AUTOTAGGING PROJECT

Ozge Pervane*

*Email: ozgedincel@gmail.com, LinkedIn: linkedin.com/in/ozgeper

Introduction

Tällt Ventures is scanning the market to help their clients to access the latest trends and startups. We purposed to develop a machine learning model for their market search engine called **Sønr.** In this project, I developed a multi-classification machine learning model which can classify the startups automatically into their trend categories.

Product

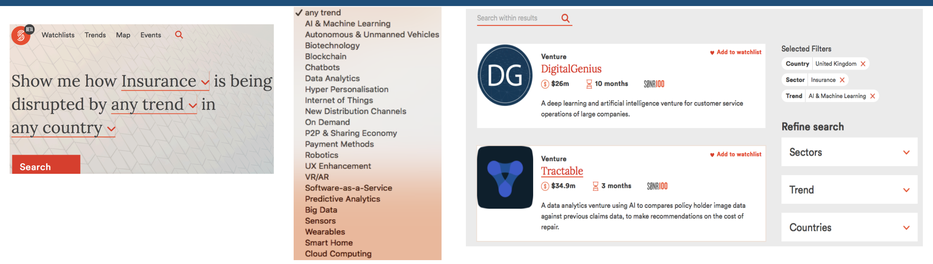

Figure 1: Sønr search engine 'https://sonr.global/dashboard'

Main Challenges

- **Highly Imbalance Data:** Imbalance ratio in terms of class proportions are 1:1:1:1:1:1:2:3:4:4:5:14:16:50. Imbalance data can cause overfit to the dense class which is "UX enhancement" in this data.

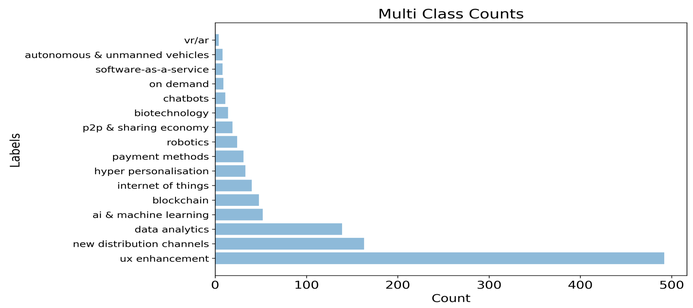

Figure 2: Multi class counts of the imbalance data

- **Scraped Text Data** include unnecessary words which can cause noise and extra features. Raw text data need to be extracted regarding feature.
- **Multi Labelled Data:** Data has multi-label classes which means a company can use more than one trend.

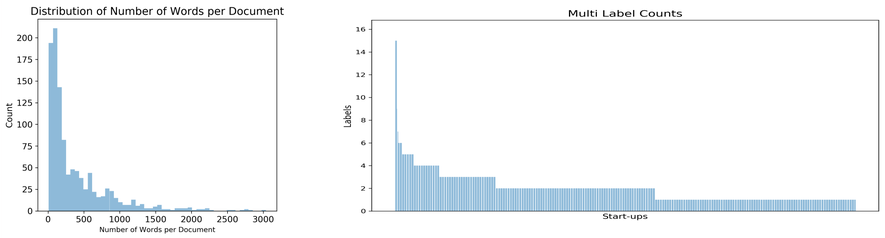

Figure 3: Distribution of scraped words and multi label class per document

Approach

Data Collecting and Preparing Steps,

- Collecting and merging data from MySQL database
- Data Exploratory Analysis
- Web Scraping from company website: body and about us sections and Facebook pages by using Scrapy
- Cleaning the scraped text dataset.

Modeling: NLP and Machine Learning with Python

- Pre Processing of the text: Replacing acronyms, synonyms and repeating characters, lemmatize, stemming, removing stop words, spelling corrections.
- Feature Engineering: Word2Vec, Latent Semantic Analysis (LSA), TF-IDF, N-Gram Range.
- Modeling: Cross validated data used with OneVsOne or OneVsRest Classifiers for Logistic Regression, Random Forest, SVC, Naive Bayes or KNN.
- Evaluate the Model Performance: Confusion Matrix and ROC Curve

Results

- Scraped raw text data is extracted by NLTK and replacing codes, and text pre-processing steps had a significant effect to decrease the accuracy of the model.
- Random Over Sample method adjusted the distribution of the data which decreases class imbalance.
- One-vs-one multiclass strategy worked best for the Logistic Regression classifier.
- Baseline accuracy is 0.45, and the model accuracy is 0.63 which means this model increases the performance around 40%.

