LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Fake News Analysis and Detection in German

Bachelor Thesis in Computational Linguistics at the LMU Munich
Faculty for Languages and Literatures; 5th Jan. 2021

Presented by Laurin B. Gerhardt
Supervised by Antonis Maronikolakis
Examined by Prof. Dr. Hinrich Schütze

Motivation

– Fake news are a far-reaching problem

→ Examples

– ‘fake news’ = news content that is factually wrong, often intentionally so

Problem: Expertise and time required

– Can’t keep up

→ Use machine learning to help

→ Most datasets for English

Analysis → Datasets

Fake News Dataset German (FNDatasetGerman)

→ ~62 thousand entries from 2008 to 2018

→ ~4.6 thousand of which are fake news

→ 3 sources for ‘real’ news, 4 for ‘fake’

(Stöckl, 2020)

GermanFakeNC

→ 490 articles (URLs, 65 not valid)

→ almost all contain false information

→ 33 valid sources, 12 invalids

(Vogel and Jiang, 2019)

Analysis → Results

	Legitimate		Fake	
	Title	Body	Title	Body
mean	7.000	395.54	11.05	317.520
std	2.006	206.428	3.632	198.061
25%	6	239	9	239
median	7	366	11	286
75%	8	511	13	344

Table 3.2.: Tokens per Text

	Legitimate		Fake	
	Title	Body	Title	Body
mean	6.614	5.615	5.894	5.08
std	4.350	4.280	4.175	3.805
25%	3	3	3	3
median	6	4	5	4
75%	9	8	8	7

Table 3.3.: Letters per Token

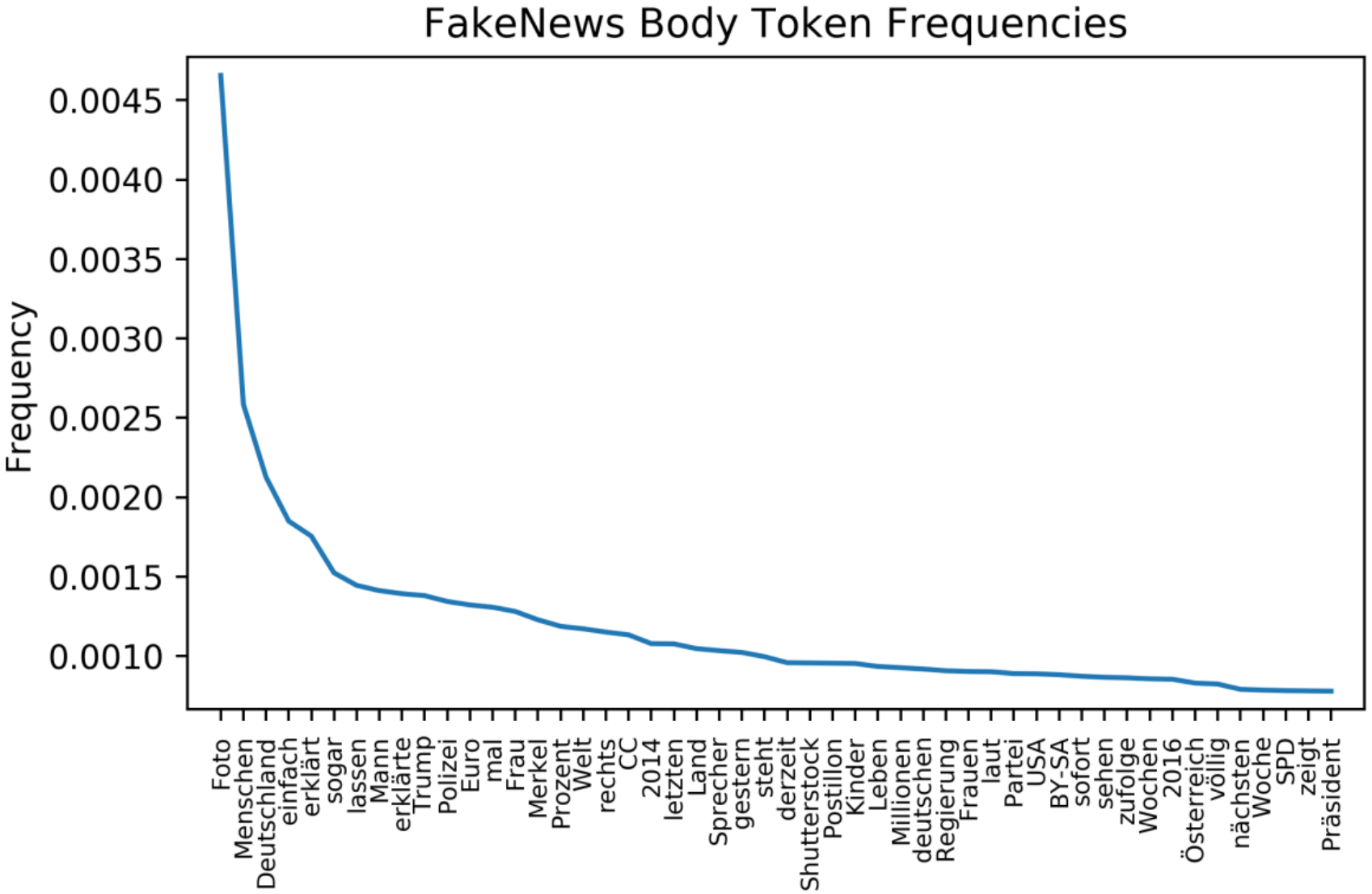
	Legitimate		Fake	
	Title	Body	Title	Body
mean	1.185	24.049	1.555	20.989
std	0.455	14.951	0.808	19.105
25%	1	14	1	14
median	1	21	1	18
75%	1	31	2	23

