Grenze International Journal of Engineering and Technology, Jul Issue

An Application for Detecting and Avoiding Redundant Messagesin Social Media

Dr. Shankar Gowda B N[1], Lavanya C[2] and Aishwarya C[3]
[1-3]Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore, India
Email: bnsgowda@gmail.com, lavanyanaik2002@gmail.com, aishwaryac21121999@gmail.com

***Abstract*—On social media we browse through enormous messages. Most of the messages tendto be redundant. Browsing through this redundancy is simply time consuming and attention seeking. The messages may consist of many day-to-day slangs and short forms that are ambiguous. The spectacular usability of the social media applications can be improved by overcoming redundancy. The standard form of slangs or short form words enhances the understanding ability of messages. This paper discusses an application that has been developed to avoid redundancy by standardizing the messages. The application deploys machine learning techniques to preprocess and analyze the messages. Henceforth, the redundant messages are restricted from appearing on the application user interface. By this the spectators get the unique and standard messages view.**

***Index Terms*— Redundant messages, Social media, Dictionary, Android.**

I. INTRODUCTION

Social media has become an important aspect in everyone's day-to-day life. Social media applications like Facebook, Twitter, WhatsApp, Telegram, Instagram, ShareApp, LinkedIn, Messenger, Snapchat etc., have gained a lot of attention. A lot of common messages are shared by many people on these platforms. Similarly media files like images, videos etc are also shared repeatedly [5]. Therefore a person active on social media receives enormous messages with varying redundancy [2]. These redundant messages don't weigh much importance. However, the redundancy in media consumes a lot of storage which is unnecessary [3].

Whatsapp is a commonly used social application to communicate. An exported file from Whatsapp is used as the input data [4] for the application. The messages in the input data are both unique and redundant. The redundant messages may have different forms but mean the same [1][8].

Standard meaning like 'hello' can be inferred from 'hie', 'hiii', 'heloo', 'hlo' etc., similarly 'How are you?' can be inferred from 'how are u?', 'hw r u?', 'how r u?'. 'hru?'etc. Though the spellings and the grammar mayvary they pose a single meaning [9][10]. This is the basic level of redundancy we come across commonly. We find many forms for a single phrase and slang consisting of fewer words or letters. Overcoming this redundancy is basic. The detected redundancy should be avoided effectively to convince the spectator [3][6].

The application deploys ML modules [1] topreprocess the data. The messages are thoroughly segregated unique and redundant from time to time. The messages appear only once based on time sequence. This convinces the spectator that no redundant messages have appeared. In group chats the redundant message appears once however the sender is not tagged with it. The application provides all message view for group chats so that the

Grenze ID: 01.GIJET.7.2.525

spectator can refer to it for his knowledge.

As discussed above standard meanings can be extracted from phrases, slangs and short forms. A standard dictionary is created to train the application. When a phrase or slang is detected the application asks the user to teach itthe meaning. Therefore the dictionary can be customized based on the user as to what he wants.

II. Literature Survey

The paper [1] describes the professional chat application where the messages are interpreted into intended meaning to find any redundant or inappropriate terms in the message that may include vulgar words, etc., using NLP. However it omits the commonly used slangs or short forms of words. The paper [2] describes the android application built to create separate sections for unique and redundant messages in group chat of Whatsapp. Though redundancy is detected it is not handled. The paper [3] presents a scheme to remove the duplicate files based on proposed image matching scheme which suits mobile environment using hash function and Huffman code to build unique code for each image. Scheme implements creating index file and removing daily images. The drawback is the minor changes in the image like size, cropping etc, is considered as a new image. The paper [4] aims to study and analyze theWhatsapp conversations. The analysisdiscovered that both students and professionals use Whatsapp to achieve their academic and business goals. In paper [5] the research was done to check the images that circulated the news on Whatsapp is redundant. It uses theautomatic detection tools using the machine learning techniques to detect the news in the images. The drawback is research checked all the redundant images, its space efficiency and time efficiency reduced. The paper [6] portrays the two stage sampling selection strategy and simple map reduce concept is proposed. Blocking, comparison and classification stage is used to analyze the data in these three stages, identify the duplicate data from the original data. The paper [7] proposes a 2-step approach to address this new ECPE task, which first performs individual emotion extraction and cause extraction via multi-task learning, and then conduct emotion-cause pairing and filtering. The emotion extraction and the cause extraction are independently modeled by two Bi-LSTMs. But the mistakes made in the first step will affect the results of the second step. The paper [8] explores the approach on a wide range of benchmarks for natural language understanding. They proposed a general task-agnostic model, which outperformed discriminatively trained models that use architectures specifically crafted for each task. The paper [9] analyzes more than 50 string matching methods in terms of strengths, weaknesses, and efficiency with respect to applications, which can helpful to identify suitable string matching algorithms. It is tedious choosing an appropriate string matching algorithm for current applications, integrating several algorithms, and modifying algorithms. The paper [10] reviews for unstructured data mining. The entire process of unstructured information parsing can bedivided three major steps which are relationship-based extraction, template-based extraction and deep learning-based extraction.

