Improving Emotion Detection Through Translation of Text to ML Models Trained in Different Languages

MIDDLE TENNESSEE STATE UNIVERSITY

Qualifying Exam – Computational Data & Sciences

By Richard Hoehn

MTSU – August 2023

Hello

Richard Hoehn

•

Living Franklin

•

MTSU Grad 2005, Vanderbilt Grad 2009

•

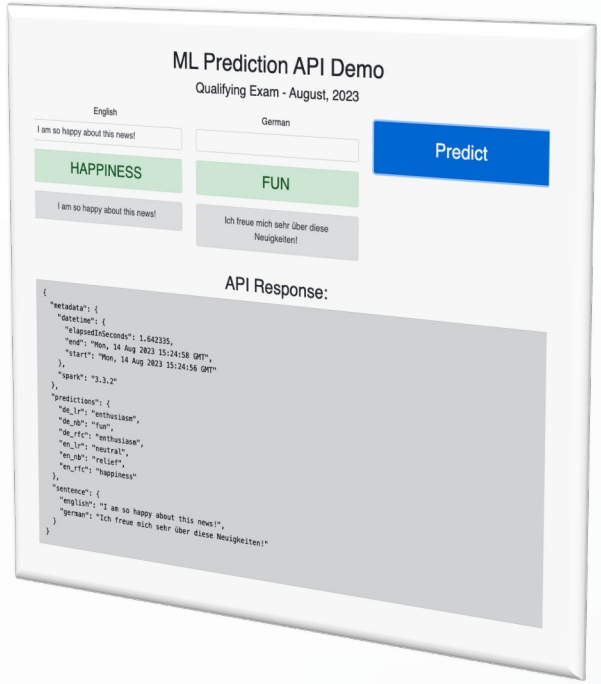
Last 15 years in SW development for Retail & Logistics Applications

•

Primarily in Databases, API / PubSub Integration Work, and Business Reporting Apps

Introduction & Agenda

- **Introduction**
  - Research on Emotion Detection (ED) in Text and improving Prediction Rates
  - Specifically by extending Data through Translation to Different Languages
  - Training Multiple ML Models on Original & Extended Data to Compare Prediction Rates
  - and Finally in Real-Time Translate Text to process in Parallel for further expansion
- **Agenda**
  - Significance of Emotion Detection
  - Challenges, Motivation and Scope of Research
  - Methodology – Including Code Review and Demonstration!
  - and finally Analysis of Results, Conclusion, and Future Work



Significance of Emotion Detection

- Emotions can vary depending on the theoretical framework or model being considered. One well-known model is the Plutchik's Wheel of Emotions, which proposes eight (8) primary emotions.
  - Joy / Happiness
  - Sadness
  - Anger
  - Worry / Fear
  - Surprise
  - Disgust / Hate
  - Trust / Love
  - Enthusiasm

- By accurately identifying and understanding emotions from text data, ML applications can assist in improving:
  - User Experiences (chat-bots),
  - Decision-Making Processes
  - and Overall human-machine interactions in a positive manner[4, 5], with most of these interactions being processed in real-time.
- ED is still a growing field in Text, Video, and Image reading. The market for ED software and services is estimated to reach **$3.8billion[2] by 2025.**

Challenges:
Data Scarcity & Language Fragmentation

- Emotion detection data requires primarily supervised learning data!
- Unlike Sentiment Analysis (SA) the availability of large datasets for training purposes of ML models is much smaller[3].
- Many datasets that are available are in many cases in multiple languages - not all are in English since emotions that are linked to text are contextual in nature.

Motivation for Research and Evaluation of Dataset Extending Impact

- ED is still a growing field in Text, Video, and Image reading. The market for ED software and services is estimated to reach $3.8billion[2] by 2025.
- ED spans most all domains such like psychology,
- By use of ML models the analysis of human emotions at scale, providing valuable insights into individual and collective emotional states both in real-time but also for measuring sentiments from the past versus the current time.
- In Chowanda et al. paper they believe that ”*Emotions hold a paramount role in the conversation, as it expresses context to the conversation*.”, this means that emotions are a part of a conversation and with that are needed to ensure valid analysis of a conversation.

