IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A REVIEW ON CONTENT BASED SMS CLASSIFICATION USING REGULAR EXPRESSION AND PATTERN MATCHING

Sumeet Bhangale
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Nikhil Kolhe
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Pravin Kakad
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Aniket Bandgar
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Vivek Waghmare
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Mangesh Ghonge
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Amol Potgantwar
Sandip Institute of Technology & Research Centre
Savitribai Phule Pune University
Nashik, India

Abstract:

Short Message Service (SMS) is a comprehensive mobile phone service for users to communicate with others, a faster and more convenient way of communication. However, it has some limitations such as: the inability to search and categorize SMS and there is room for improvement, thereby solving real-time text messaging problems. Our system provides basic text messaging functionality as well as various features such as categorizing messages based on personal, social, transactional, and custom categories with colour codes within the app, bill due date reminders, and highlighted messages.

Encryption is paramount when sensitive data is transmitted over the network. Various encryption algorithms such as AES, DES, RC4, and others are available for this. The most widely used algorithm is the AES algorithm. We have developed an application on the Android platform that allows the user to encrypt messages before they are transmitted over the network. We used the Advanced Encryption Standards algorithm for data encryption and decryption. This app can run on any device with the Android platform. This application provides safe, fast, and strong encryption of data. A large amount of data confusion and proliferation occurs during encryption, making it very difficult for an attacker to interpret the encryption pattern and the plaintext form of the encrypted data.

Keywords*: *SMS (Short Message Service), Regular Expressions, Pattern Matching, Tokenization, Categorization, End-to-End Encryption, AES (Advanced Encryption Standard).

Introduction:

Proposed Text Messenger which solves real time problems of text messaging. Our system provides core functionalities of text messaging and beside to that various facilities like categorization of messages based on personal, social, transactional and user defined categories with colour codes. As the number of SMS users increases, short messaging service (SMS) becomes more popular and more widely used for personal messaging and authentication (mobile banking). SMS messages are among the fastest and most convenient ways to communicate promotions and advertisements to users. [1] [2]

To develop a system to classify and categories different types of messages. That will save users time for identifying important and non-important messages. Where user can keep an eye on his personal/ banking transactions. This application also automatically store and remind to user about his due statuses for different things Ex. bills, So that user can never miss his dues. [2] [3]

The algorithm used for End-to-End encrypted transmission is Advanced Encryption Standards algorithm. [4] This application is developed on Android platform. The later part of the paper explains the working of SMS, the Categorization of SMS and the AES algorithm and the working of our developed application. [4]

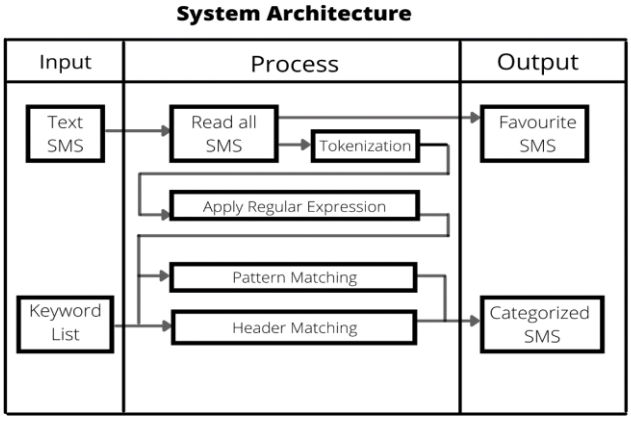
SMS - Short Message Service:

SMS is an Abbreviation of Short Message Service, it's a technique of communication that sends text between cell phones, or from a laptop or hand-held to a cell phone. The "short" term refers to the utmost size of the text messages: a hundred and sixty characters (letters, numbers or symbols within the Latin alphabet). For different alphabets, like Chinese, the maximum SMS size is seventy characters. [4]

Working of - Short Message Service:

Whenever you send an instant message, it initially goes to a close by cell overshadow a pathway called the control channel, and afterward into a SMS community (SMSC). The SMSC resends that message to the pinnacle nearest to the beneficiary, and afterward it goes to their telephone. SMS likewise sends information related with the message, including the length of the message, design, time stamp, and objective. [4]

Methodology:



There are some methodologies for Categorization of text SMS, such as Rule based systems, Machine Learning based system, Hybrid systems. In our proposed system we are using Regular Expression and Pattern Matching for categorization of SMS and we are also using AES algorithm for encrypting and decrypting SMS. The entire Process can be divided into several modules such as capturing SMS, Pre-processing, Classification and Encrypting SMS. The basic illustration of modules is to classify SMS into different categories is shown in the figure followed:

