AUTO SYNONYM SUGGESTION BASED ON WHATSAPP CHAT

Dande Aryan Srivatsava

Department of Electronics and Computer Engineering,
Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad

Abstract

Social Media applications have been one of the salient innovative ideas that have been implemented and executed in a large scale. Applications like "Whatsapp" has been one of the biggest fortune companies across the globe that has been continuously emerging day by day. These applications are now serving as platforms which help people communicate and interact with the people they like to interact with..The purpose of communication could be of different reasons. And for the users of these applications whose mother tongue is not English, who communicate with other users for professional purposes in internationally generalized language, which is English, finds it hard to articulate the context they are supposed to explain or convey. Observing this issue faced by a lot of people, an idea was developed to partially help the users and make them comfortable while having a conversation virtually on their keypad or keyboard using this algorithm that helps the user explore new words in the dictionary and improve his vocabulary skills by getting suggestions by the software.

Keywords: *English.,user,communicate,interact.*

Date of Submission: 09-07-2022

I. INTRODUCTION

Field intro.

1.Natural Language Processing is the field in artificial intelligence which focuses on interpreting the human language with respect to the designed process or model. In this field many innovative algorithms have been designed to make use of human language inputs in as many applications as one can. Semantic analysis on reviews, Parts of speech tagging, Naïve bayes classification using probabilistic models and recommendations with collaborative filtering based on reviews, Auto correction, Named entity recognition, Machine translation and many more.

2.As part of this research one has to observe the problems using the ability of cognitive thinking and build a solution, draft the algorithm, implement and execute with deployment finally. In the current scenario, every part of the world relies on a communicative or an interactive platform such that the users can be connected in the best possible way. Using the connections, the users or the clients can contact their clients effortlessly. The impact of the communication platforms like whatsapp, facebook, twitter and instagram are monotonically increasing day by day.

Problem context.

3.It is known that whatsapp application is the most used platform for communication between two users or among a group of people. For people who work professionally always use these communication platforms to interact with their superiors, colleagues and fellow employees in their office. So it is mandatory for them to have good communication skills which should never let them feel embarrassed for misspelling a word or not articulating their intentions or not letting the other person know his weaknesses.

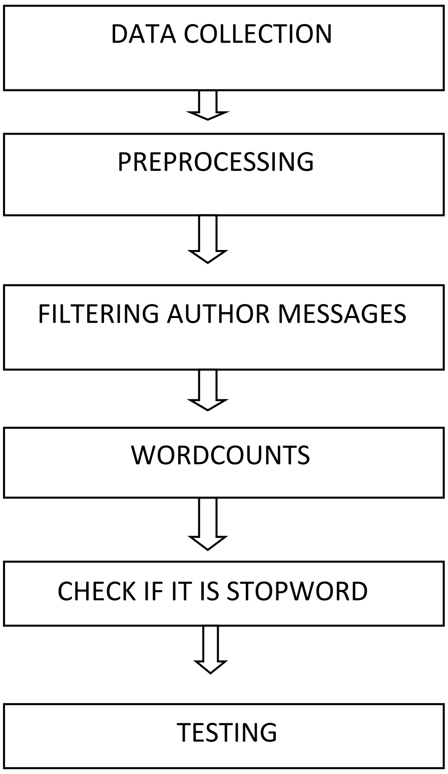
4. Communication/Social media interactive application software's are being employed to generate remote accessibility to applications and devices and inter change of documents and content or messages in text, AV formats among the devices or clients. These processes also contain document transferable programs, conversation through keypad/keyboard and instant fast messaging features.

5. Whatever may be the context whether it is detailing information, confessing an issue, explaining a topic and while having normal non-professional conversation the user should be able to use the words in the vocabulary in the best possible way. So all the users who are not good at vocabulary generally use the words again and again as they are not aware of the words in the dictionary with same meaning .So this algorithm is designed to suggest synonyms for the words that are repetitively used by the user many times.

II. PROPOSED METHOD

This method has mainly six stages. 1. DATA COLLECTION 2.PREPROCESSING 3.FILTERING THE MESSSAGES OF AUTHOR 4. WORDCOUNTS 5.CHECKIF IT IS STOPWORDS 6.TESTING

FLOWCHART



DATA COLLECTION

This Data is collected by extracting the conversation of two persons; their names are “Tanay kamath “and “Dheeraj kulkarni” in whatsapp. This is text file that was provided in open source. Also a particular user’s own whatsapp conversation is implicitly backed up and stored everyday in our mobile’s memory. Based on the particular mobile app settings, one can also periodically retrieve the chat/data which can also be media files such as audio clips or video clips, URL’s to cloud or Google drive based on the operating system of the device.

Steps to extract chat of a user:

1. Open the application(whatsapp)
2. Click/tap on the button – search bar
3. Type the name of the user who you want to extract the chat with.
4. Open the chat
5. Tap More options :
6. Click ”More”

7. Tap Export chat.
8. A text document will be mailed to the mail which is linked to the whatsapp account or to the mail which is in sync with the Google account. The final delivered .txt file will be in the format: ” XX/XX/XX, X:XX am/pm – “Message content” “

SAMPLE DATA ELEMENT:

{ 27/01/2020, 7:44 pm - Dheeraj Lalwani (TSEC, CS): Something or the other So that we can build our problem solving skills }

PREPROCESSING:

At this stage the whatsapp chat data is considered for preprocessing such that it should be ready to be used for the count phase. These levels of process starts with the process of data cleaning (cleaning), followed by the process of extracting the dates, times, authors and messages. Then The text file is loaded in the form of non-binary mode using the utf-8 encoding (Unicode Transformation Format).Function Getdatapoint - takes the input of a given whatsapp chat as a set of lines in the format as shown above. It splits the message into two parts using regex module with ‘-’as separator. Then using the split function Date, time, author and message are extracted separately and represented in a data-frame.