III. Problem Statement

The messages of social media like WhatsApp get accumulated with lot of redundant or duplicate messages. There is a great need to reduce the real time problem of redundant messages without missing out essential details.

IV. Proposed Method

Fig. 1 system architecture depicts the working model of the application. The input file is uploaded into the application. The application is divided into the following modules.

1. Message extraction: The imported file is read line by line. Each line is validated by checking for unusual contents like ‘This message was deleted’, ‘< Media omitted> ’, ‘ Vidya added +91 9876543219 ’, ‘Nandini removed +91 8765493219 ’, ‘Aishwarya left’ etc., such contents are removed to get just messages.
2. Dictionary search module: The messages are processed based on the criteria of their length. Only the short messages willchecked for the key-value in the dictionary. If the message is found then key meaning of that message will be retrieved. In the case of new slangs or phrases being detected, they are piled intoa list. Fuzzy logic is applied to compare the strings in the dictionary with accuracy more than 85%. Hence, the relevant key is retrieved from the dictionary. The processed messages along with its key andother attributes like word length, count,redundant, continued etc., is inserted into the firebase database.
3. Comparison module: The new messages are compared with previous messages to check redundancy. The fuzzy logic is applied to match the strings with the accuracy of 87% and above. If the keys ought to match

the messages are considered redundant. If the logic fails the message is claimed unique.

4. Unique message and redundant message module: The attributes of compared messages from the above module are updated based on uniqueness and redundancy.
5. Message display module: Based on the occurrences of the messages from time to time the unique messages are displayed to the user in black colour. The redundant messages are displayed only once in green color. The redundant messages that occurmore than once are tagged with count of the redundancy. The senders of suchmessages can also be viewed.

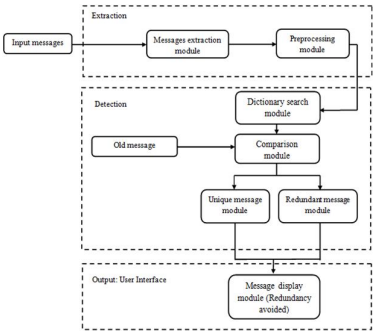

Figure 1: Architecture

V. ALGORITHMS

Preprocessing algorithm:-

Input: Valid lines from the file

Output: Extracted message class

Description: In this process date, time, author, message are extracted from the received lines. Computations are performed on the messages this to get word count, letter count and id. This is required to differentiate short messages and long messages for further process.

Algorithm:

Step 1: Start

Step 2: The date and time are extracted fromthe line.

Step 3: If line starts with date and time:

extract the message and the sender.

Else:

Considered as continuation of previous msg.

Step 4: computations are performed on the message and the previous message isupdated.

Step 5: Only the short messages are sent fordictionary check. The long messages are simply checked for redundancy.

Step 6: Stop

Dictionary check algorithm:-

Input: Message

Output: Key of the message

Description: The short message is comparedusing fuzzy technique with matching dictionary. The mapped keys of the messages are returned.

Algorithm:

Step 1: Start

Step 2: For short message get the validdictionary from the database.

Step 3: Search for the key in the dictionary using fuzzy sort token ratio.

Step 4: If the message is found in the dictionarywith an accuracy of 85% or above return thekey of the message.

Step 5: For long message simply the message itself is returned as key

Step 6: In case of failure of any of the above steps return false.

Step 7: Stop

Comparison algorithm:-

Input: Key of the message.