Table 3.4.: Sentences per Text

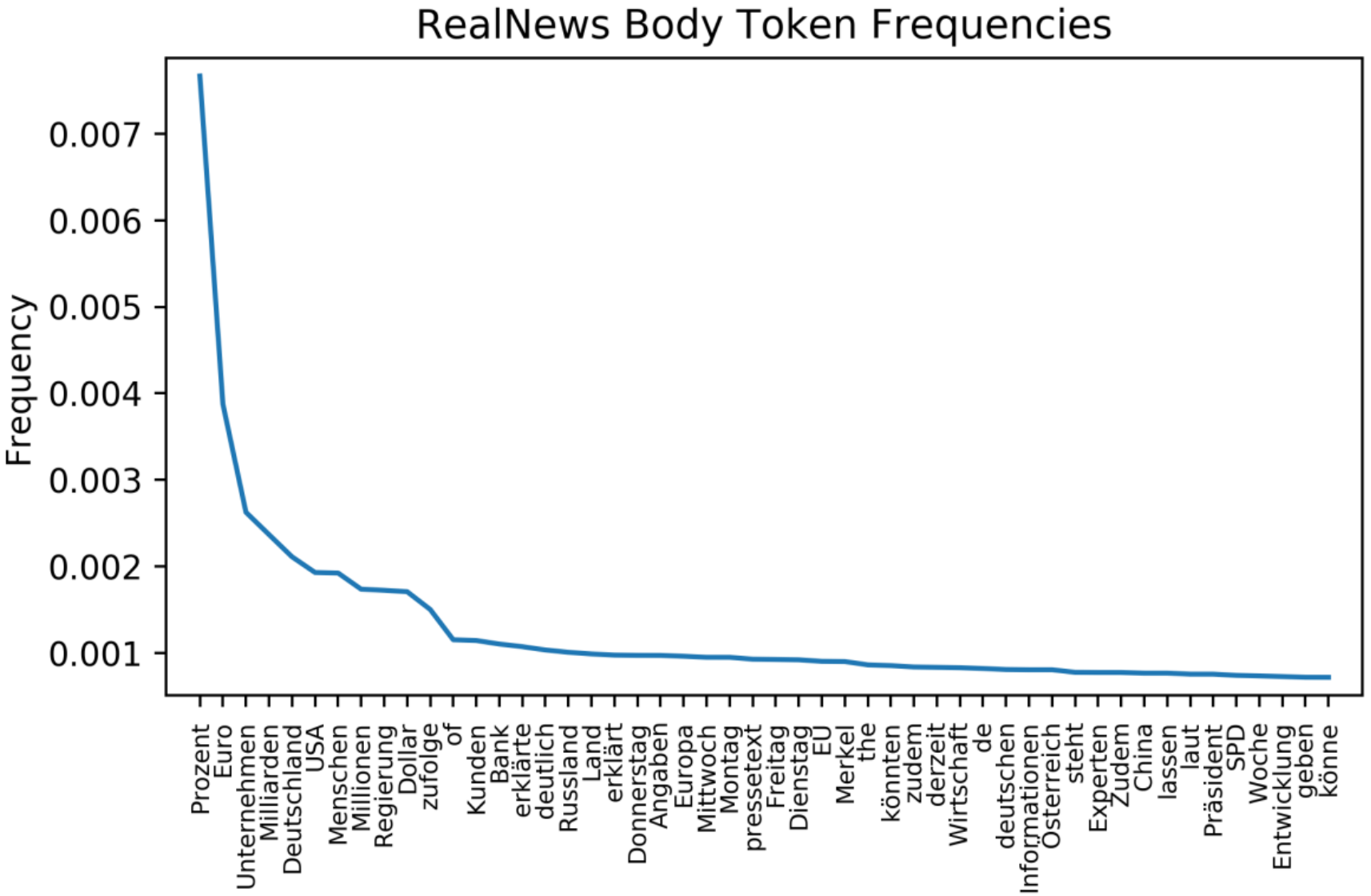
	Legitimate		Fake	
	Title	Body	Title	Body
mean	5.906	16.406	7.106	15.125
std	2.384	9.930	3.884	11.289
25%	4	10	4	7
median	6	16	7	13
75%	7	22	10	21

