In our experiment we have tried to classify the 20NewsGroups data using both supervised and semi supervised classifiers. We have to extract meaningful information from the contents of 20NewsGroups data so that we can train a classifier on it. To achieve this we have extracted features using Bag of Words, TF-IDF, Parts of Speech tagging, Word Embedding (GloVe) and Topic Modelling(LDA).

Since Bag of Words sometimes fail to capture the syntactic relations between words we have used different versions of BoW like bigrams and trigrams. We have also used POS tagging to extract features based on parts of speech. We are using GloVe as a method of word embedding feature extraction for our data as it outperforms other models such as word analogy, word similarity, and named entity recognition. We were also able to extract unique topics from the data as a set of words using LDA.

We implemented Self learning with SGD classifier, Label propagation, Multinomial Naive Bayes and Naive Bayes with EM.

Comparing with SGD classifier and Naive Bayes, the SSL model had better accuracy score. But Naive Bayes with EM performed better than self training and label propagation.

However, when TF-IDF and POS / TF-IDF and GloVe (features) were combined label spreading and label propagation we were getting very low accuracy scores.

We have learnt from this experiment that for text classification, even when we have less pre-labelled training data we can use semi supervised classifiers like Naive Bayes with EM and Label Propagation to classify an unlabelled instance with an average accuracy of 70%.

A ML model is only as good as the features it learns. So the performance of our models could improve if we had domain expertise which could help us do a better feature extraction.

| Feature Extraction | Supervised Accuracy | Semi-Supervised Accuracy |
| --- | --- | --- |
|  |  |  |
| TF-IDF + POS | 69.8% | 70.5% |
| LDA |  |  |
| Word Embedding |  |  |
|  |  |  |