News Group Identification using Semi-Supervised Learning

Team 09

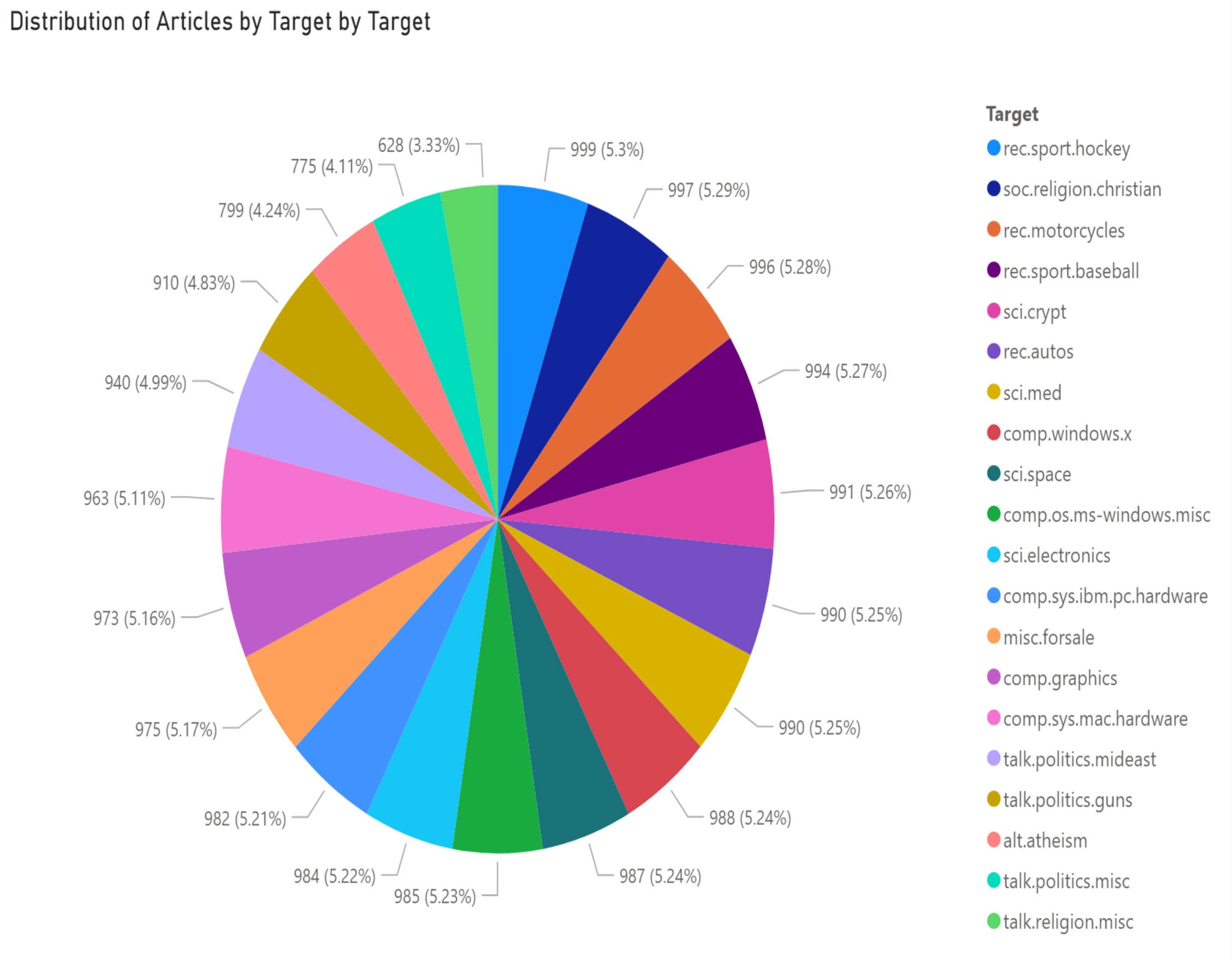
Otto-von-Guericke-Universität Magdeburg

MOTIVATION AND OBJECTIVES

- Labelled data is difficult and expensive to generate as compared to unlabeled data.
- Goal is to improve performance of classifier using labelled and unlabeled data.
- Implement and observe the performance of Expectation maximization SSL Algorithm.
- Compare performance of EM algorithm with baseline Label spreading algorithm.

DATASET

- Dataset consists of 20000 news articles with 20 different target classes.
- We use 60% articles to train the model, 20% to for model validation and 20% for model testing.
- Dataset Split - Initially we train the model with 20% of labelled data and 80% of unlabeled data. In next phases we slowly increase the amount of labelled training data and check the performance.



FEATURE EXTRACTION

Extracted features:

- Bag of Words: Count occurrences of words
- Skip-Grams: Count occurrences of word pairs while skipping words in between
- Part of Speech tagging: Assign a word a role it plays in a sentence e.g. noun
- LDA topic model: Learn a model on the dataset, that can generate a number of possible topics for a word
- Word2Vec: Learn a model on the dataset, that can generate a context vector for a word

Total number of feature elements: 4.16 million

Select 2500 elements for each extracted feature based on occurrence (removed frequently and rarely occurring) and combine them into a training vector with 12500 elements.

