Machine Learning Nano Degree

Capstone Project Proposal

Domain Background:

As a human being, we can quickly grasp a conversation, an opiniated article, or a book and interpret it. Sometimes, even humans find it difficult to interpret the general sentiment of all the writers – for instance, it might take forever to understand the widely mis-understood character of Karna (in the epic Mahabharata); one has to go through the most popular negative, neutral and positive interpretations of different writers before coming to a conclusion. It is similar to understanding the sentiment of people during general election or before release of a popular new product like an iPhone. For computer, it is even more challenging to understand and analyze the natural language. It will be a great technological feat if computers can understand the language that we speak on a regular basis at a granular level. There have been many advancements in this direction and is popularly known as 'Natural Language Processing (NLP)".

NLP is the ability of a computer to understand, process and analyze the natural human language like English, Spanish or Chinese by using the context, text and speech as input. There are many components to NLP and can be broadly categorized into Named Entity Recognition (NER), syntactic analysis, semantic analysis, sentiment analysis and pragmatic analysis. This project is focused on Named Entity Recognition and hence will not discuss other topics.

Problem Statement:

Named Entity Recognition is essentially an information extraction task. It involves segmenting a sentence to identify and extract entities such as a person, geographical location, organization, events etc., One of the main challenges is to match different variations of an entity and cluster them as the same. One instance would be a person's name with first and second name – both of them should be clustered as same. Some popular applications of NER include – classifying news articles, efficient search algorithms, content recommendations (practically applied in recommending news articles based on reader's history and interests), customer support (in categorizing complaints), research papers (segmenting papers based on tags) etc., A search for Named Entity Recognition in Google Scholar show more than 15,000 results in 2018 alone.
(https://scholar.google.com/scholar?as_ylo=2018&q=named+entity+recognition&hl=en&as_sdt=0,36)

This shows the amount of active research happening in the field. Also, there have been many applications in biomedical area recently (Wang, Xu, Chen Yang, and Renchu Guan. "A comparative study for biomedical named entity recognition." *International Journal of Machine Learning and Cybernetics* 9.3 (2018): 373-382.)

My primary motivation to choose this project is to work on something novel that has not been covered in the MLND. This gives me an opportunity to learn about NLP and know about other areas where ML is being applied widely. The goal for this project is to identify the named entities from the annotated corpus data taken from Groningen Meaning Bank (GMB).

Dataset:

In this project, the goal is to create a good machine learning model to classify named entities in the text. This is an information extraction project where we try to extract the following classes:

- geo: Geographical Entity
- org: Organization
- per: Person
- gpe: Geopolitical Entity
- tim: Time Indicator
- art: Artifact
- eve: Event
- nat: Natural Phenomenon
- O: Other

Each of the above classes has two sub-classes and hence a total of 17 classes. There are 4 columns in the dataset

1. Sentence #
2. Word
3. POS (part-of-speech tag)
4. IOB tag (Inside-Outside-Beginning tag)

There is a total of 2999 sentences with 66161 words. Each word is taken as a data point and relevant features will be generated. However, care must be taken when splitting the data into test/train data sets to avoid data bleeding. The words are first converted to sentences of which 2400 sentences (80% of the data) are taken into training set and rest of the 600 sentences will be considered as test set. While dealing with neighboring words, care will be taken not to include next/previous sentence's information.

We need to generate new features (feature engineering) and try to predict column_4 (IOB tag) using a good ML classifier.

The dataset is taken from

https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus#ner_dataset.csv

Solution Statement:

There are many open source NER classifiers like Stanford CoreNLP, SpaCy etc., But those packages are not utilized here and everything is built from scratch. The general idea is generate new relevant features and then apply a regular ML classifier like gradient boosting or randomforest algorithms. An essential thing in NLP problems is to identify the context and hence our features should include a way capture the context of the particular word – for instance the previous and next word etc., Also parts of speech, punctuation, root words are some features that are helpful. Details of the features and algorithm to generate them will be discussed in the project.

Benchmark model:

A simple baseline model would be a memorization model where we create a dictionary of words and their tags: just remember the most common named entity for every word and predict that. Incase, we don't know that word predict it as non-entity (Other).

Evaluation metric:

This is a classification problem and the dataset has 85% of the words classified as 'O'. Fraction of words belonging to each category is as follows:

O 0.849700
B-geo 0.031287
B-org 0.018697
I-per 0.018651
B-gpe 0.018591
B-tim 0.017533
B-per 0.016732
I-org 0.013996
I-geo 0.006257
I-tim 0.005048
B-art 0.000801
B-eve 0.000680
I-eve 0.000559
I-art 0.000514
I-gpe 0.000514
B-nat 0.000302
I-nat 0.000136

Because of the skewness in the data, prediction accuracy is not a right measure for algorithm effectiveness. I will use Precision, Recall & F1-score as the evaluation metrics. Since, it is a multi-class classification problem, it is required to look at these scores for each of the categories.

Project Design:

Outline of the project is as follows:

1. Data Exploration & Filling missing data: Skewness in the data and missing values will be explored
2. Baseline model: A baseline "memorization algorithm" will be evaluated
3. Feature Engineering: New features will be generated that capture the context, root words, parts of speech, stop words etc.,
4. Supervised learners: Some supervised classifiers like randomforest and gradient boosting will be tried and explored for the best algorithm
5. Evaluation of classifier: The best classifier will be evaluated using learning curves (data limiting?) and Complexity curves (high bias / high variance ?)
6. Tuning the model: The chosen model will be tuned for optimal performance using GridSearch
7. Variable importance & Sensitivity: Importance of different features and sensitivity of the model will be analyzed
8. Conclusions and Remarks