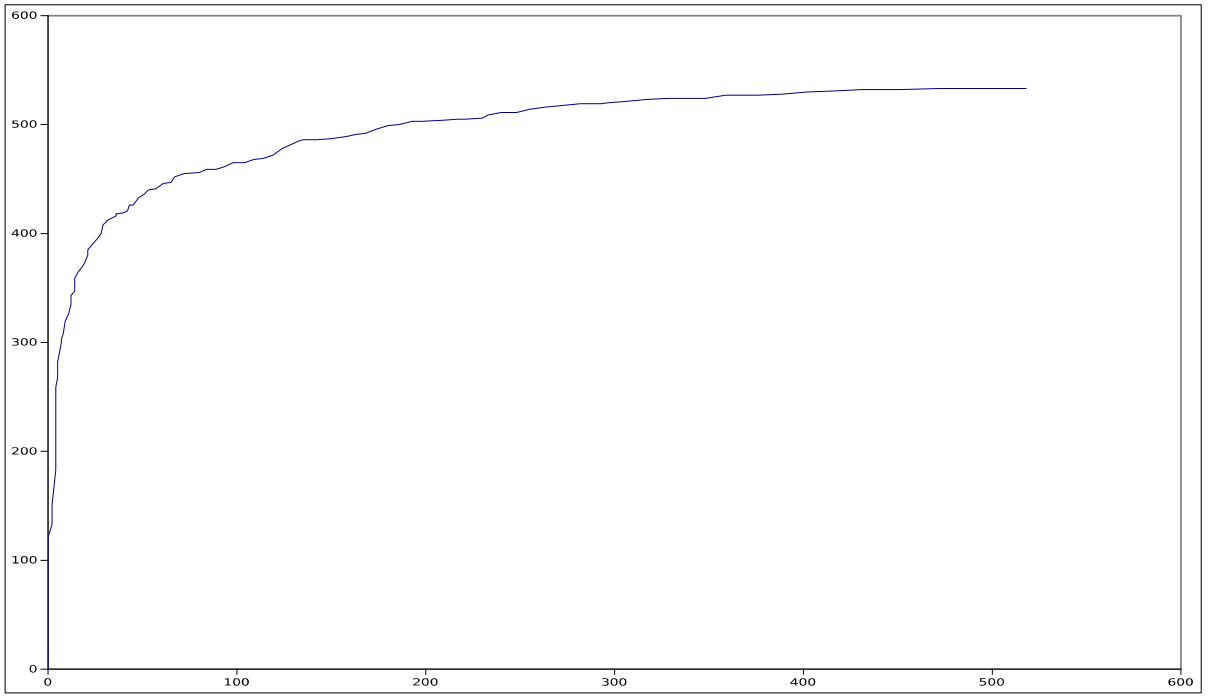
Thesis Meeting
09.06.2015
Sebastian Köpsel

Overview

- More Features
- Data stuff yesterday



Amount-of-features

- Dots
- Words
- NonstopWords
- Stopwords
- Syllables
- Senteces

Has-Features

- Questionmarks
- Exclamationmarks
- Numbers
- Symbols
- (Sensitve content)

Other Features

- Ratio Stopworte zu allen Worten
- User von dem der tweet ist
- Testing vs Wordlists
  - Self reference(I, mine, me etc.)
  - Reader reference(you, your etc.)
  - Affectwords
  - Animals
  - Abbreviations

Other Features

- Maincontent length of associated Webpage
- Type of media
- Mentioned Users
- Sentiment Feature - SentiWordNet 3 and StanfordCoreNlp
- Readability
- Max word length

Weka Results

- Procedure
  - All Features
  - StringToWordVector with word 3-gramms
  - X^2 to reduce number of attributes to 100
  - Balanced corpus

Results

	J48	Logistic Regression	Random Forrest	Naive Bayes
Accuracy in %	83.6	86.7	85.9	85.1
ROC Area Weighted Avg.	0.866	0.936	0.928	0.918