Fig. System Architecture

1. Tokenization:

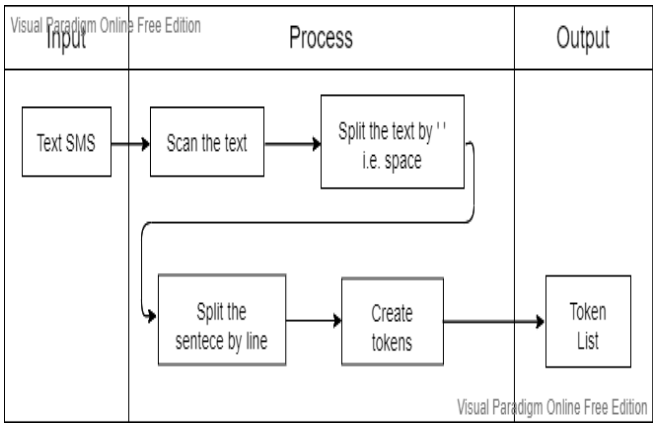


Fig. Tokenization

Tokenization is the method of exchanging sensitive information for non-sensitive data referred to as "tokens" which will be utilized in a piece of information or internal system while not conveyance it into scope. though the tokens are unrelated values, they preserve bound parts of the first data—commonly length or format—so they'll be used for uninterrupted business operations. the first sensitive data is then safely kept outside of the organization's internal systems. not unlike encrypted data, tokenized data is unreadable and irreversible. This distinction is especially important: as a result of there being no mathematical relationship between the token and its original number, tokens cannot become to their original type while not the presence of additional, one-by-one keep data. As a result, a breach of tokenized surroundings won't compromise the first sensitive data. As represented previously, a token may be a piece of information that stands sure another, a lot of valuable pieces of information. Tokens have just about no worth on their own they are solely helpful as a result they represent one thing bigger, equivalent to a Mastercard primary account variety (PAN) or Social Security number (SSN). a decent analogy is a poker chip. rather than filling a table with money (which will be simply lost or stolen), players use chips as placeholders. However, the chips will be used as money, although they're stolen. they have to 1st be changed for their representative value. Tokenization works by removing the dear knowledge from your surroundings and exchanging it with these tokens. Most businesses hold a minimum of some sensitive data at intervals in their systems, whether or not or

not it's Mastercard data, medical information, Social Security numbers, or anything that needs security and protection. exploitation tokenization, organizations can still use this data for business functions while not acquiring the chance or compliance scope of storing sensitive data internally. [1]

2. Regular Expressions:

A regular expression is one character or a lot of difficult patterns. Regular expressions could be used to perform all kinds of text operations such as text search and text replace operations. [5]

Java does not have any built-in class for Regular Expression however, we are able to import the java.util.regex package to perform operations on regular expressions.

The package includes the subsequent classes:
Pattern class - Defines a pattern (to be utilized in a search)

Matcher class – won't to seek for the pattern.

PatternSyntaxException class - Indicates software error in an exceedingly regular expression pattern performance.

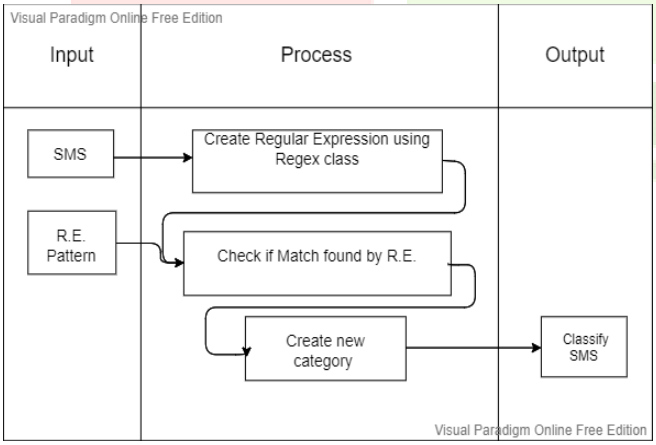


Fig. Regular expressions

we apply regular expressions as followed:

i. Regular Expressions Patterns:

The first parameter of the Pattern.compile() method is the pattern. It describes what is being searched for.

