Important Information Extraction from Short-Text Messages

Abhay Parihar
Department of Information Technology
Indian Institute of Information Technology, Allahabad
Prayagraj, Uttar Pradesh, India
iit2020074@iiita.ac.in

***Abstract*—This literature review explores the topic of information extraction from short text messages, which has become increasingly important due to the growing use of social media and mobile messaging platforms. The review begins by defining short text messages and discussing the challenges associated with extracting information from them. It then examines various approaches that have been used to extract information from short text messages, including natural language processing techniques. The review also discusses the different types of information that can be extracted from short text messages, such as pattern matching (RegEx), named entity recognition, keyword extraction and topic modeling. Finally, the review summarizes the current state of the field, identifies areas for future research, and discusses the potential applications of information extraction from short text messages in various domains, including healthcare, marketing, and security.**

***Index Terms*—Information Extraction, Short Text Messages, Natural Language Processing, Named Entity Recognition, Topic Modelling**

I. Introduction

The widespread use of social media and mobile messaging platforms has resulted in an exponential increase in the volume of short text messages being generated every day. These short text messages contain a lot of information, such as transactions, bill information, and topics of discussion, that can be leveraged for various applications. However, extracting this information from short text messages is not an easy task due to their inherent brevity and informal nature. Therefore, information extraction from short text messages has become an important research topic in the field of natural language processing.

This literature review delves into the various approaches that have been used to extract information from short text messages. The review begins by defining what is meant by short text messages and highlighting the challenges associated with extracting information from them. It then discusses the natural language processing techniques that have been employed to extract information from short text messages, including named entity recognition, keyword extraction, and topic modeling. The review also sheds light on the different types of information that can be extracted from short text messages.

The importance of information extraction from short text messages is highlighted by the potential applications of the extracted information in various domains such as healthcare, marketing, and security. The review summarizes the current state of the field, identifies areas for future research, and discusses the potential applications of information extraction from short text messages. Overall, this literature review provides valuable insights into the challenges, techniques, and potential applications of information extraction from short text messages.

II. Literature Review

A. *Design and Implementation of SMS Extraction and Analysis System [1]*

The literature suggests that Regular Expressions (regex) are widely used in text processing for pattern matching and data extraction from short text messages. The syntax rules described by a single string are used to match a series of sentences and retrieve or replace text that fits a particular pattern.

Different types of messages require different regular expressions, as shown in Table 1. For instance, in bank SMS messages, key information such as the bank name, bank card number, transaction amount, transaction time, and transaction details are necessary. However, some of this information needs to be formatted to show in charts for easier analysis.

TABLE I
Regular Expression for Bank SMS

Function	Regex Expression	Result
Extract Card	Suffix \w+	Suffix:6666
Extract Bank	(?<=[)[\4e00-\u9fa5]+(? =])	the bank name
Extract Left Money	Acc. Bal.:\s*Rs.\d+.\d+	Acc. Bal.: Rs. 75.32

B. *Information extraction from short text messages [4]*

The literature suggests that classification of messages into categories and extraction of information using neural networks is a promising approach in the field of natural language processing. The authors of the paper found that preconfigured extractors are not feasible and that preprocessing and annotation of messages are time-consuming tasks.

One of the main challenges in the work was the preprocessing and annotation of messages, which required a significant amount of time and effort. The authors suggest that annotating more data and familiarizing oneself with the character of the messages before embarking on the annotation task could improve the consistency and accuracy of annotation decisions.

The literature suggests that classification and information extraction tasks have various applications, including sentiment analysis, customer feedback analysis, and chatbot development. The proposed approach in the paper could be applied to different domains and could provide valuable insights into text data.

Overall, the literature suggests that the proposed approach in the paper is promising and could be further improved by addressing the challenges in preprocessing and annotation and by exploring different neural network architectures. The application of the proposed approach could provide valuable insights and improve various text-based tasks.

C. Named Entity Recognition for Short Text Messages [3]

The paper presents a named entity recognition (NER) system based on regular expressions and corpus-driven classifiers. However, the authors faced a major challenge in collecting a large SMS corpus due to the sensitive and personal nature of the messages. The lack of SMS corpora limits the possibility of training a good model that can deal with most cases. Additionally, the unique properties of SMS messages, such as brevity and lack of context, make recognition complex.

Furthermore, the authors note that their part-of-speech (POS) tagger is trained on newspaper text, which differs in style and language from text messages, resulting in degraded tagger accuracy. These challenges highlight the need for specialized training data and models that can handle the specific characteristics of SMS messages.

Overall, The paper highlights the difficulties in applying classical machine learning techniques to SMS NER and the importance of considering the unique properties of SMS messages. Future research in this area could focus on developing better methods for collecting SMS corpora and exploring more effective models for SMS NER.

D. Extracting Information from Short Messages [4]

The authors note several areas where the system could be improved, including handling more complex sentence structures, fuzzy word checking, and learning sentence structures. Additionally, the authors suggest extending the context mechanism to hold more of the history, handling negative or conflicting information, and managing synonyms at the data level.

Overall, while the system described in literature used by authors represents a significant step forward in the extraction of information from short text messages, there is still much work to be done to improve its accuracy and scalability. Further research is needed to develop systems that can handle the unique challenges presented by short text messages and other forms of unstructured data.

TABLE II
LITERATURE REVIEW TABLE

S.no	**Author**	**Year of publication**	**Paper Title**	**Observations**
1.	Ahmed Abbas, Josef Holmberg	2019	Information extraction from short text messages[4]	Promising approach with challenges in preprocessing and annotation, and need to explore different architectures for improvement.
2.	Yu-Yang WAN, Hui XU, and Rong-Rong FANG	2016	Design and Implementation of SMS Extraction and Analysis System[1]	Extraction can be done using Regex in some cases
3.	Tobias Ek, Camilla Kirkegaard, Håkan Jonsson and Pierre Nugues	2011	Named Entity Recognition for Short Text Messages[2]	NER technique to find words and their labels from SMS
4.	Richard Cooper, Sajjad Ali and Chenlan Bi	2005	Extracting Information from Short Messages[4]	Light weight Information extraction component which uses pattern matching

REFERENCES

[1] Wan, Yu-Yang, Hui Xu, and Rong-Rong Fang. "Design and Implementation of SMS Extraction and Analysis System." In ITM Web of Conferences, vol. 7, p. 04019. EDP Sciences, 2016.

[2] Ek, Tobias, Camilla Kirkegaard, Håkan Jonsson, and Pierre Nugues. "Named entity recognition for short text messages." Procedia-Social and Behavioral Sciences 27 (2011): 178-187.

[3] Cooper, Richard, Sajjad Ali, and Chenlan Bi. "Extracting information from short messages." In Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005. Proceedings 10, pp. 388-391. Springer Berlin Heidelberg, 2005.

[4] Abbas, Ahmed, and Josef Holmberg. "Information extraction from short text messages." LU-CS-EX 2019-18 (2019).