Table 3.5.: Tokens per Sentence

Analysis → Results



Analysis → Results



Analysis → Results

Article Bodies	our findings
Fake news articles are shorter	agree
Fake news uses fewer technical words	unclear
Fake news contains fewer quotes	agree
Fake news contain fewer nouns	agree
Fake news contain more adverbs	agree
Fake news (FN) and real news contain equally many determiners	FN has fewer
Titles	
Fake news titles are longer	agree
Fake news titles contain more proper nouns	agree
Fake news titles contain fewer nouns	agree
Fake news titles contain fewer determiners	opposite

(Horne and Adalı, 2017)

Detection → Methodology

– Test set size = 33.3%

– Classes: True (= fake news), False (=real news)

– Classifiers: Logistic Regression, Random Forests, Complement Naive Bayes

– Text Representations:

binary BoW, wordcounts, tf.idf

→ 9 experiments in two settings each:

‘default’ (default parameters),

‘optimized’ (inverse class weighting, stopword removal)

Detection → Results

Classifier	Text Representation	F1 default	F1 optimized
Complement Naive Bayes	binary BoW	0.95	0.95
	word-counts	0.96	0.96
	tf.idf	0.92	0.92
Logistic Regression	binary BoW	**0.99**	**0.99**
	word-counts	**0.99**	0.98
	tf.idf	0.97	0.96
Random Forest	binary BoW	0.93	0.95
	word-counts	0.93	0.95
	tf.idf	0.94	0.96

Table 4.1.: (minor) F1-scores for default and optimized parameters respectively

Detection → Results

Classifier	Text Representation	default		optimized	
		Legitimate	Fake	Legitimate	Fake
Complement Naive Bayes	binary BoW	0.99	0.54	0.99	0.58
	word-counts	0.98	**0.68**	0.98	**0.70**
	tf.idf	1.00	<.01	1.00	<.01
Logistic Regression	binary BoW	0.99	**0.92**	0.99	**0.96**
	word-counts	0.99	**0.92**	0.99	0.95
	tf.idf	0.99	0.73	0.96	0.94
Random Forest	binary BoW	1.00	0.22	0.99	0.47
	word-counts	0.99	0.22	0.99	0.49
	tf.idf	0.99	**0.28**	0.99	**0.52**

Table 4.2.: Accuracy-Scores for each class for default and optimized parameters respectively

Detection → Results

		Accuracy	
	F1	Legitimate	Fake
default	0.94	0.99	0.40
optimized	0.96	0.93	0.66

Table 4.3.: Performance of the Logistic Regression Classifier using the binary BoW text-representation measured in (minor) F1-score and Accuracy score of each class

Concluding Remarks

– Substantial differences in human written text

→ Differences are similar in English and German

– Machine learning can help

Limitations

– Classification may rely on style and wording
(see Schuster et al.,2020)

– Dataset(s) used are imbalanced

References

– Horne, B. D. and Adalı, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398.

– Schuster, T., Schuster, R., Shah, D. J., and Barzilay, R. (2020). The Limitations of Stylometry for Detecting Machine-Generated Fake News. Computational Linguistics, pages 1–12.

– Stöckl, A. (2020). Fake News Dataset German, Version 1. https://www.kaggle.com/astoeckl/fake-news-dataset-german. Last Access 5. Dec. 2020.

– Vogel, I. and Jiang, P. (2019). Fake News Detection with the New German Dataset ”GermanFakeNC”. In Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings, pages 288–295.