SIGEVO Summer School,
10-12 July 2019 - Prague

# FAKENEWS

Lucia Cavallaro, Kathrin Kefer, Adam Johanides, Etor Arza, Evžen Šírek
supervised by Marc Schoenauer

# Initial thoughts on the topic “Fake News”

- The context:
  - **Textual** (as articles, tweets, etc.)
  - Images
  - Videos
- The approaches:
  - Network Science
  - **ML using classifiers**
- The target:
  - Containing the spread by the author
  - **Identify the fake news by the contents**

# Final Problem Formulation

- Main Goal: Try to discriminate Fake News from real ones
- Main Focus: Contents of articles
- Dataset: https://www.kaggle.com/c/fake-news/data
  - 20k articles

# General Classification Approach

1. Generate the features using the two different preprocessing approaches
    1.1. Topic Classification: apply best trained model to the real problem dataset
    1.2. doc2vec: transform dataset to vector representation

2. Apply the same 10 classifiers to the specified features
    2.1. Only Topic Classification Features
    2.2. Only doc2vec Features
    2.3. Both, Topic Classification and doc2vec features together

3. Evaluate the classifiers for their accuracy

# Topic Classification Preprocessing

- Data Basis for Training: Reuters-21578 News Article Dataset
  - 21578 instances of articles that are labelled with 135 categories (e.g. Business, Politics,..)

1. Train 10 different classifiers with partly different configurations [1]
   - LinearSVC, Decision Tree, Random Forest, kNearestNeighbour, SVM, Logistic Regression, Naive Bayes, AdaBoost, LDA
   - Features: each word with its number of appearances, that appears at least 3 times in the text and is no stop word like e.g. “and”, “it” etc → 26147 features in total

2. Choose best of the 10 classifiers and apply Hyperparameter Tuning to get the best combination of parameters
   - LinearSVC

# Topic Classification - Classifier Comparison

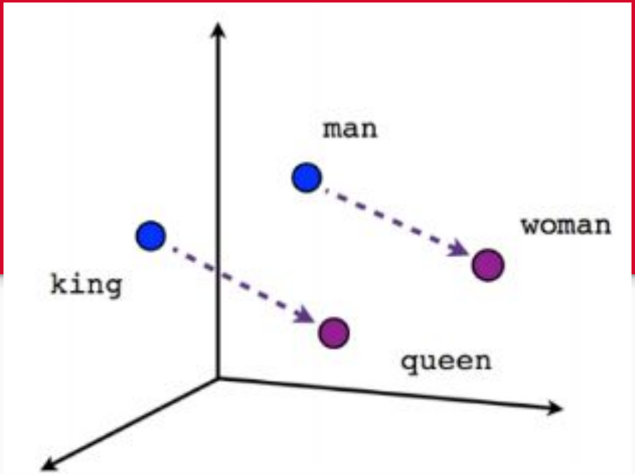
| | Accuracy [%] | F1 [%] |
|---|---|---|
| LinearSVC | 81.05 | 84.04 |
| Logistic Regression (C=1000) | 80.79 | 84.10 |
| kNN (n=5) | 72.97 | 76.07 |
| kNN (n=3) | 72.28 | 75.43 |
| Logistic Regression (C=1) | 67.47 | 67.21 |
| Random Forest (200 trees) | 65.75 | 64.36 |
| Random Forest (50 trees) | 64.79 | 63.69 |
| Decision Tree | 55.75 | 53.23 |
| Naive Bayes | 43.86 | 47.98 |
| SVM linear | 33.55 | 29.67 |