Brackets are used to find a range of characters:

Expression	Description
[abc]	Used to find single character from the contents between the brackets.
[^abc]	Used to find single character that is NOT available between the brackets
[0-9]	Used to find digit from the range 0 to 9 between the brackets.

ii. Meta-Characters:

Metacharacters are characters with a special meaning:

Metacharacter	Description
|	Look for a match for any of the patterns separated by |, such as cat|dog|fish.
.	Look for only one instance of any character.
^	Look for a match as the first character of a string, as in: ^Hello
$	Look for a match at the end of the string, as in: World$
\d	Look for a digit
\s	Look for a whitespace character
\b	Look for a match at the start of a word like \bWORD or at the end of a word like WORD\b.
\uxxxx	Look for the Unicode character indicated by the hexadecimal number xxxx.

3. Pattern Matching:

In software engineering, design matching is the demonstration of actually looking at a given succession of tokens for the presence of the constituents of some examples. Rather than design acknowledgment, the match for the most part must be accurate: "it is possible that it will or won't be a match." The examples by and large have the type of either successions or tree structures. Utilizations of example matching incorporate yielding the areas (if any) of an example inside a symbolic

arrangement, to yield a few parts of the matched example, and to substitute the coordinating example with another symbolic succession (i.e., search and supplant).
Grouping designs (e.g., a text string) are frequently portrayed utilizing normal articulations and matched utilizing methods, for example, backtracking.

4. SMS - Categorization:
Categorization can be done as per the following:

i. Company.
ii. Transactional.
iii. Due-Dates.
iv. OTP.
v. Favorites.

5. End to End Encryption: In our proposed system, we used the AES algorithm to encrypt and decrypt SMS by both sender and receiver side.

6. AES Algorithm:

The AES Encryption algorithm (also known as the Rijndael algorithm) is a symmetric block cipher algorithm with a block/chunk size of 128 bits. It converts these individual blocks using keys of 128, 192, and 256 bits. Once it encrypts these blocks, it joins them together to form the ciphertext. [4]

It is based on a substitution-permutation network, also known as an SP network. It consists of a series of linked operations, including replacing inputs with specific outputs (substitutions) and others involving bit shuffling (permutations). [3] [4]

A. SubBytes Step:

This step is equal as SubBytes step of AES algorithm. In the S-Box Substitution step, every byte withinside the matrix is reorganized the usage of an 8-bit substitution container. This substitution container is called the Rijndael S-container. This operation offers the non-linearity withinside the cipher. The S-container used is derived from the multiplicative inverse over GF (28), recognised to have appropriate nonlinearity properties. To keep away from assaults primarily based totally on simple algebraic properties, the S-container is built with the aid of using combining the inverse feature with an invertible affine transformation. The S-container is likewise selected to keep away from any constant points (and so is a derangement), and additionally any contrary constant points. [7] This step reasons confusion of statistics withinside the matrix. S-Box Substitution is executed one by one for LPT and RPT. This is step one of iterative spherical transformation. The output of this spherical is given to the subsequent spherical. [3]

B. ShiftRows Step:

The ShiftRows step is performed on the rows of the state matrix. It cyclically shifts the bytes in each row by a certain offset. The first row remains unchanged. Each byte of the second row is shifted one position to the left. Similarly, the third and fourth rows are shifted by two positions and three positions respectively. The shifting pattern for block of size 128 bits and 192 bits is the same.[3]

C. MixColumns Step:

In the MixColumns step, the **4** bytes of every column of the nation matrix are mixed the use of an invertible linear transformation [5]. A randomly generated polynomial is organized in a four*four matrix. The equal polynomial is used during decryption. Each column of the nation matrix is XOR-ed with the corresponding column of the polynomial matrix. The end result is up to date withinside the equal column. The output matrix is the enter to AddRoundKey.[3]

D. AddRound Key:

A spherical key's generated through appearing diverse operations on the cipher key. This spherical key's XOR-ed with every byte of the nation matrix. For each spherical a brand-new spherical key's generated using Rijndael‟s key scheduling algorithm. [3]

Decryption of the Proposed Algorithm:

The encryption set of rules is known as the cipher and the decryption set of rules because the inverse cipher. In addition, the cipher and the inverse cipher operations need to be accomplish**ed** in the sort of manner that they cancel every other. The rounds keys need to additionally be utilized in opposite order. [4] The Cipher Text which is fashioned of 256-bit 4*eight Matrix is the enter for the decryption process. [3]

Implementation and Snapshots of the Application:
SMS - Categorization:
Categorization can be done as per the following:
i. Company:
To categorize SMS in the Company tab we used the following Regular expressions Patterns.
"[A-Z][A-Z] – [A-Z][A-Z][A-Z][A-Z]"