Output: Redundant/Unique message detected Description: The key of the message iscompared with the key of previous messageusing the fuzzy string matching logic withaccuracy of 87% and above. This process detects unique and redundant messages via theabove comparison.

Algorithm:

Step 1: Start

Step 2: Compare the new message key with previous message key by fuzzy set token ratio with accuracy of 87% and above

Step 3: If the keys match the message is considered redundant.

Step 4: If failed the message is considered unique.

Step 5: Stop

Unique Message algorithm:-

Input: Unique message Output: Unique message class

Description: The unique message class is updated. The parent of the unique message is updated to itself. The count of the message is set to 0 and the previous message class is updated to the key and the message itself.

Algorithm:

Step 1: Start

Step 2: Set the parent as the message itself.

Step 3: Set the count of the message to 0.

Step 4: Set the rest of the attributes likeredundant, continued etc to null

Step 5: Update the previous message class tothe current message class.

Step 6: Update the message class in thedictionary.

Step 7: Stop

Redundant message algorithm:-

Input: Redundant message Output: Redundant message class

Description: The redundant message class isupdated. The parent of the message is updated to the previous message. The count of the message is updated. The redundant factor is set and the previous message is retained.

Algorithm:

Step 1: Start

Step 2: Set the parent as the previous message.

Step 3: Set the count of the message based onits occurrence.

Step 4: Set the redundant factor

Step 5: Retain the previous message class

Step 6: Update the message class in thedictionary.

Step 7: Stop

VI. Result And Discussion

The results from the above experimentation are discussed in this section along with tables and graphs. The table 1 shows the time taken for dictionary check of different type of messages. Three types of messages are considered for dictionary check and comparison module. They are long messages and short messages. Short messages are further categorized into message found in the dictionary and not found in the dictionary. Long messages take negligible time for the process.

It happens so because long messages do not undergo dictionary check however they are compared for redundancy. Short messages undergo dictionary check and comparison module. The keys of short messages can be found at the beginning, at the end or anywhere in between in the dictionary. The time complexity of comparison module is based on this search. When the keys of short messages are not found in the dictionary the timecomplexity is the highest because the entire dictionary is searched. The Fig. 2 shows the graph of efficiency of dictionary searches and comparison module. It can be deciphered that processing of long messages is efficient than short messages. The execution time of shortmessages found in dictionary is less than that not found.

TABLE I. THE TABLE SHOWS THE TIME TAKEN FOR DICTIONARY CHECKAND COMPARISON MODULE OF DIFFERENTTYPE OF MESSAGES DICTIONARY

Phrases	Average time in ms
Peace out	13.4
Zero	12.4
Cannot talk	2
Gd morning	5.4
Not in dictionary	26.2
Atleast	17.8
Practical	15.4
Dictionary not found in	23.4
Long message	1
Long message	1
Long message	1
Long message	1

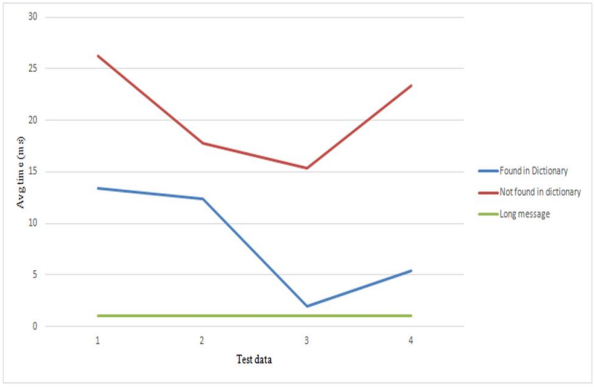

Figure 2: Graphical representation of time efficiencyof check and comparison module for different types of messages

The Fig. 3 shows graphical representation of the accuracy score of token sort ratio on test data and dataset. The token sort ratio function sorts the strings alphabetically and then joins them together, ratio is then computed. The peaks indicate highest possible accuracy for considered test data with dataset.

The Fig. 4 shows graphical representation of the accuracy score of token sort ratio on test data and dataset. The token set ratio function sorts the strings alphabetically and then joins them together; it takes out the common tokens before calculating the ratio between the strings. The peaks indicate highest possible accuracy for considered test data with dataset.

