h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Bridging the Syn-to-Real Gap in Microorganism Detection Using Blended Synthetic Data

Sebastian Jörz, Stefan Höreth, Antonio Jorba, Eva Brucherseifer

Darmstadt University of Applied Sciences, Computer Science

Initial Situation

Hate Speech Detection using NLP

The *DeTox* dataset, which contains tweets classified as Hate Speech (0/1), serves as the basis.

- **Number of Datasets:** 9251
- **Labels:** 0 (Not Hate Speech), 1 (Hate Speech)
- **Class Imbalance:**
  - Class 0 (Not Hate Speech): approx. 71.6%
  - Class 1 (Hate Speech): approx. 28.4%

The Dataset

Validation Dataset: 80% Training, 20% Validation. The F1-Scores refer to the validation dataset.

Text	Label
Wenn man 4 AFD'ler braucht um eine Banane zu schälen, 5, um eine Glühbirne in eine Fassung zu drehen, Ähm, da muss man leider die Querdenker und Herr Liefers mit dazu nehmen. Dann könnte es gerade so reichen. https://t.co/teI4b6ooUR	0
@RockInTheBrand @ABaerbock Stimmt. Er ist zu intelligent!	0
Nazis mit Teenagern zu symbolisieren ist GENAU DAS RICHTIGE. Denn das SIND SIE. Nicht das Teenager dumm wären (man muss hier ja manchmal für Idioten sowas noch extra erklären), sondern Teenager, bei denen es im Hirn nicht mehr VORWÄRTS (in Richtung Reife eines Erwachsenen)...'	1

Tabelle 1: Example data from the dataset

Data Preprocessing

Goal: Improve text quality and reduce noise to increase model performance.

- Convert to lowercase
- Remove URLs and mentions (e.g., '@username')
- Remove hashtag symbols (e.g., '#'), but retain the text
- **Stemming**:
  - Reduce words to their root form
  - e.g., 'universal', 'university' and 'universe' to 'univers'
  - Advantage: Fast processing
  - Disadvantage: Loss of information and meaning

Text	Label
wenn man 4 afd'ler braucht um ein banan zu schal , 5 , um ein gluhbirn in ein fassung zu dreh , ahm , da muss man leid die querdenk und herr lief mit dazu nehm . dann konnt es gerad so reich .	0
stimmt . er ist zu intelligent !	0
nazis mit teenag zu symbolisi ist genau das richtig . denn das sind sie . nicht das teenag dumm war (man muss hier ja manchmal fur idiot sowas noch extra erklar) , sond teenag , bei den es im hirn nicht mehr vorwart (in richtung " reif ein erwachs ") ...	1

