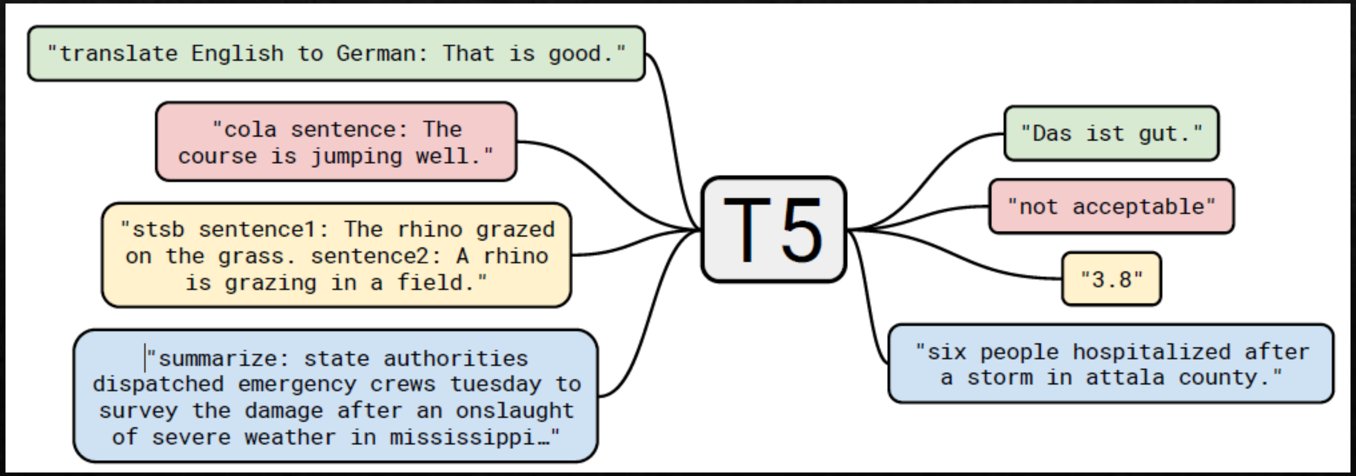
Moderat: Language Models for Fair and Explainable German Comment Moderation.

Isadora White

Background on T5 Model

- Developed by Google to be ultimate transformer model for transfer learning
- Text-to-Text model which can be used for many different tasks from translation to classification



T5 Training and Architecture

- Trained on C4 dataset: Colossal Clean Crawled Corpus
  - A massive English dataset with inappropriate words filtered out (aka swear words)
- Follows classic Encoder-Decoder Architecture (whereas BERT is an encoder-only model)
- Experimented with taking away the decoder portion of the T5 model as well

Accuracy Scores on the RP Datasets

Model Name	RP-Crowd-3	RP-Crowd-2	RP-Mod
bert-base-german-cased	0.8381	0.8027	**0.7377**
XLM-Roberta-base	0.8135	0.8044	0.7199
GermanT5/t5-efficient-oscar-german-small-el32	**0.8476**	0.8137	0.7367
GermanT5/german-t5-oscar-ep1-prompted-germanquad	0.8214	**0.8338**	0.7216
Google/mt5-small	0.7881	0.7756	0.7003
Google/mt5-base	0.8087	0.7938	0.7174
Encoder-T5	**0.8476**	0.809	0.7346

Results & Takeaways

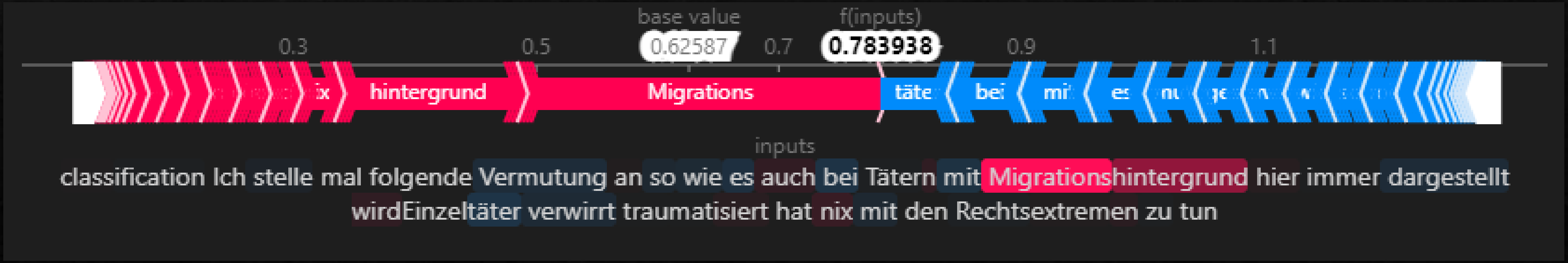
- Single language (German) models out-performed multi-lingual language models
- Encoder-decoder architecture outperformed encoder-only architecture
- Larger models outperformed smaller ones
- Models perform better on Crowd-Worker datasets (RP-Crowd) than on Moderated Datasets (RP-Mod)

SHAP Values Overview

- Explanation method for ML algorithms
- Attribution: assigns each word in the comment a score
- The scores sum up to the probability that the comment is problematic
- Technique derived from game theory's Shapley values

Explanations & Patterns of False Positives

- Used SHAP Values to find words which contributed the most to false positive classifications
- Found top 200 words which contributed to the false positive predictions
- Top 20: `[' Migranten',' Nazi‘, 'arme‘, 'Muslim','monster‘, 'ackt',' Psychiater','sy',' Juden','gewalt','flüchtling‘]`



Underrepresented Words

- Harmless words are contributing greatly to problematic examples
- Negative examples are underrepresented in the dataset

Word	Positive (hate) Examples	Negative Examples
Migranten	126	30
Arme	37	36
Muslim	59	25
Psychiater	10	1

Resolving Issue with False Positives

- Created a new validation dataset with the false positive words where 50% were positive and 50% were negative
- Resampled the dataset so that for each of the words which contributed to the false positives had an equal number of positive and negative examples
- Retrained the model on the new resampled dataset
- Accuracy on the new validation dataset increased from 50% to 66%

Intro to HASOC Competition

- Hate speech classification competition associated with the FIRE conference
- Includes tasks for English, Hindi, Marathi, and German
- Dataset consists of Tweets
- Test dataset becomes available on August 10th
- Registration deadline is August 13th

Task 2B: ICHCL GERMAN Codemix Binary Classification.

A task focused on hate speech and offensive language identification is offered for German. It is a coarse-grained binary classification in which participants are required to classify tweets into two classes, namely: hate and offensive (HOF) and non- hate and offensive (NOT).

- **(NOT) Non Hate-Offensive -** This post does not contain any Hate speech, profane, offensive content.
- **(HOF) Hate and Offensive -** This post contains Hate, offensive, and profane content.

Ideas for Future Work on HASOC Dataset

- Verify the results from the RP datasets on HASOC dataset
- Increase performance through data augmentation
  - Using emojis as features
  - Creating new comments by replacing words with their synonyms
- Cross-validate models trained on the different datasets