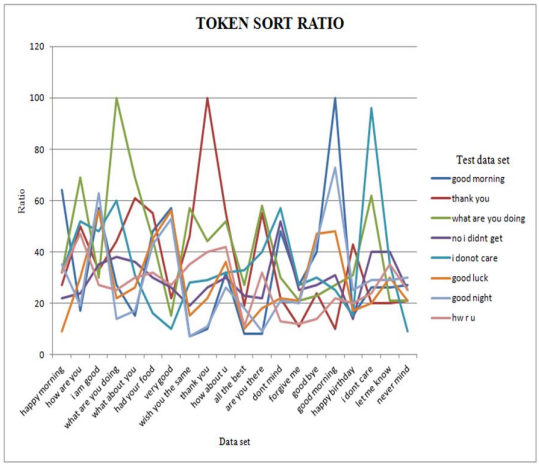

Figure 3: Graphical representation of the accuracyscore of token sort ratio on test data and dataset

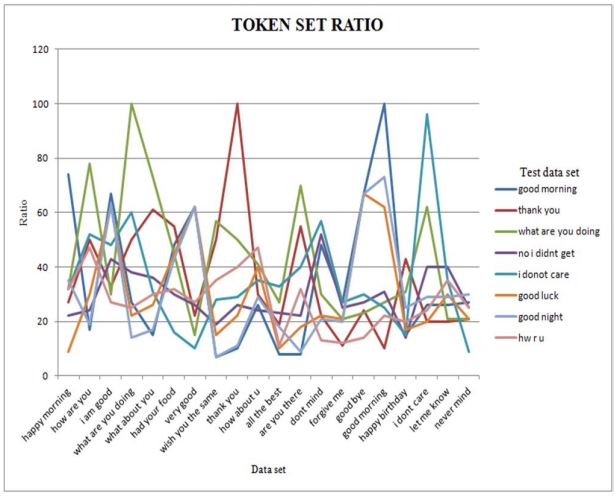

Figure 4: Graphical representation of the accuracy score of token set ratio on test data and dataset

VII. CONCLUSIONS

The problems and the treats that were discussed in the existing system and its drawback can be solved using this effective Machine Learning models. The problem of redundancy can be overcome by the application. Besides detecting, effective avoidance measures are also included. The integrated application can be used with other social media applications to detect and avoid redundancy. The application can evolve semantic analysis and emotions for better performance in the future. Hence using social media applications are easier and better.

ACKNOWLEDGMENT

This research was supported by Bangalore Institute of Technology. We thank the support of Department of Computer Science and Engineering that provided insight and expertisewhich greatly assisted the research. We acknowledge the information from the literature survey of the listed papers.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] Ammar Asaad, Ali Adil Yassin Alamri, “A New Scheme for Removing Duplicate Files from Smart Mobile Devices: ImagesasaCase Study, CUESJ, Volume-3 pp: (5- 13), 2019
[4] Sana Shahid, “Content Analysis of WhatsApp Conversations: An Analytical Study to Evaluate the Effectiveness of WhatsApp Application in Karachi”, IJMJMC, Volume-4, pp: (14-26), 2018.
[5] Julio C. S. Reis, PhilipeMelo, Kiran Garimella, Jussara M. Almeida, Dean Eckles,FabricioBenevenuto, “A Dataset of Fact- Checked Images shared on WhatsApp During the Brazilian and Indian Elections”, ICWSM, pp:(903-908),2020.
[6] Anbarasi. M, Karthika. S K,“An Approach to Reduce the Data Duplication using Simple Map Reduce Algorithm”, NCICCT- 2016, Volume 4 Issue 19.
[7] Rui Xia, Zixiang Ding, “Emotion-Cause Pair Extraction: A New Task To Emotion Analysis In Text”, arXiv [cs.CL],4 June 2019.
[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, “Improving Language Understanding By GenerativePre-Training”,[online],2019.
[9] Saqib Hakak, Amirrudin Kamsin, Palaiahnakote Shivakumara, “Exact StringMatching Algorithms”,, IEEE Access.
[10] Shubham Jain, Amy de Buitleir, and Enda Fallon, “A Review of Unstructured Data Analysis and Parsing Methods”, ConferencePaper, March 2019.