Topic 0
mail
program
ftp
file
window
available
image
use
windows
graphics
Topic 1
moon
space
gov
nasa
access
net
satellite
orbit
launch
digex

Libraries and implementation details

- Scikitlearn as main machine learning library
- Gensim for LDA and Word2Vec model training
- NLTK for text data preprocessing functions
- Custom implementation of scikitlearn's Vectorizer class for each feature

SEMI SUPERVISED LEARNING STRATEGY

Expectation Maximization:

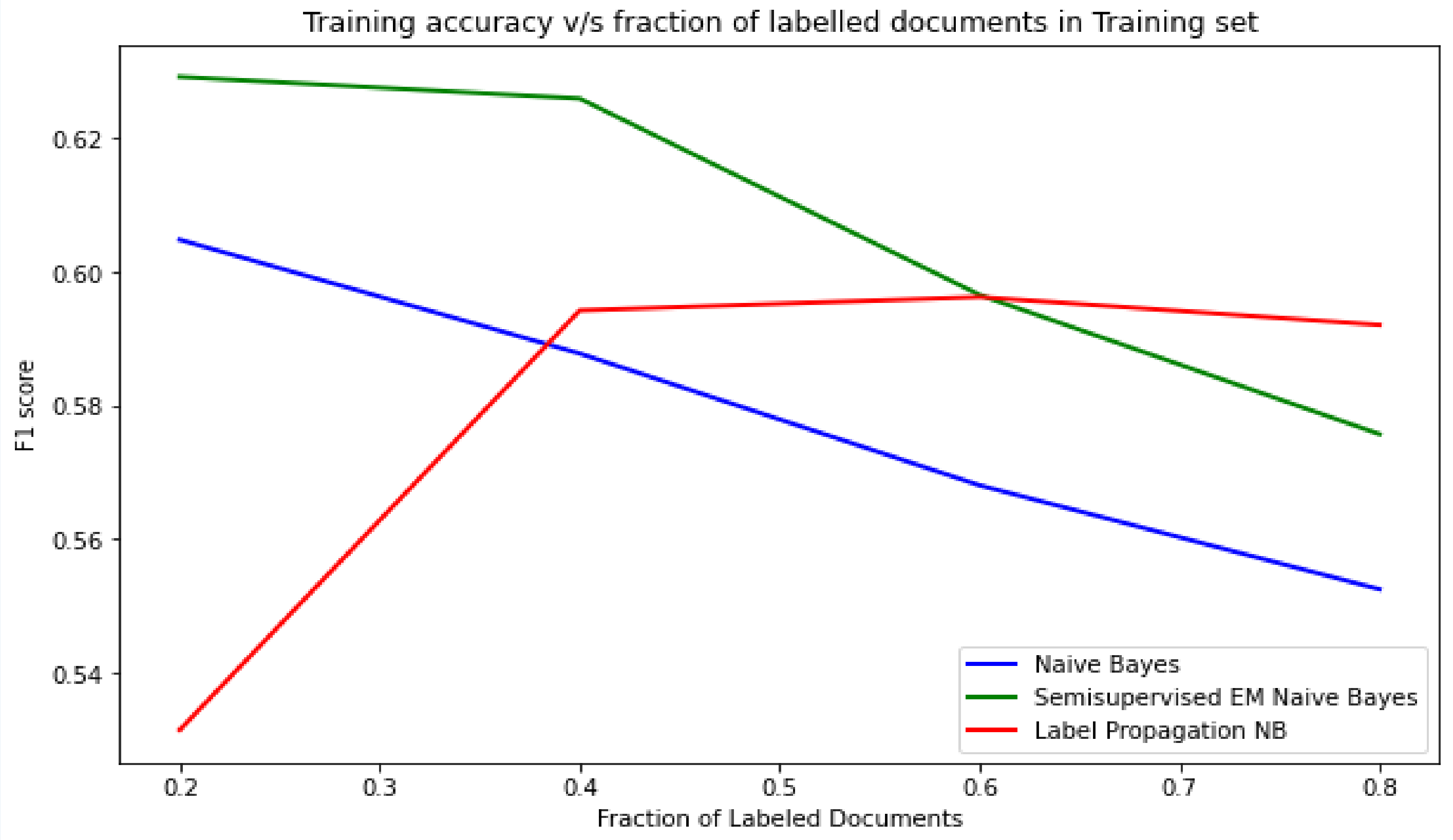
Expectation Maximization (EM) is an Semi-Supervised Learning approach where the model maximizes the probabilities of unlabeled data belonging to a class. This is done in two steps:

- E-Step: Estimate class membership on unlabeled data
- M-Step: Re estimate model parameters on labeled and the now labeled unlabeled data

Repeat the this until the class probabilities (expectation) converge.