Objectives & Scope of Research

The research project's objectives were three-fold:

- The first is to translate English data (feature & label) to German in order to extend the original German dataset for ML training purposes. Will the added text lead to better predictions?
- Similar to the first, can by translating German data to English and extending an original English dataset increase the predictability of English ML model?
- And Lastly shifting the focus to real-time translation and its impact on prediction. Can by translating in real-time an input to multiple languages improve the predictability based on the combined output of two models.

In summary, this research project investigated innovative ways to enhance the predictability of Emotion Detection models in both English and German. With these three (3) objectives from above the scope was to **P**rocure, **T**ranslate, **T**rain, and **E**valuate benefits of extending datasets by translation to improve emotion detection.

Literature Review

- Our research considered publication only past 2015, thereby providing an up-to-date perspective on ED analysis.

”Emotions hold a paramount role in the conversation, as it expresses context to the conversation.” **[1]**

- This means that emotions are a part of a conversation and with that are needed to ensure valid analysis of a conversation
- Emotions can differ across **A**ge groups, **G**enders, **C**ultures, and **L**anguages[6]
- Fragmentation caused by different languages further exacerbates the issue, as it reduces the size and diversity of data available for training, resulting in limited cross-lingual generalization and potentially biased models.[7]

Methodology

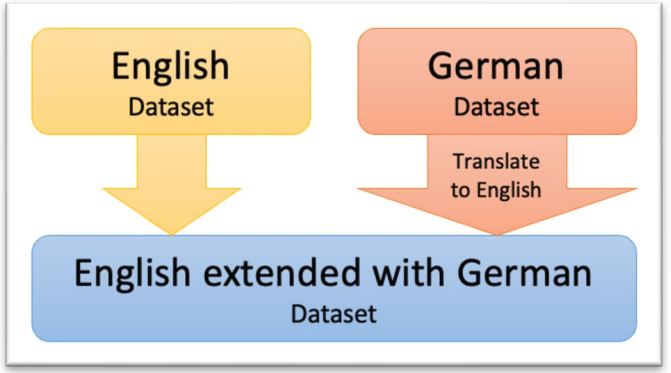
Data Procurement

Parsing & Cleanup

Dataset Translation & Extending

ML Training & Testing

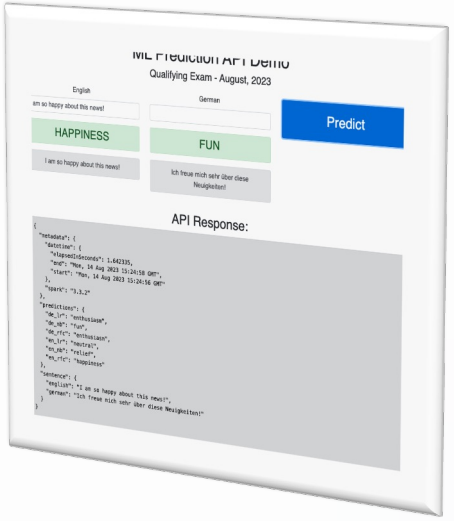
API for Real-Time Translation & Prediction



File Details

Name	Row Count	Type
English	38,000	CSV
German	2,500	JSON

```python
# Using Deep Translator to leverage Google Translate
# Link: https://cloud.google.com/translate/docs/reference/libraries/v2/python
from deep_translator import GoogleTranslator

sentence = 'Chocolate milk is so much better through a straw.'

translated = GoogleTranslator(source='auto', target='de').translate(sentence)
print(translated) # Schokoladenmilch schmeckt durch einen Strohhalm viel besser.

translated_back = GoogleTranslator(source='auto', target='en').translate(translated)
print(translated_back) # Chocolate milk tastes much better through a straw.
```



MIDDLE TENNESSEE STATE UNIVERSITY

Methodology

Data Procurement

- The data procurement was relatively straight forward and once found by using multiple Google search terms for the Emotion Detection in English and German text languages.
- The German data was obtained from the dataset built by ETH's Emotion and Stance Detection for German Text[4].
- The English dataset was downloaded from Kaggle[8] based on Tweets collected in 2021.
- Unfortunately there was a large quantity difference between German and English!