CLEANING THE UNNECESSARY DATA:

As the desired result is only based on the count of the words which have synonyms that can be suggested to the user if they are repetitively used, we don’t have to include the emojis and urls/website links that are used between the messages. Also media files such as audio and video files are not needed for consideration. Here we define and create functions that help find out the messages containing emojis, links and media files.

CHAT ANALYSIS :

As part of the overall work, here a analysis is made on all the unique features such as emojis count ,links count and media files count.

```
 Stats of Dheeraj Lalwani (TSEC, CS) -
Messages Sent 1732
 Average Words per message 6.596997690531178
Media Messages Sent 116
Emojis Sent 0
Links Sent 20
```

SELECTING THE AUTHOR :

As we want to analyze the messages and make suggestions for a particular user we have to filter out the messages of other user. Using author column, mentioning the name of a particular user we can extract the messages of only that particular user.

EXAMPLE:
[df["Author"]=="Tanay Kamath"]

TOKENIZING :

For all the messages, we have to tokenize the words for all the sentences (messages) in the list. Applying lambda function to all the sentences in the list, the resultant tokenized form of sentences will be produced as output.

```
df["Message"].apply(lambda s : s.split(' '))
```

WORDCOUNT : We design a function "word_count" which takes the whole list of tokenized sentences and outputs the dictionary of count.

EXAMPLE:

The word 'So' is used by the user (tanay kamath) for 111 times

After tokenizing the sentence, it is followed by calculating the word count which is also called as word frequency in terms of probabilistic models. While considering the word count, it is differentiated between types and *tokens*. Types are the words which occur only once, which are unique, on the other hand tokens are the repetitive words that occur frequently. For each word in each sentence we take the count of how many times each word has been used by the user.

STOPWORDS:

Using the library "wordcloud" we can extract the set of stop words. Stop words are the words which are generally not considered for suggestions as they are the words are quite often used in any kind of communication.

Stop word Examples : so,here,are,you….

SYNONYM SUGGESTION:

Here the threshold is assigned with a value of 20. This means if the count of the input is greater than 20 suggestions will be displayed as output. The library "nltk" is imported and then the "wordnet" corpus is downloaded from nltk library. So given the input word synonyms are suggested.

III. EVALUATION

As count for all the words is stored in the dictionary, we do have the list of stop words we can take the input and retrieves the count of that word and checks if the word is stop word or not and then determines the threshold value and also if it is greater than the threshold value and suggests synonyms for the word if it is not in the list of stopwords and if its count is greater than the threshold value.

SAMPLE OUTPUT :

```
Input word = "Also"
```

```
This word is used by the user for 34 times. These are the other alternate words
which could be used by the user instead.
{'also', 'besides', 'likewise', 'too', 'as_well'}
```

IV. CONCLUSION

This algorithm makes the client get to know new words for the appropriate word he gives as input while having a conversation and suggestions will be displayed based on his previous chat count . It determines the word frequency of each word that is being used by the client and suggests synonyms for the words that are repetitively used. Using this auto suggest feature in "ON" mode in any communication platform such as whatsapp, grammarly, linkedin or discord the user will be able to improve his vocabulary skills.

.

V. REFERENCES

[1]. Ravishankara K, Dhanush, Vaisakh, Srajan I S, “International Journal of Engineering Research & Technology (IJERT)”, ISSN: 2278-0181, Vol. 9 Issue 05, May-2020 [2]

[2]. https://www.analyticsvidhya.com/blog/2021/06/build-web-app-instantly-for-machine-learningusing-streamlit/

[3]. Lakshminarayanan, S. Prabhakaran, “Dogo Rangsang Research Journal”, UGC Care Group I Journal, Vol-10 Issue-07 No. 12 July 2020

[4]. J. Chen, H. Huang, S. Tian, and Y. Qu, “Expert Systems with Applications Feature selection for text classification with Naïve Bayes,” Expert Syst. Appl., vol. 36, no. 3, pp. 5432–5435, 2009

[5]. Pujadayanti, M. A. Fauzi, and Y. A. Sari, “Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Naïve Bayes dan N-gram,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. November, 2018.

[6]. Ernawati S Yulia E R Frieyadie and Samudi, 2019 Implementation of the Na¨ıve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies 2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018 Citsm p. 1–5.

[7]. 7.Sarkar K, 2018 Using Character N-gram Features and Multinomial Na¨ıve Bayes for Sentiment Polarity Detection in Bengali Tweets Proc. 5th Int. Conf. Emerg. Appl. Inf. Technol. EAIT 2018 p. 1–4.

[8]. Nguyen V H Nguyen H T Duong H N and Snasel V, 2016 n -Gram-Based Text Compression 2016.

[9]. Indrayuni E and Wahyudi m, 2015 penerapan character n-gram untuk sentiment analysis review hotel menggunakan algoritma naive bayes 8