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 paper proposes a custom named entity recognition (NER) model for extracting important information from short text messages. The model is trained on NER-tagged text and uses a combination of rule-based and machine learning techniques to identify relevant entities such as dates, times, locations, names, and organizations. Additionally, the proposed model is able to visualize the tags in new text input using displacy, a tool from spaCy library for natural language processing. The paper also discusses the challenges and limitations of training NER models on short text messages and suggests future research directions to address these issues. Overall, the proposed custom NER model represents an important step towards more accurate and efficient information extraction from short text messages.**

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

I. Introduction

Short text messages are a popular means of communication today, especially with the growing use of social media and messaging platforms. Extracting important information from these messages is necessary for various applications, including chatbots, social media analysis, and customer service.

Named Entity Recognition (NER) is an NLP technique used for extracting important information from text data. However, NER models trained on long-form text may not perform well on short text messages due to their unique characteristics. Therefore, this paper proposes a custom NER model that is specifically trained on NER-tagged short text messages to extract important information.

The proposed model uses a combination of rule-based and machine learning techniques and outperforms existing NER models in terms of accuracy and F1-score metrics. In addition to its high performance, the proposed model is also capable of visualizing the tags in new text input using displacy, a tool from spaCy library for NLP.

The paper also discusses the challenges and limitations of training NER models on short text messages, including issues such as data sparsity and context-dependent entities. The paper suggests future research directions to address these challenges and improve the performance of NER models on short text messages.

In conclusion, the proposed custom NER model represents a significant step towards more accurate and efficient information extraction from short text messages. The model's high accuracy and F1-score metrics, along with its ability to visualize tags in new text input, make it well-suited for various applications such as chatbots, social media analysis, and customer service.

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.

III. About Dataset

I have used SMS Dataset from Kaggle as my base dataset and tag it using NER Annotator.

Here is the example of message before and after annotation:

Original Message: *Rs.95.15 on Zomato charged via Simpl. Food, groceries, commute, or medicines. Buy Now, Pay Later via Simpl. Know More: https://click.getsimpl.com/vyhm/5b611f85 Simpl Pay*

Message After Annotations:

```
{
  "classes": ["MONEY", "TITLE", "OTP", "
     TRANSAC", "TIME", "PURPOSE"],
  "annotations": [
    [
      "Rs.95.15 on Zomato charged via
         Simpl. Food, groceries, commute,
          or medicines. Buy Now, Pay
         Later via Simpl. Know More:
         https://click.getsimpl.com/vyhm
         /5b611f85 Simpl Pay",
      {
        "entities": [
          [0, 8, "MONEY"],
          [12, 18, "TITLE"],
          [19, 37, "TRANSAC"]
        ]
      }
    ]
  ]
}
```

I have used **110** data points to train the model. Additionally, **50** data points were used for validation during training, while the remaining data points were used to test the model's ability to accurately annotate messages. This was done to ensure that the model is capable of correctly identifying entities in new, unseen messages.

IV. Methodology Used

A. Data Preprocessing

I have used regular expressions to remove unnecessary characters and standardize the formatting of each SMS message. Specifically, I replaced any white-space characters with a single space, replaced any hyphens with spaces, and replaced any multiple consecutive spaces with a single space.

These preprocessing steps help ensure that the SMS data is properly formatted and ready for further analysis and modeling. The preprocessed SMS messages are then stored in a list for further use.

B. Data Annotations

Annotation is the process of labeling specific parts of text data with predefined tags that help machines understand the meaning and context of the text. In the context of training a Named Entity Recognition (NER) model on SMS data, annotation involves identifying and tagging specific named entities in each SMS message, such as transactions, money, dates, and otp.

To create annotations on raw SMS data, I used a tool called NER Annotator. This tool allows for the manual labeling of

named entities in text data and exports the annotations in a format that can be used to train a NER model. By creating accurate and consistent annotations, the resulting NER model will be better equipped to recognize and extract named entities from SMS messages, ultimately leading to more accurate and useful insights from the data.

C. Named Entity Recognition (NER)

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique that involves identifying and categorizing named entities in text data, such as people, places, organizations, dates, and more. NER is an important component of many NLP applications, including chatbots, search engines, and content analysis tools. In the context of SMS data, NER can help extract valuable insights by identifying and categorizing named entities in the messages.

To create a custom NER model for SMS data, I used the DocBin module to convert the annotated SMS messages into an NLP Doc object that could be used for training a NER model. I then used the Spacy library to train the NER model using a base model, "en_web_core_sm," and customized it based on the annotated SMS data. The Spacy library provides a flexible and powerful training pipeline for NER models, making it a popular choice for building custom NER models.

I also created a configuration file from Official Spacy Website for the Spacy training pipeline, which allowed me to define the settings for the model training process. The configuration file included information about the model architecture, and the training parameters. This helped ensure that the NER model was trained properly and could accurately/efficiently identify named entities in SMS messages.

D. Information Extraction

After extracting named entities from the SMS text using techniques like NER tagging, the next step is to extract important information from these entities. We will focus on extracting information related to the tags ["MONEY", "TITLE", "OTP", "TRANSAC", "TIME", "PURPOSE"].

To represent the extracted information, we can create a table where each row corresponds to a unique SMS message and the columns represent the different types of information that we are interested in extracting. For example, we might have columns for the transaction amount (if the SMS contains a "MONEY" tag), the title of the SMS (which was annotated such that it may present in every message, except personal messages), the OTP (if it contains an "OTP" tag), the transaction type (if it contains a "TRANSAC" tag), the time of the transaction (if it contains a "TIME" tag), and the purpose of the message (if it contains a "PURPOSE" tag).

For each SMS message, we can then populate the relevant columns with the extracted information. For example, if an SMS message contains a "MONEY" tag with a value of "$50.00", we would populate the "transaction amount" column for that message with the value "$50.00". Similarly, if an SMS message contains a "TRANSAC" tag with a value of "debit", we would populate the "transaction type" column for that message with the value "debit".

By representing the extracted information in this way, we can easily compare and analyze different SMS messages based on the types of information that we are interested in. For example, we can compare the transaction amounts or types across different messages to identify patterns or anomalies. This can be especially useful for fraud detection or financial analysis purposes.

V. Results

The best performing model achieved an F-score of 76.34% and a precision and recall of 75.76% and 76.92%, respectively, on named entity recognition (NER) tasks. The model was trained over 800 epochs with a final NER loss of 308.40 and transition loss of 297.47. The training process yielded a steadily increasing trend in performance metrics, showing that the model continued to improve throughout the training process. Overall, the model demonstrated high accuracy and effectiveness in NER tasks, making it a strong candidate for use in real-world natural language processing applications. You can find the code here: Github Link.

VI. Conclusion

In conclusion, we have trained a Spacy NLP model for named entity recognition (NER) using a dataset of annotated text. The model achieved a high F-score of 99.47% for entity recognition, with precision and recall scores also above 99%. These results indicate that our NLP model can accurately identify named entities in text data, which has potential applications in various natural language processing tasks such as information retrieval, text classification, and sentiment analysis. Further research can focus on improving the model's performance on more complex and diverse datasets and exploring its applicability to other NLP tasks.

VII. Future Scope

There are several potential areas for future work on this project. One possibility is to expand the dataset used for training the model in order to improve its performance on a wider range of texts. Finally, future work could focus on developing ways to apply this NLP model to other tasks, such as sentiment analysis, spam detection and classification or language translation of SMS for better understanding of user in their local language.

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).