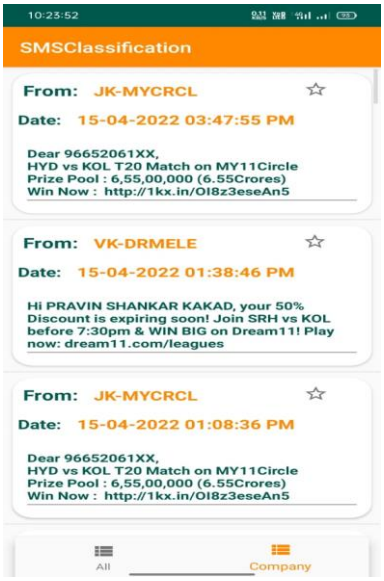


Fig. Snapshot Company Category

ii. Transactional:

To categorize SMS in the Transactional tab we provide some pre-defined keyword lists to the system for example- “Rs”, “Credit”, “Debit” etc. with both upper and lower cases and post that categorize them in the Transactional tab with Green color for Credit Amount and red for Debit.

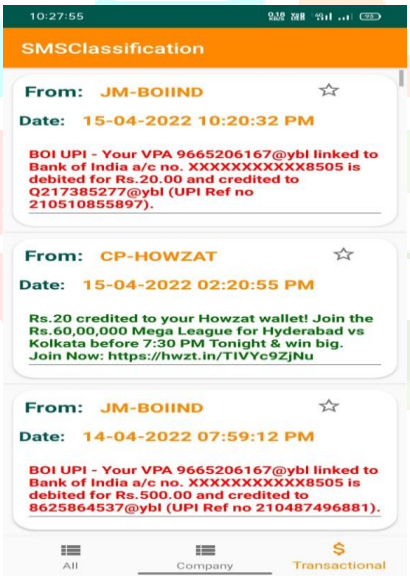


Fig. Snapshot Company Category

iii. Due Dates/Reminders:

To categorize SMS in the Due Dates tab we used the following Pattern and Regular Expressions:

Pattern:

```
"^(?:(?:31(\\/|-|\\.)(?:0?[13578]|1[02]|(?:Jan|Mar|May|Jul|Aug|Oct|Dec)))\
\1|(?:(?:29|30)(\\/|-|\\.)(?:0?[1,3-9]|1[0-
2]|(?:Jan|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec))\\2))(?:(?:1[6-9]|[2-
9]\\d)?\\d{2})$|^(?:29(\\/|-|\\.)(?:0?2|(?:Feb))\\3(?:(?:(?:1[6-9]|[2-
9]\\d)?(?:0[48]|[2468][048]|[13579][26])|(?:(?:16|[2468][048]|[3579][26])
00)))$|^(?:0?[1-9]|1\\d|2[0-8])(\\/|-|\\.)(?:(?:0?[1-
9]|(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep))|(?:1[0-
2]|(?:Oct|Nov|Dec)))\\4(?:(?:1[6-9]|[2-9]\\d)?\\d{2})$"
```

Regular Expressions:

"\\d\\d\\d\\d-\\d\\d-\\d\\d"
"\\d\\d\\d\\d/\\d\\d/\\d\\d"
"\\d\\d/\\d\\d/\\d\\d\\d\\d"
"\\d\\d-\\d\\d-\\d\\d\\d\\d"

Fig. Snapshot Due-Dates Category

4. OTP:

To categorize SMS in OTP Tab we did the same process as a transactional category with a different keyword list.

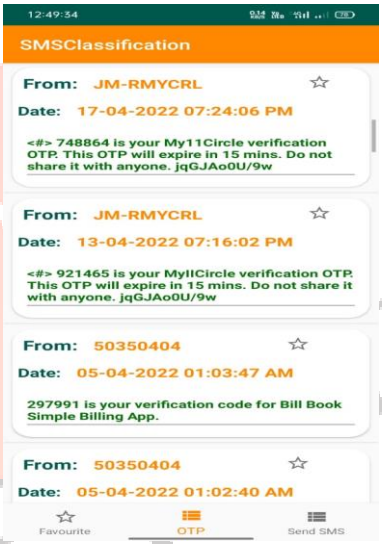


Fig. Snapshot OTP Category

Conclusion:

This application can perform the core tasks of the text messaging application like Sending and Receiving text messages. Along with this our application is solving the real time problems occurred while messaging text messages and classification it into different tabs.

This applications system provides a unique solution of categorizing the text messages based on various pre-defined categories gives user flexible option to organize SMS.

