Machine Learning Project Report: Fake News Detection

Martin Wagner, Philipp Hoffmann
Applied ML – Ludwig-Maximilians-Universität München
wagner.mar@campus.lmu.de, philipp.hoffmann@campus.lmu.de

July 14, 2025

1 Task Overview

Briefly describe the dataset or model you worked on, the goal of the project, and why the task is challenging. For example, challenges may include complex preprocessing, large dataset size, class imbalance, poor performance of baseline approaches, or the need for more complex models to achieve good results.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2 Methods

We compared the performance of logistic regression and linear svm models (mainly on the collection of titles) using different vectorization techniques from natural language processing. We used `TfidfVectorizer` and `CountVectorizer` from `scikit-learn` for basic text vectorization. Given a vocabularly of tokens (e.g. words), bag-of-words assigns to a document d for each token t, the term-frequency $\mathrm{tf}(t, d)$, defined as the count of term t in d divided by the total number of terms. A bit more evolved is Tf-Idf-vectorization, which also takes into account the frequency of how often a term occurs in a fixed corpus $D = \{d_1, \ldots, d_n\}$ of documents $d_1, \ldots, d_n$. Given a term t, the inverse-document frequency of t in D is $\mathrm{idf}(t, D) := 1 + \log\left(\frac{1+n}{1+\mathrm{df}(t)}\right)$, where $\mathrm{df}(t)$ is the number of documents in D containing the term D. Then the Tf-Idf-score of a term t with respect to the collection D is defined as

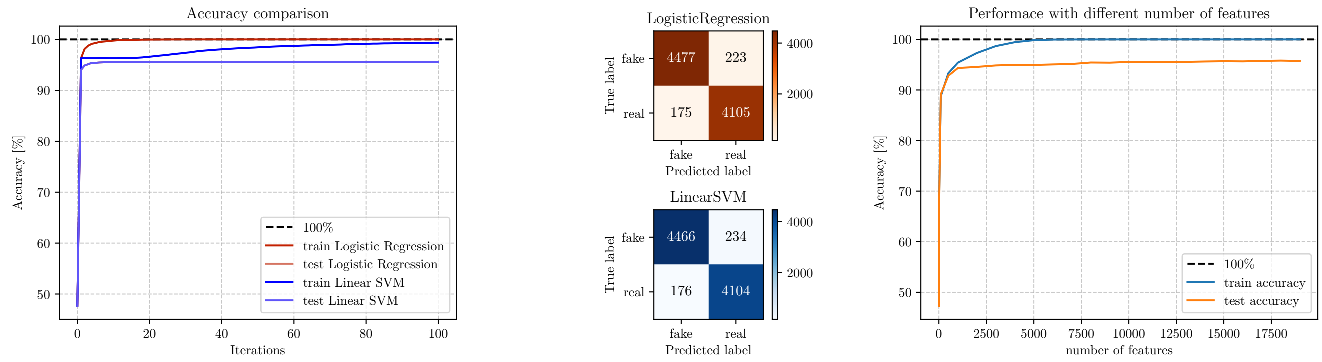
$$\text{tf-idf}(t, d, D) := \mathrm{tf}(t, d)\,\mathrm{idf}(t, D).$$

Note that d does not need to be contained in the collection D of documents itself. The tools from `scikit-learn` also apply $L2$-normalization to the obtained vectors, which we retained. Additionally, we applied z-score normalization afterwards, as this gave us better results, in particular for linear SVMs. We adopted models in courselib to support sparse matrix calculation and implemented methods to adopt for feature shifts (due to z-score normalization) without destroying the computational efficiency resulting from the sparsity of the unshifted feature matrices. Moreover we implemented custom tokenizers using `nltk` that support for example stemming or lemmatiztation, i.e. reduce words to their basic form. To test all of these options more compactly, we implemented a multi-column-vectorizer which supports vectorization of text data of one or multiple columns of a dataframe with different vectorization options.

3 Experiments and Results

Present your results. Use figures, tables, and metrics: Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus

tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.



(a) (b) (c)

Figure 1: figure

Model	batch size	Train Accuracy [%]	Test Accuracy [%]	Training Time [s]	Precision	Recall	F1-Score
LogisticRegression	full	100.0	95.77	0.63	0.95	0.96	0.96
LinearSVM	full	99.36	95.22	0.88	0.94	0.96	0.95
LogisticRegression	100	100.0	95.19	8.89	0.94	0.95	0.95
LinearSVM	100	96.38	91.37	23.0	0.91	0.91	0.91

Table 1: table

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

4 Discussion

Summarize key findings, insights, or issues. Optionally, suggest future work or limitations. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien.

