USING NAMED ENTITY RECOGNITION MODELS FOR STARTUP DOMAIN

Università di Pisa

TEXT ANALYTICS

Muhammad Umair Naeem

Supervised by : Prof. Andrea Esuli

INTRODUCTION

Named Entity Recognition (NER) is one of the fundamental undertakings in natural language processing. Most notable NER data sets consist of information from news archives with three kinds of named elements or entities marked: persons, organizations, and locations. For these kinds of named entities, the best-in-class NER strategies as a rule give noteworthy outcomes. Nonetheless, in explicit spaces, the exhibition of NER frameworks can be a lot lower because of the need to present new sorts of entities, to build up the standards of their marking, and to annotate them reliably.

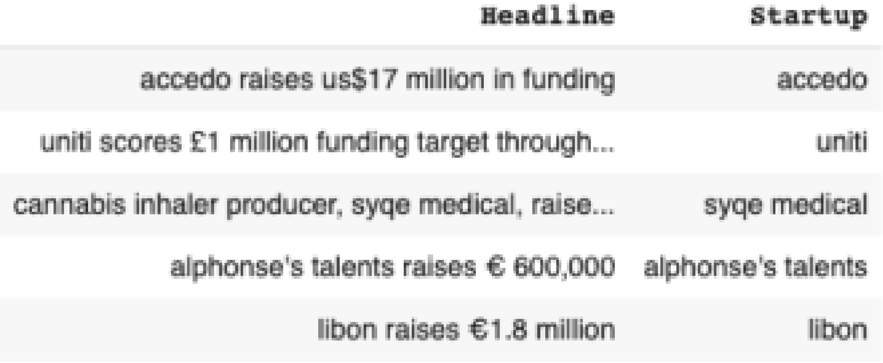
In this paper, we discuss the NER task in the startup domain. The task at hand was to identify the named entity ‘ORG’ from news headlines.

The effort of this paper was to find the process that makes data collection easy, improving data accuracy and reducing data redundancy using different NER models and knowing which natural-entity-recognition model is more accurate and accessible in the process.

We experimented with different models to analyze the performance and to choose the most suitable model for our task.

DATA UNDERSTANDING

The dataset we have used for training the models in this project is collected manually. This included going through every headline and finding the name of the startup which received funding.



Headline	Startup
accedo raises us$17 million in funding	accedo
uniti scores £1 million funding target through...	uniti
cannabis inhaler producer, syqe medical, raise...	syqe medical
alphonse's talents raises € 600,000	alphonse's talents
libon raises €1.8 million	libon

Our data set consisted of 2797 entries with ‘Headline‘ and ‘Startup’ as columns, after dropping one row out as it was not appropriate for our task. The data was already cleaned so we did not need the process of data cleaning and also there were ‘0’ null values present in the data.

MODELS USED IN STARTUPS’ NAMED ENTITY RECOGNITION

spaCy

spaCy[1] is designed specifically for production use and helps you build applications that process and “understand” large volumes of text. It can be used to build information extraction or natural language understanding systems or to pre-process text for deep learning.

SIMPLE TRANSFORMERS

Simple Transformer[2] models are built with a particular Natural Language Processing (NLP) task in mind. Each such model comes equipped with features and functionality designed to best fit the task that they are intended to perform. The high-level process of using Simple Transformers models follows the same pattern.

- Initialize a task-specific model
- Train the model with train_model()
- Evaluate the model with eval_model()
- Make predictions on (unlabelled) data with predict()

Models we have explored from this library:

- DistilBERT

DistilBERT[3] is a method to pre-train a smaller general-purpose language representation model, by knowledge distillation during the pre training phase it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

- RoBERTa

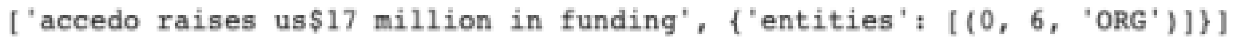
RoBERTa[4] builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

DATA PRE-PROCESSING

For our work we required a single annotation to define the named entity which is the startup's name. As mentioned above the data was clean, one of our preprocessing tasks was to do 'Case Normalization'. Then we used two libraries namely: Spacy and Simple transformers for implementing all our models.

Preparing data for spaCy model

Spacy needs the data with every heading as a list containing the entries as shown below.