This system does not need any Internet connection to work effectively.

References:

[1] S. Ballı and O. Karasoy, "Development of content-based SMS Classification Application by using Word2Vec feature based extraction," *IET Software*, vol. 13, p. 10, 2018.

[2] H. Padhiya and P. P. Rekh, "Improving Accuracy of Text Classification for SMS Data," *International Journal for Scientific Research & Development*, vol. 1, no. 10, p. 4.

[3] N. Saxena and N. S. Chaudhari, "EasySMS: A Protocol for End-to-End Secure Transmission of SMS," *IEEE*, vol. 9, p. 12, 2014.

[4] R. Rayarikar, S. Upadhyay and P. Pimpale, "SMS Encryption using AES Algorithm on Android," *International Journal of Computer Applications*, vol. 50, p. 06, 2012.

[5] M. CUI, R. BAI, Z. LU, X. LI, U. AICKELIN and P. GE, "Regular Expression Based Medical Text," *IEEE Accsess*, vol. 07, 2019.

[6] P. Pimpale, R. Rayarikar and Sanket , "Modifications to AES Algorithm for Complex Encryptions," *International Journal of Computer Science and Network Security*, vol. 11, 2011.

[7] O. B. S. Karasoy, "Developing mobile application for content base spam SMS filtering and comparison of classification algorithms," *Int. Artificial Intelligence and Data Processing Symp*, 2016.

[8] Castiglione, A., De Prisco, R., De Santis, A.: 'Do you trust your phone?'. ECommerce and Web Technologies, Linz, Austria, September 2009, pp. 50–61

[9] Ho, T., Kang, H., Kim, S.: 'Graph-based KNN algorithm for spam SMS detection', J. Univers. Comput. Sci., 2013, 19, (16), pp. 2404–2419

[10] Church, K., Oliveira, R.D.: 'What's up with Whatsapp?: comparing mobile instant messaging behaviors with traditional SMS'. 15th Int. Conf. HumanComputer Interaction with Mobile Devices and Services, Mobile HCI, Munich, Germany, 2013

[11] Delany, S.J., Buckley, M., Greene, D.: 'SMS spam filtering: methods and data', Expert Syst. Appl., 2012, 39, (10), pp. 9899–9908

[12] Priyanka Pimpale, Rohan Rayarikar and Sanket Upadhyay, "Modifications to AES Algorithm for Complex Encryption", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.10, October 2011.

[13] J.Daemen and V.Rijmen, AES Proposal: Rijndael, NIST"s AES home page, http ://www:nist:gov/aes."Announcing the Advanced Encryption Standard (AES)",Federal Information Processing Standards Publication 197, November 2001

[14] Delany, S.J., Buckley, M., Greene, D.: 'SMS spam filtering: methods and data', Expert Syst. Appl., 2012, 39, (10), pp. 9899–9908

[15] J.Daemen and V.Rijmen, AES Proposal: Rijndael, NIST"s AES home page, http ://www:nist:gov/aes."Announcing the Advanced Encryption Standard (AES)",Federal Information Processing Standards Publication 197, November 2001

[16] R. E. Anderson et al., "Experiences with a transportation information system that uses only GPS and SMS," in Proc. IEEE ICTD, no. 4, Dec. 2010.

[17] K. Yadav, "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," in Proc. Workshop Hotmobile, 2011, pp. 1–6

[18] M. Densmore, "Experiences with bulk SMS for health financing in Uganda," in Proc. ACM CHI, 2012, pp. 383–398.

[19] J. Hellström and A. Karefelt, "Participation through mobile phones: A study of SMS use during the Ugandan general elections 2011," in Proc. ICTD, 2012, pp. 249–258

[20] K. Park, G. I. Ma, J. H. Yi, Y. Cho, S. Cho, and S. Park, "Smartphone remote lock and wipe system with integrity checking of SMS notification," in Proc. IEEE ICCE, Jan. 2011, pp. 263–264.

[21] Xinmiao Zhang and Keshab K. Parhi, "Implementation Approaches for the Advanced Encryption Standard Algorithm", 1531-636X/12, IEEE 2002.

[22] Yun Yang, Y. W. (2010). The Improved Features Selection for Text Classification. 2nd International Conference on Computer Engineering and Technology. IEEE

[23] Mita K. Dalal, M. A. (2011). Automatic Text Classification: A Technical Review. International Journal of Computer Applications

[24] Mehar, H.S. 2013. SMS Spam Detection using Machine Learning Approach.. International Journal of Information Security Science 2