SK[AI] IS THE LIMIT: SERVICE AND MANUAL PREDICTION

Srikanth Gurram Anastasia Sycheva Sai Yaswanth

ETH Zurich, Switzerland

ABSTRACT

Find a solution to predict which service is the incident about and which manual would work best to solve the incident

1. IDEAS CONSIDERED

- Use TF-IDF weighted word embedding centroids as a doc embedding proxy
- Construct TF-IDF from word vector similarity
- LSTM model with pretrained word embeddings
- BERT Models
- Doc2Vec embeddings as features for classical ML models

Upon quick implementation of the above ideas, the last two showed most promise. Hence, they have been the focus of our work since.

2. DATA PREPARATION

- Clean the subject - removing the prefixes and alphanumeric strings.
- Clean the body.
- Filter the mails based on language.

3. MODELS

BERT: Bidirectional Encoder Representation for Transformers have been proven to be the state of the art models when it comes to any NLP task. Our best performing model was a BERT based classification model initialized with pretrained weights. We use $dbmdz/bert-base-german-uncased$ *by Hugging Face as our pretrained weights.*

We also trained a doc2vec model that converts the incident email into a embedding vector which is used by an Support Vector Classifier to perform predictions. We did not achieve impressive results through this approach. But nevertheless, it's a good model to consider.

4. RESULTS

4.1. BERT Models

- *Validation F1 score: 0.71*
- *Train F1 score: 0.97*

4.2. Doc2Vec + SVC Models

- *Validation F1 score: 0.45*
- *Train F1 score: 0.51*