Tabelle 2: Example data after preprocessing with Stemming

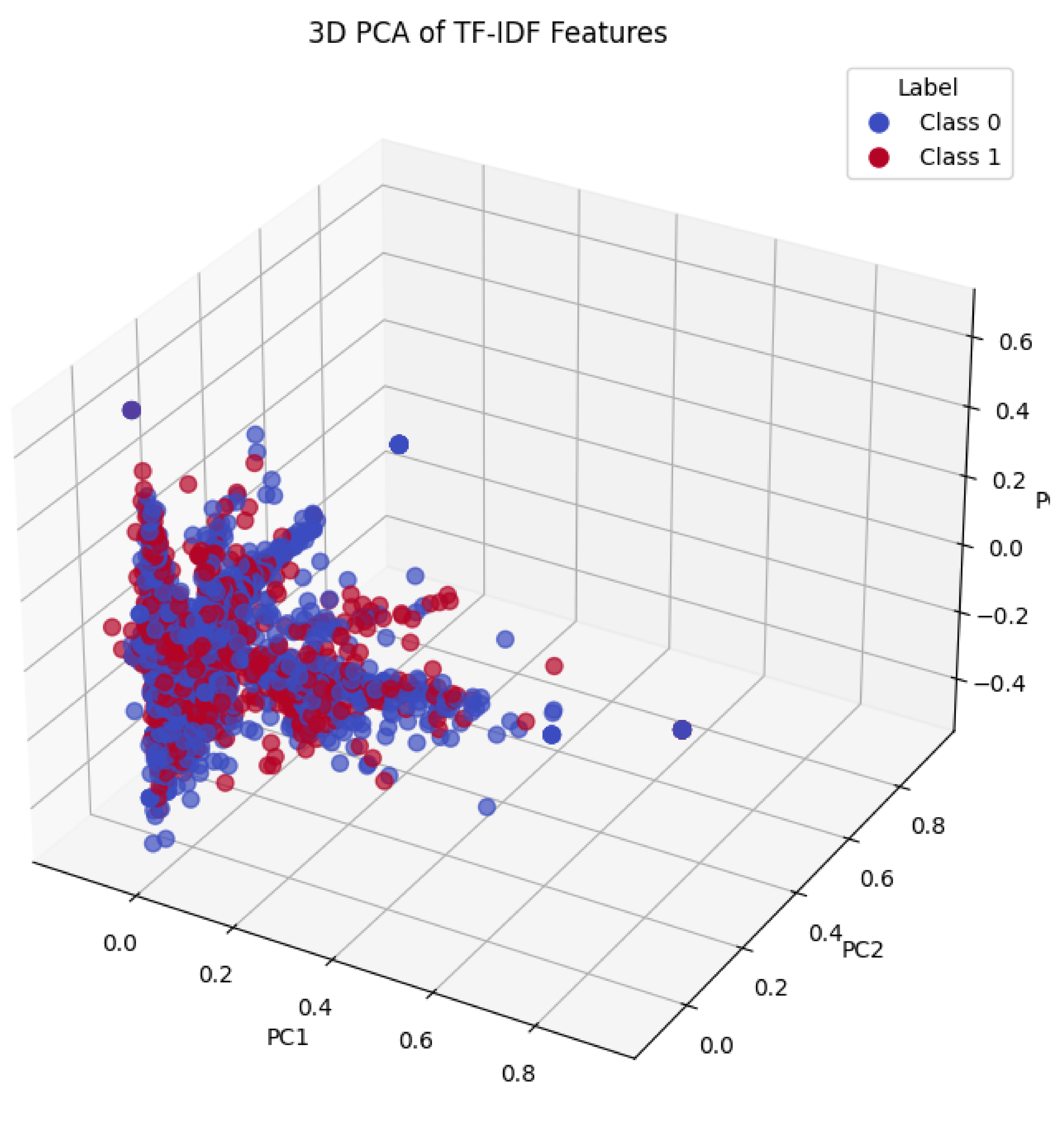
- **Lemmatization**:
  - Reduce words to their base form (lemma)
  - e.g., 'better' to 'good', 'running' to 'run'
  - Advantage: Better preservation of meaning and context
  - Disadvantage: More time-consuming than stemming

Text	Label
4 Afd 'ler brauchen Banane schälen 5 Glühbirne Fassung drehen ähm Querdenker Herr Liefers nehmen reichen	0
stimmen intelligent	0
nazis Teenager symbolisieren genau richtig Teenager dumm sein manchmal Idiot sowas extra erklären Teenager Hirn vorwärts Richtung reif erwachsen	1

Tabelle 3: Example data after preprocessing with Lemmatization

First Analysis

- **TF-IDF:** Term Frequency-Inverse Document Frequency for word weighting
- **Visualization:**
  - PCA (Principal Component Analysis) for dimensionality reduction
  - Scatterplot of the 3 PCA components



- **Results:**
  - PCA shows no clear separation between classes.
  - The data are highly overlapping, indicating high classification complexity.

First 'Classic' Models

Traditional Machine Learning models were tested as a basis

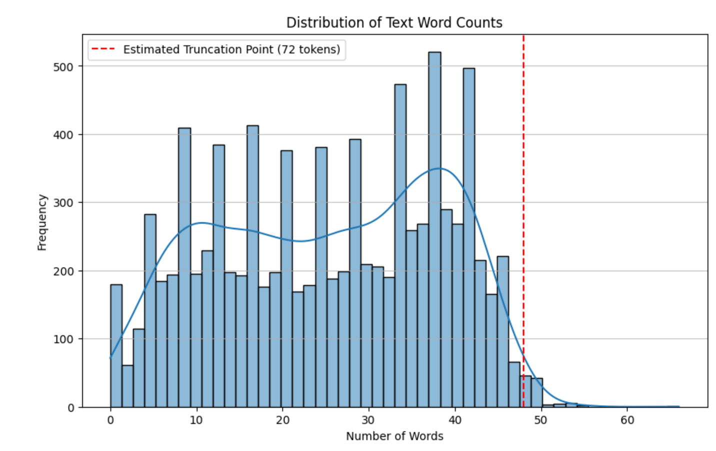
- **Vectorization Methods:**
  - TF-IDF (with German stopwords)
  - FastText Embeddings (with 'word2vec-google-news-300')
- **SMOTE** (Synthetic Minority Over-sampling Technique): For handling class imbalance.
- **Classification Models:**
  - Random Forest (F1-Score up to: 0.35)
  - XGBoost (F1-Score up to 0.44)
- **Experiments:**
  - Different preprocessing (with/without Stemming, Lemmatization, etc.)
  - Different tokenizer hyperparameters
  - Hyperparameter optimization of the models using 'GridSearchCV'

