IMPORTANT INFORMATION EXTRACTION FROM SHORT TEXT MESSAGES

Presented By:
Abhay Parihar (IIT2020074)

Introduction

Proposed custom NER model for extracting important information from short text messages

Short text messages are popular means of communication and require special NER models

Model uses combination of rule-based and machine learning techniques

Outperforms existing NER models in accuracy and F1-score metrics

Model has potential for various applications such as chatbots, social media analysis, and customer service

Literature Review

Author	Year of Publication	Paper Title	Observation
Ahmed Abbas, Josef Holmberg	2019	Information extraction from short text messages[4]	Need to explore different architectures for improvement.
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
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
Richard Cooper, Sajjad Ali and Chenlan Bi	2005	Extracting Information from Short Messages[4]	Light weight Information extraction component which uses pattern matching

About Dataset

- SMS Dataset from Kaggle
- Annotate the dataset on following entities:
  - "MONEY",
  - "TITLE",
  - "OTP"
  - "TRANSAC"
  - "TIME"
  - "PURPOSE"
- **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

About Dataset (Contd.)

- **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"]
        ]
      }
    ]
  ]
}
```

Methodology Used

- **Data Preprocessing:**
  - used regular expressions to remove unnecessary characters
  - replaced any white-space characters with a single space
  - replaced any hyphens with spaces
  - replaced any multiple consecutive spaces with a single space
- **Data Annotations:**
  - Labeling specific parts of text data with predefined tags
  - 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.

Methodology Used (Contd.)

- **Named Entity Recognition (NER):**
  - To create a custom NER model for SMS data, I used the DocBin module
  - Used the Spacy library to train the NER model using a base model, "en_web_core_sm".
  - Customized it based on the annotated SMS data
- **Information Extraction:**
  - We will focus on extracting information related to the tags ["MONEY", "TITLE", "OTP", "TRANSAC", "TIME", "PURPOSE"]
  - 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.

Results

- **F1 Score:** 79.03%
- **Precision:** 83.05%
- **Recall:** 75.38%
- **Overall Score:** 0.79
- **Final NER loss:** 359.89
- **Final Transition Loss:** 334.01

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).