Primary Approaches to Text-based Classification

There exist two prevailing types of solutions to text-based classification problems. One is a rules-based text classification and the other is the machine-learning-based text classification.

In the rules-based text classification method, one employs the use of manually constructed logical rules to classify text into specific categories. Without the use of a rules-engine, such an

implementation would resemble an extensively built and potentially nested if-else statements [1]. Because of the manual nature of building and maintaining these set of logical rules it is expected to

come with some maintenance overhead as business rules evolve over time.

For machine-learning-based text classification, one trains and tunes a supervised machine learning model to process a corpus of words found among the dataset to learn how to classify bodies of text.

This path requires significant but non-repetitive upfront time and effort to build the training data set as well as fine-tuning the training process to teach the algorithm how best to classify bodies of text

[1]. However, it is expected to better handle the nuances of the language it has been trained on as compared to the rules-based approach.

Typical Natural Language Processing (NLP) requires a preprocessing stage where the source raw bodies of text are cleaned of unproductive elements and features are extracted in a manner that a machine learning algorithm can ingest and process [2, 3]. This is a necessary dependency for the machine-learning approach but it may also prove useful for the rules-based approach.

Purpose of Rules-based Approach

The purpose of the Rules-based approach in the context of using a rule engine is to abstract away the implementation of the rules at code level. This frees us from having to write complex if-else

statements and enables us to reconfigure the rules freely for sake of experimenting which is crucial for this initial stage of the project [4]. There exists an open-source Python library where we can

accomplish just this very goal in [5].

In the context of the use case of performing property and loan segmentation on our own data set, it is the first step to use building a training data set for the sake of training a machine learning model

to take over the classification of properties and their associated loans.