Vectorization	Stop Words	Tokenizer	Features	Train Accuracy [%]	Test Accuracy [%]	Precision	Recall	F1-Score	Vectorization Time [s]
tf-idf	None	default	19639	100.0	95.73	0.95	0.96	0.96	0.55
tf-idf	None	basic	25010	100.0	96.79	0.97	0.97	0.97	5.08
tf-idf	None	stemming	17433	100.0	96.63	0.97	0.96	0.96	12.27
tf-idf	None	lemmatization	20975	100.0	96.69	0.96	0.97	0.97	38.76
tf-idf	english	default	19367	100.0	94.39	0.94	0.94	0.94	0.55
tf-idf	english	basic	24740	100.0	96.22	0.96	0.96	0.96	5.01
tf-idf	english	stemming	17287	100.0	95.76	0.96	0.95	0.96	10.67
tf-idf	english	lemmatization	20913	100.0	96.08	0.96	0.96	0.96	29.7
bag-of-words	None	default	19639	100.0	95.66	0.96	0.95	0.95	0.55
bag-of-words	None	basic	25010	100.0	96.94	0.97	0.97	0.97	5.09
bag-of-words	None	stemming	17433	100.0	96.75	0.97	0.96	0.97	11.96
bag-of-words	None	lemmatization	20975	100.0	96.71	0.97	0.96	0.97	33.67
bag-of-words	english	default	19367	100.0	94.54	0.94	0.94	0.94	0.54
bag-of-words	english	basic	24740	100.0	96.37	0.96	0.96	0.96	5.14
bag-of-words	english	stemming	17287	100.0	95.76	0.96	0.95	0.96	10.84
bag-of-words	english	lemmatization	20913	100.0	96.18	0.96	0.96	0.96	30.02

Table 2: table

n-gram range	Features	Train Accuracy [%]	Test Accuracy [%]	Vectorization Time [s]	Training Time [s]
1-1	19639	100.0	95.73	0.57	0.51
1-2	212673	100.0	95.67	1.55	1.14
1-3	497241	100.0	94.24	2.76	1.52
2-2	193034	100.0	92.42	1.19	0.7
3-3	284568	100.0	72.8	1.38	0.9

Table 3: table

Data	Features	Train Accuracy [%]	Test Accuracy [%]	Vectorization Time [s]	Training Time [s]
title	19639	100.0	95.73	0.56	0.42
text	112215	99.98	98.69	12.08	4.2
title+text	131854	100.0	99.0	11.97	4.33

Table 4: Table

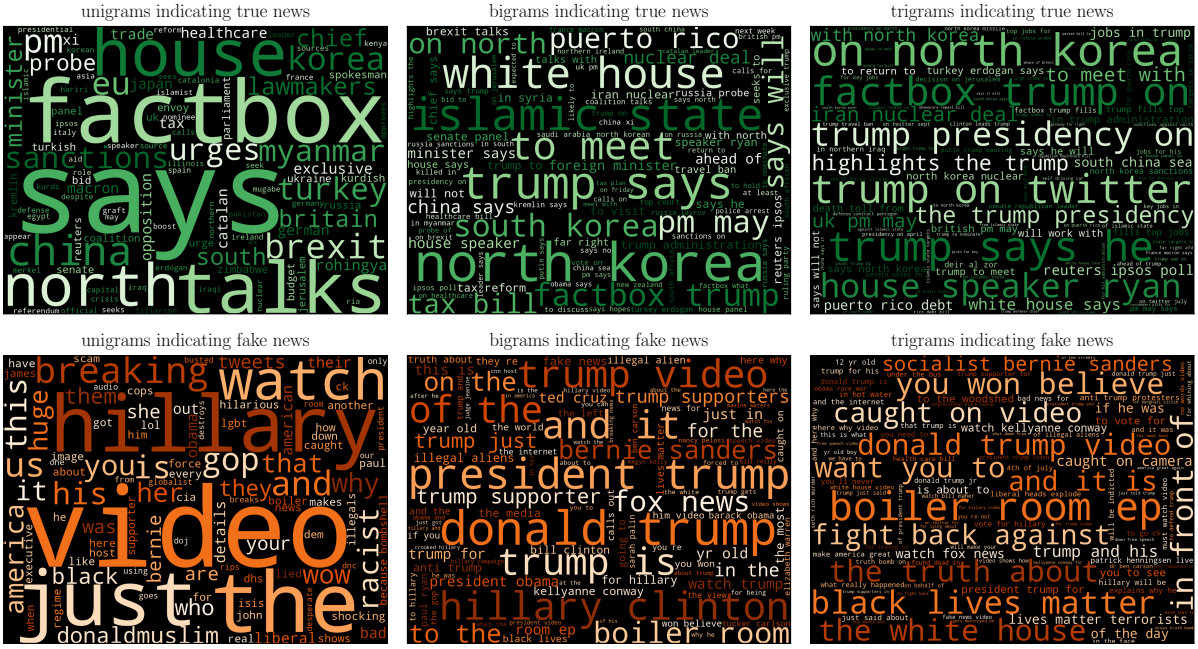


Figure 2: figure

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.