Class	Precision	Recall	F1-Score	Support
0	0.74	0.87	0.80	1325
1	0.43	0.24	0.31	526
Accuracy			0.69	1851
Macro Avg	0.58	0.56	0.56	1851
Weighted Avg	0.65	0.69	0.66	1851

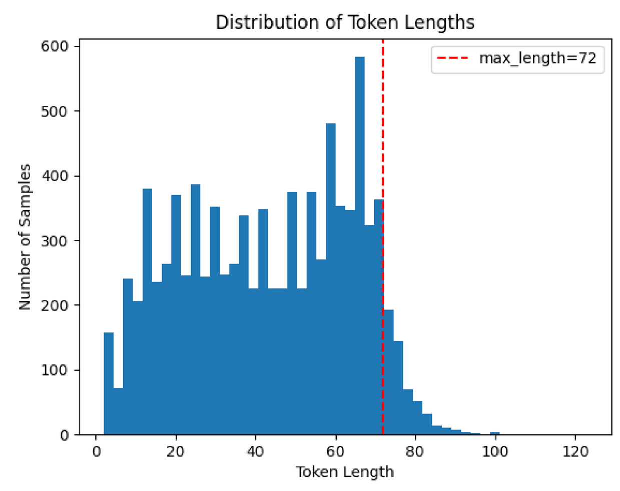
Tabelle 4: Classification Report of a Random Forest (F1-Score=0.30 for Class 1) on the validation dataset.

First Approach with BERT

- **Model:** Based on 'bert-base-german-cased'
- **AutoModelForSequenceClassification:** Model with a head for classification.
- **Tokenization:**
  - 'AutoTokenizer' with 'bert-base-german-cased'.
  - 99% Quantile of word lengths: 48 words.
  - 95% Quantile of token lengths: 73 tokens.
  - Exploratory analysis with different token lengths
  - Best result with 72 tokens => 5.77% of the data is truncated.



Distribution of word lengths in tweets



Distribution of token lengths in tweets

- **Class Imbalance Handling:**
  - Inverse frequency weights applied to weight the minority class.
  - weight_0 = num_samples / (2 * label_counts[0])
  - weight_1 = num_samples / (2 * label_counts[1])
- **Evaluation Metric for Best Model:** F1-Score, due to class imbalance.
- **Regularization:** Early Stopping
- **Best F1-Score:** 0.64

BERT Transformer & Embeddings with XGBoost

- **Model:** 'deepset/gbert-large'
- **Tokenization:** 'AutoTokenizer' with 72 tokens
- **Embeddings:**
  - 'AutoModel' with 'deepset/gbert-large'
  - Extraction of the last Hidden States as Embeddings
- **XGBoost Classifier:**
  - Use of extracted Embeddings as input
  - Hyperparameter optimization with 'GridSearchCV'
  - Class imbalance handling using 'scale_pos_weight'
- **F1-Score:** 0.58

Class	Precision	Recall	F1-Score	Support
0	0.82	0.88	0.85	1308
1	0.65	0.52	0.58	543
Accuracy			0.78	1851
Macro Avg	0.73	0.70	0.71	1851
Weighted Avg	0.77	0.78	0.77	1851

Tabelle 5: Classification Report of the XGBoost Model with BERT Embeddings on the Validation Dataset

Evaluation

Evaluated on a separate test dataset with 1,018 data.

Results (macro-F1-Score):

- **ChatGPT:** 0.20
- **RF:** 0.44 - 0.55
- **XGBoost:** 0.40 - 0.63
- **BERT:** up to 0.68
- **BERT + XGBoost:** up to 0.72

Conclusion and Outlook

- Best F1-Score on own validation dataset: 0.64 with BERT.
- Best F1-Score on test dataset: 0.72 with BERT-Embeddings and XGBoost.
- **Hate Speech Detection** is a complex task requiring careful data preprocessing and model selection.