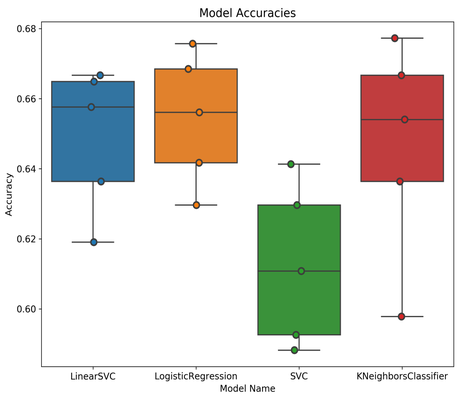

Figure 4: Comparing Model Accuracies

Evaluate Model Performance

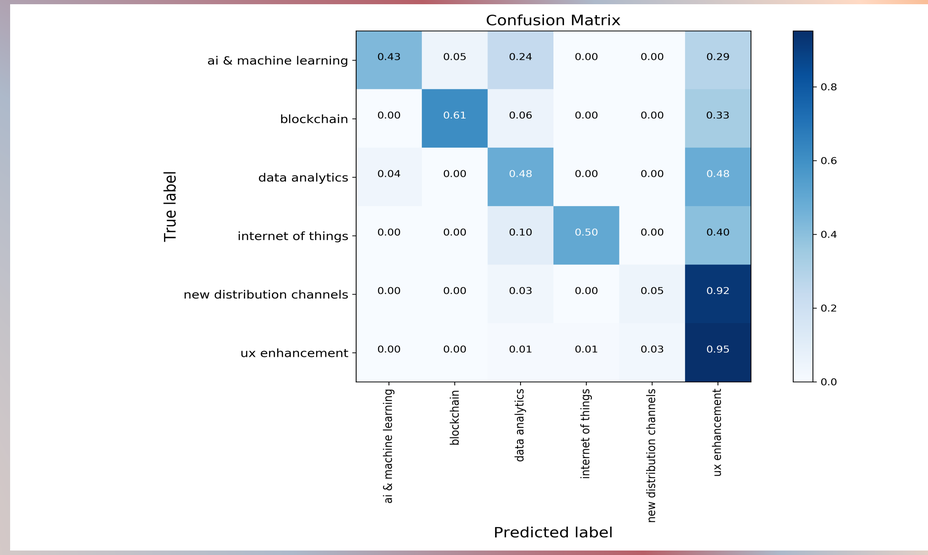

Figure 5: Normalized Confusion Matrix evaluates the quality of the output of the classifier

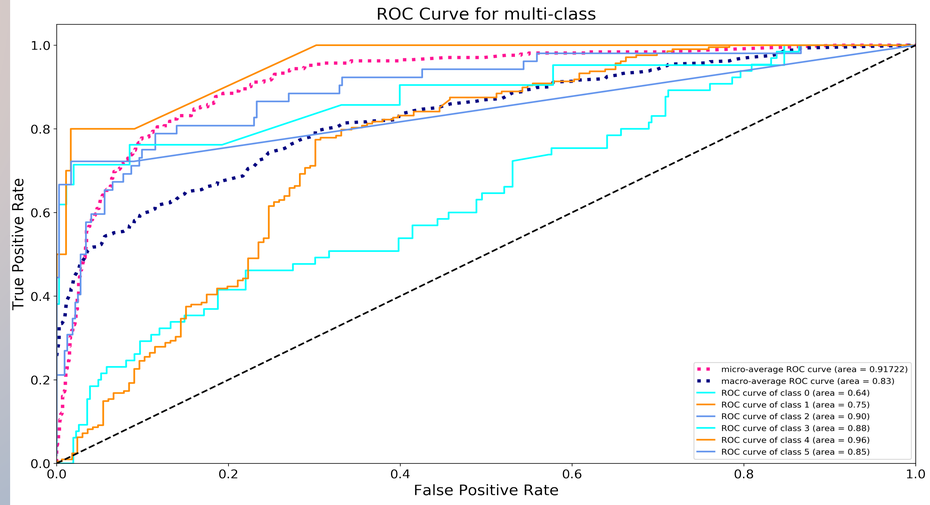

Figure 6: ROC Curve shows the performance of the classification model

Conclusion

This model is developed for the single label multi-classification model. Developed model increased the accuracy of the baseline around 40%. This project showed that the scraped text data from companies websites and Facebook pages could be used to guess their trend categories.