Technische Universiteit Eindhoven University of Technology

Rumour Finder

Goal and Scope

Goal

- Find out whether a sentence is a rumour or not.
- Compare different classifiers and preprocessing techniques.

Scope

- Limited to *Tweets* of the *Zika-virus.*
- Only considered supervised learning.

Neural Network

Architectural choices

- Sigmoid neurons
- Cross-entropy cost function
- Stochastic mini-batch gradient descent
- L2 regularization (weight decay)
- Gaussian weight initialization
- Single hidden layer
- Early stopping

Pre-processing

Converting tweets to features

- Counting Vector (CV)
- Binary Counting Vector (CV)
- TF-IDF

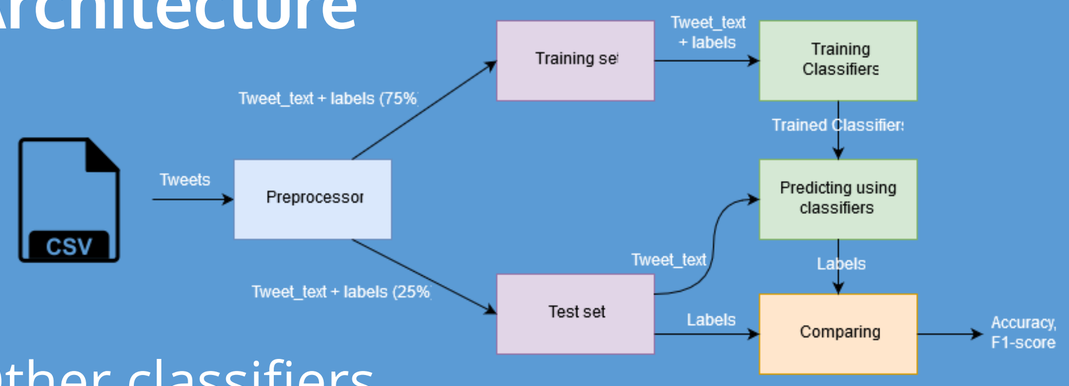
Optional pre-processing

- Removing stopwords

Other tasks

- Removes duplicate tweets from data.
- Splits data randomly into training-set (75%) and test-set (25%).

Architecture

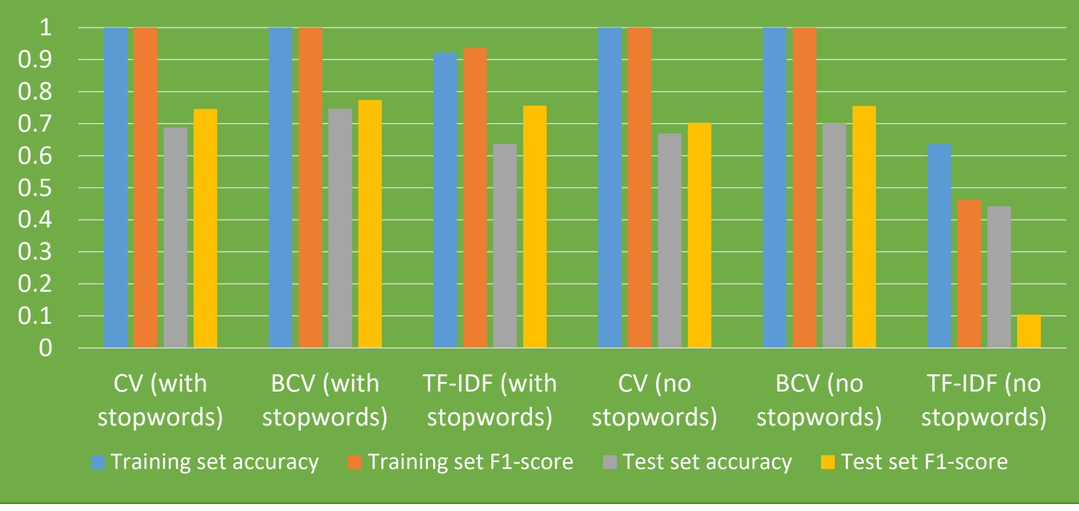


Other classifiers

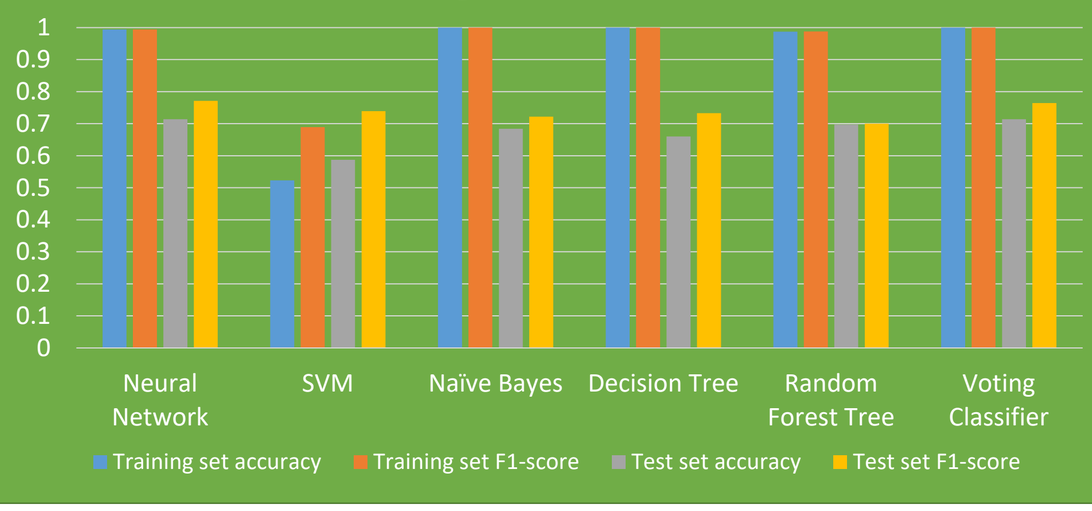
- Support Vector Machine (SVM)
- Naive Bayes classifier
- Decision Tree classifier
- Random Forest Tree classifier
- Voting classifier (hard) – Consisting of Naive Bayes, Decision Tree and Random Forest Tree

Results

Comparing pre-processing techniques



Comparing different classifiers



Discussion

Since the training score is much higher than the test score, all classifiers heavily overfit the data.

Reasonably high accuracy and F1-Score.

Limitations

- Only works with *Tweets* about the *Zika-virus*.
- No cross-validation hyper-parameter fitting has been done.
- Pre-processing techniques have only been tested with the Neural Network.

Conclusions

1. Removing stopwords reduces accuracy + score.
2. Neural Network and Voting classifiers give best results. Combination of them would be ideal.
3. Using the Binary Counting Vector results in the highest test score for the Neural Network.
4. The classifiers heavily overfit the data. Best way to solve this problem is with more training data. Therefore, it would be interesting to combine these results with unsupervised learning.

Department of Mathematics and Computer Science

Joost Visser (0828234) | 17th of January, 2017