Application and challenges in Text Classification

* What are some practical applications of text classification? Examples might include spam detection, sentiment analysis, topic modeling, document categorization, and customer service automation.

There are several practical applications for natural language processing:

1. Spam Detection: This involves automatically identifying spam emails or unwanted emails to prevent them from reaching our emails main inbox
2. Sentiment Analysis: In the world of finance, the tone, and ultimately the recommendation, of different articles can help drive the price of a stock. Therefore, it is useful to have some code read the opinion of different analysts and subject matter experts and determine whether a piece of code is positive, negative, or neutral.
3. Topic Modeling: This involves identifying general themes in large text corpora, one example of this could be customer feedback.

* What challenges often arise in text classification tasks? Consider aspects like class imbalance, ambiguous labels, overfitting, noisy or unstructured data, and domain-specific vocabulary.
  1. According to a medium article, “when the distribution of classes in a dataset is highly skewed, standard learning algorithms tend to be biased towards the majority class.” This can lead to bad predictions for less present classes.
  2. Ambiguous labels: Labels can be unclear or subjective, which makes it difficult to accurately train models based on these.
  3. Domain-Specific Vocabulary: Some domains use specific jargon / terminology, which requires models to adapt or even further train on the domain-specific data. This exposes our model to overfitting as highly specific data will likely have smaller training corpora.

Source: https://medium.com/@dinghan1995/how-to-tackle-dataset-class-imbalance-for-nlp-4453af6f6b87