# Topic Classification - Hyperparameter Tuning

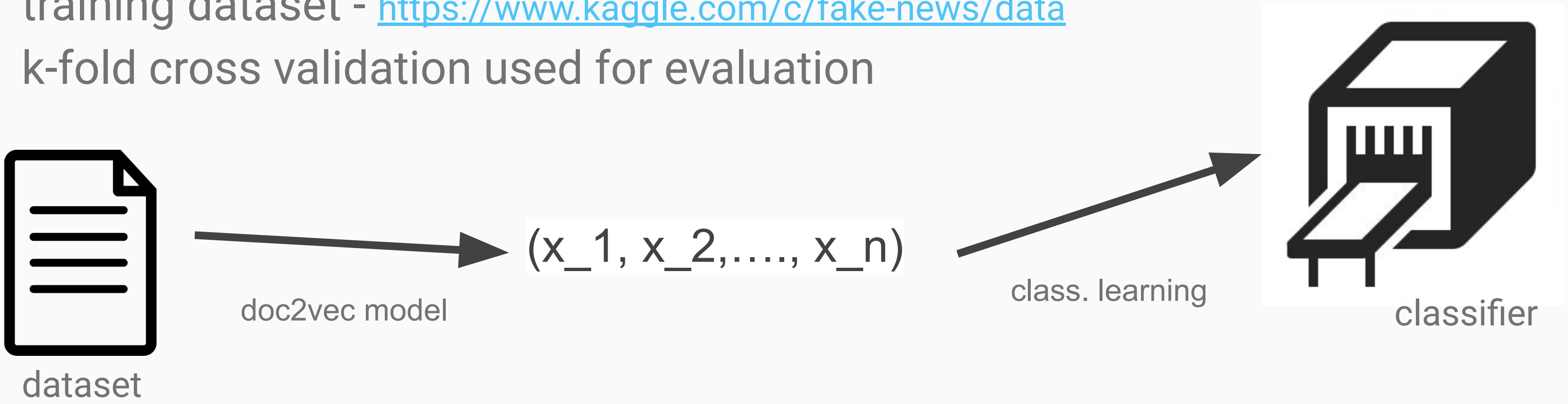
- Parameters to tune:
  - C = Penalty Parameter of the error term
    - [1,10,100,1000]
  - multi_class = determines the multi-class strategy of the LinearSVC classifier
    - "ovr" trains n_classes one-vs-rest classifiers
    - "crammer_singer" optimizes a joint objective over all classes

- Accuracy: 82.15%
- Best Parameters found:
  - C = 1
  - multi_class = crammer_singer

# Doc2Vec based classification



- using Doc2Vec (Mikilov and Le, 2014)
- based on Word2Vec
- training dataset - https://www.kaggle.com/c/fake-news/data
- k-fold cross validation used for evaluation



dataset

doc2vec model

(x_1, x_2,...., x_n)

class. learning

classifier

# Adapting the articles.

crisis prevention, and verifiable nuclear disarmament should be substituted for continuing counterproductive wars. Therefore, we, as signers of this petition, call for the immediate cancellation of the F-35 program as a whole, and the immediate cancellation of plans to base any such dangerous and noisy jets near populated areas.

# Adapting the articles.

crisis prevention, and verifiable nuclear disarmament should be substituted for continuing counterproductive wars. Therefore, we, as signers of this petition, call for the immediate cancellation of the F-35 program as a whole, and the immediate cancellation of plans to base any such dangerous and noisy jets near populated areas.

# Adapting the articles.

crisis prevention and verifiable nuclear disarmament should be substituted for continuing counterproductive wars therefore we as signers of this petition call for the immediate cancellation of the program as a whole and the immediate cancellation of plans to base any such dangerous and noisy jets near populated areas

# Adapting the articles.

crisis prevention verifiable nuclear disarmament should be substituted continuing counterproductive wars therefore we signers petition call immediate cancellation program whole immediate cancellation plans any such dangerous and noisy jets populated areas

# Adapting the articles.

crisis prevention verifiable nuclear disarmament should be substituted continuing counterproductive wars therefore we cancellation program whole cancellation plans any such dangerous noisy jets