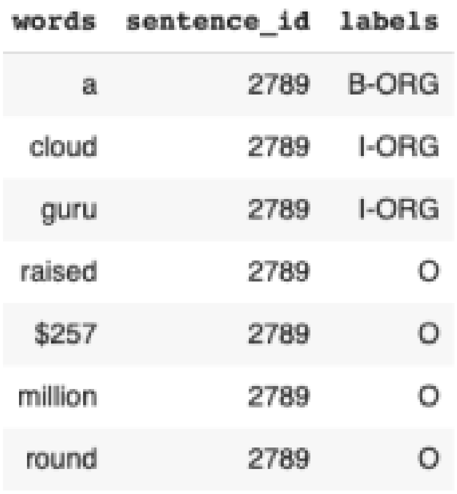
```
['accedo raises us$17 million in funding', {'entities': [(0, 6, 'ORG')]}]
```

We now have data in the form that is required for the Spacy model and the next step is to split the data into training sets and test sets.

After splitting the data we have 2238 rows of data in training and 559 rows for testing data and we have loaded the default spacy model and set up the ner ML pipeline. We got the whole data annotated with '{'ORG': 2797}'.

Preparing data for simple transformers

For simple transformers, we prepare the data in CoNLL format in the pandas dataframe. We annotated the data for the startups as 'B-ORG' for the beginning of the startup name, 'I-ORG' for the continuation of the startup's name, and 'O' for all other tokens as we were only interested in identifying the name of the startups in our use case.



words	sentence_id	labels
a	2789	B-ORG
cloud	2789	I-ORG
guru	2789	I-ORG
raised	2789	O
$257	2789	O
million	2789	O
round	2789	O

After getting the annotated data we split the data into 3 parts: train, validate and test set.

PERFORMANCE OF THE MODELS

We used our test set to do a model evaluation after training and got the following results:

Model Name	Recall	Precision F-Measure Eval_loss
Spacy	0.91256830 60 10929	0.9209558823 0.9167429094 NA 529412 236047
RoBERTa	0.87182448 03 69515	0.8483146067 0.8599088838 0.0582046538 41573 268792 025939
DistilBERT	0.92995391 70 506912	0.9273897058 0.9286700414 0.0477788950 823529 173954 8487923

The model predicts named entities inside the textual content inside the evaluation dataset. Each entity predicted with a score above the edge is in comparison to the actual entities inside the dataset to generate precision.

PRECISION

Precision[5] expresses the proportion of the organizations in the dataset that the given model says were organizations that actually were relevant organizations.

As per our workings, we observe that the distilBERT model outperforms the others with a Precision score of 0.9273897058823529.

RECALL

Recall[5] expresses the ability of the given model to find all relevant organizations in a dataset.

Here we can see that again the distilBERT model has the highest Recall value of 0.9299539170506912, amongst all the models.

F-MEASURE

In cases where we want to find an optimal blend of precision and recall we can combine the two metrics using what is called the F-MEASURE[5].

If we want to create a balanced classification model with the optimal balance of recall and precision, then we try to maximize the F-MEASURE.

Here again, we observe that the F-Measure score amongst all the models was the highest for the distilBERT model at 0.9286700414173954.

We have obtained these results for the spaCy model after having trained 200 epochs and the accuracy score for our model stands at 89.80%.

CONCLUSION

In this paper we present the results of applying spaCy, distillBERT, RoBERTa to named entity recognition to identify the named entity ‘ORG’ from news headlines. We compared the aforementioned models and the highest F-Measure was shown by the distilBERT model at 0.9286700414173954.

After analyzing the performance of all the three models we found that the distillBert model was the best in terms of Precision, Recall and F-Measure followed by spaCy and RoBERTa respectively.

DistilBERT was the clear winner as it not only performs better in performance metrics but also it’s a lightweight model so it will provide a responsive model in production.

Bibliography

1. spaCy. “What’s spaCy?” spaCy 101: Everything you need to know, spaCy, https://spacy.io/usage/spacy-101#whats-spacy.

2. Simple Transformers. “Simple Transformers.” Named Entity Recognition Specifics, https://simpletransformers.ai/docs/ner-specifics/.

3. The Hugging Face Team. “DistilBERT.” DistilBERT, https://huggingface.co/transformers/model_doc/distilbert.html.

4. The Hugging Face Team. “RoBERTa.” RoBERTa, https://huggingface.co/transformers/model_doc/roberta.html#.

5. Towards Data Science. “Beyond Accuracy: Precision and Recall.” Choosing the right metrics for classification tasks, https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c.