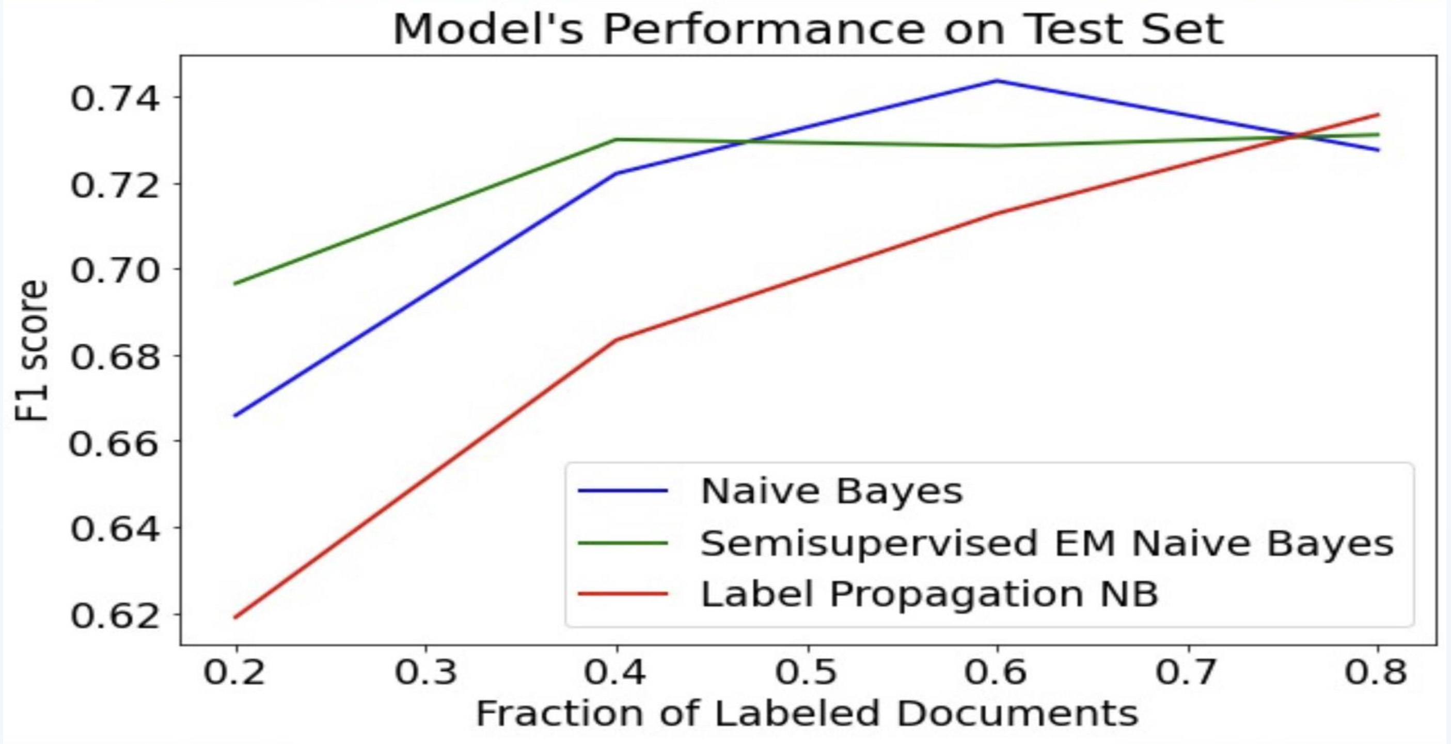
MODEL TRAINING

- Training a selection of classifiers using k-Fold Cross-Validation. In the graph below, the overall model performance is displayed. The classifiers are: Naïve Bayes, EM-Naïve Bayes, Label Propagation.
- EM-Naïve Bayes and Label Propagation are SSL-algorithms so they are trained also on the unlabeled portion of the data.
- By Comparing SSL-EM with Label propagation, SSL-EM performs better than Label propagation with least number of labeled data and in real world we always don't have much of labelled data. So we went ahead with SSL-EM as our classifier.



EVALUATION

- Testing the trained SSL-EM model on the disjoint test data set and compared it to other classifiers.
- With 40% of labeled data SSL-EM performs better compared to other algorithms.



CONCLUSIONS

- EM performs better in scenarios that have only a small ratio of labeled data.
- The higher the ratio of labeled data is, the better the other classifiers perform eventually overtaking EM-Naïve Bayes at around a 50/50 ration.
- Our EM approach is limited to Naïve Bayes which makes it vulnerable to negative values in the feature vector.
- We can say our model follows SAFE SSL strategy as performance increased by taking into account both labelled and unlabeled data compared to only labelled data for supervised algorithm.

REFERENCES

- https://github.com/jerry-shijieli/Text_Classification_Using_EM_And_Semisupervied_Learning/tree/master/code
- Kamal Nigam, Andrew McCallum, Tom Mitchell. (2002). Semi-Supervised Text Classification Using EM
- https://github.com/VXU1230/Medium-Tutorials