# Adapting the articles.

```
['crisis', 'prevention', 'verifiable', 'nuclear', 'disarmament', 'should', 'be',
'substituted', 'continuing', 'counterproductive', 'wars', 'therefore', 'we',
'cancellation', 'program', 'whole', 'cancellation', 'plans', 'any', 'such', 'dangerous',
'noisy', 'jets', 'we', 'replacing', 'any', 'basing', 'any', 'locations', 'we', 'further',
'demand', 'money', 'human', 'needs', 'us', 'customer', 'world', 'including', 'climate',
'change', 'student', 'debt', 'education', 'healthcare', 'housing', 'add', 'your',
'swanson', 'is', 'an', 'author', 'journalist', 'host', 'he', 'is', 'director', 'coordinator',
'rootsactionorg', 'books', 'war', 'is', 'lie', 'he', 'blogs', 'at', 'he', 'hosts', 'nation', 'he', 'is',
'peace', 'prize', 'follow', 'him', 'on', 'twitter', 'support', 'clicking', 'here']
```

# doc2vec/preprocess - class. comparison

| | Accuracy [%] | F1 [%] |
|---|---|---|
| LinearSVC | 87.48 | 87.48 |
| Logistic Regression (C=1000) | 87.67 | 87.67 |
| kNN (n=5) | 81.60 | 81.39 |
| kNN (n=3) | 81.99 | 81.82 |
| Logistic Regression (C=1) | 87.72 | 87.72 |
| Random Forest (200 trees) | 87.02 | 87.01 |
| Random Forest (50 trees) | 87.07 | 87.06 |
| Decision Tree | 76.19 | 76.18 |
| SVM adjusted | 89.50 | 89.50 |
| SVM linear | 87.62 | 87.62 |

# doc2vec/no preprocess - class. comparison

| | Accuracy [%] | F1 [%] |
|---|---|---|
| LinearSVC | 74.93 | 74.89 |
| Logistic Regression (C=1000) | 74.81 | 74.79 |
| kNN (n=5) | 76.50 | 76.49 |
| kNN (n=3) | 75.82 | 75.81 |
| Logistic Regression (C=1) | 74.79 | 74.76 |
| Random Forest (200 trees) | 78.35 | 78.34 |
| Random Forest (50 trees) | 77.94 | 77.94 |
| Decision Tree | 71.13 | 71.12 |
| Naive Bayes | 59.45 | 57.33 |
| SVM linear | 74.79 | 74.72 |

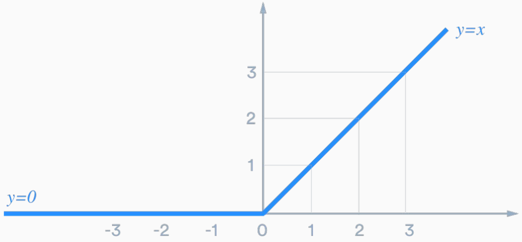
# Results - doc2vec/preprocess + deep learning

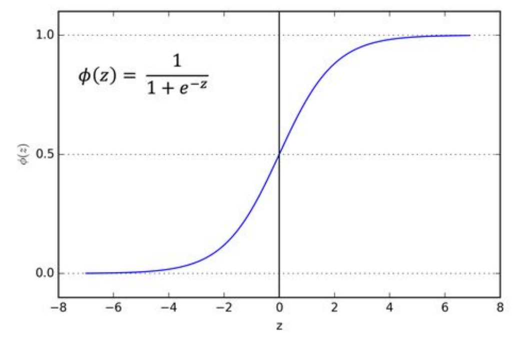
Accuracy on a training set: 97.45%

Accuracy on a test set: 89.23%

Activation function: ReLU, Sigmoid

Binary classification





# Overall Results - Preprocessing Approach Comparison

- Scripts are set up & ready.. but algorithms didn’t run through so far (they take some time…)

- → So no results available for comparing the preprocessing approaches

# Outlook

- Finish the final and overall comparison of the preprocessing approaches
- Set up some sort of framework to be able to continuously/regularly and automatically train the models for continuous improvement/adaptation (language is changing!)
- Apply the trained models in real life e.g. on some sort of news website or similar
- Take satyre into account

# Thank you!

# References

[1] https://martin-thoma.com/nlp-reuters/