Barzilay, ''Deriving machine attention from human rationales,'' in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 1903–1913.

16. Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah, and A. Shah, "Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis," 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bangkok, Thailand, 2018, pp. 1-4, doi: 10.1109/ICETAS.2018.8629198.

17. M. Piyaneeranart and M. Ketcham, "Modeling for Robot-Driven Prototype Automation," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-6, doi: 10.1109/iSAI- NLP.2018.8692828.

18. Naresh, E., B. P. Vijaya kumar., Niranjanamurthy, M., Nigam, B. (2019). Challenges and issues in test process management. Journal of Computational and Theoretical Nanoscience, 16(9), 3744–3747.

19. Naresh, E., Vijaya Kumar, B. P., Naik, M. D. (2019). Survey on test generation using machine learning technique. International Journal of Recent Technology and Engineering, 7(6), 562–566.

20. Rayudu, Dadi Mohankrishna, Naresh. E and Vijaya Kumar B. P, “The Impact of Test-Driven Development on Software Defects and Cost: A Comparative Case Study”, International Journal of Computer Engineering and Technology (IJCET) 5, no. 2 (2014).

21. M. Nayak, N. E, S. P. Shankar and A. B. J, Cognitive Computing in Software evaluation,&quot; 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 509-514, doi:

22. 10.1109/DASA51403.2020.9317134.https://ieeexplore.ieee.org/document/9317134.