File Details

Name	Row Count	Type
English	38,000	CSV
German	2,500	JSON

Emotion Datasets Labels

English		German		Used
Name	Count	Name	Count	
Boredom	179	—	—	NO
Love	3842	Vertrauen	316	YES
Relief	1526	—	—	NO
Fun	1776	—	—	NO
Hate	1323	Ekel	29	YES
Neutral	8638	Unklar	314	YES
Anger	110	Ärger	226	YES
Happiness	5209	Freude	140	YES
Surprise	2187	Überraschung	369	YES
Sadness	5165	Traurigkeit	184	YES
Worry	8459	Angst	154	YES
Enthusiasm	759	Antizipation	774	YES
Empty	827	—	—	NO

MIDDLE TENNESSEE STATE UNIVERSITY

Methodology

Parsing, Cleanup, & Emotion Linkage

- By use of Jupyter Notebooks the English (csv) and German (JSON) files were.
- For processing I opted to use Pandas.
- Google Translator was use for Translation

```
# Emotion Panda DataFrame
# This was predetermined by review of the emotion from English to German
emotion_key = {
"boredom": "---",
"love": "Vertrauen",
"relief": "---",
"fun": "---",
"hate": "Ekel",
"neutral": "Unklar",
"anger": "Ärger",
"happiness": "Freude",
"surprise": "Überraschung",
"sadness": "Traurigkeit",
"worry": "Angst",
"enthusiasm": "Antizipation",
"empty ": "---",
"---": "Keine"
}

# Create a DataFrame from the emotion_key dictionary
df_emotions = pd.DataFrame(emotion_key.items(), columns=['emotion_en', 'emotion_de'])
```

```
# This google API take a Sentence and converts to German
def translate(sentence, dest_lang):
try:
translator = Translator()
translator.raise_Exception = True
translation = translator.translate(sentence, dest=dest_lang)
time.sleep(0.5) # Add a delay (This is due to rate limit of 1/s)
return translation.text
except Exception as e:
print(f"Translation Error: {e}")
return None
```

MIDDLE TENNESSEE STATE UNIVERSITY

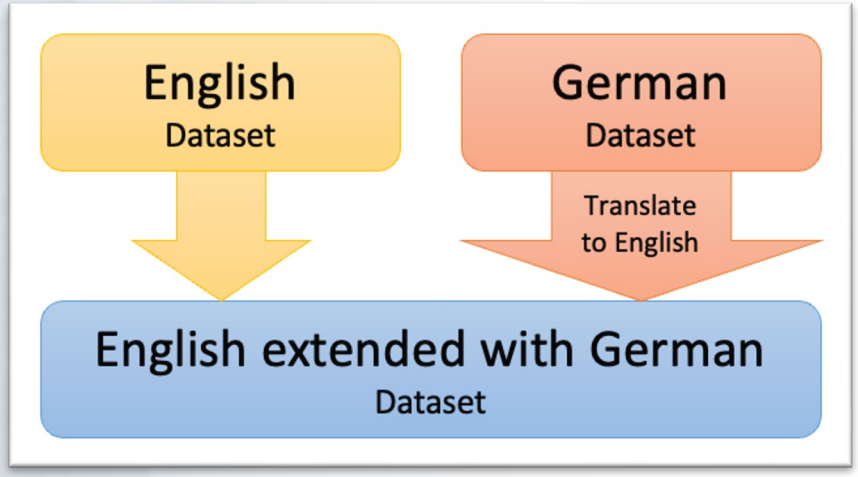
Methodology
Translation Application

```
# 1
# Merge German Emotions onto English
df_en = pd.merge(df_en, df_emotions, on='emotion_en', how='left’)

# 2
# Add German Sentence Column
df_en["sentence_de"] = ""

# 3
# Randomly select 1500 rows
df_en = df_en.sample(n=1500, random_state=2023)

# 4
# Save original to Disk
df_en.to_csv('./data/pd_en.csv', index=False)


# 5
# Iterate over the rows with tqdm to show the progress
for index, row in tqdm(df_en.iterrows(), total=df_en.shape[0]):
    # 6
    # Call Translation
    sentence = translate(row["sentence_en"], 'de’) # To German (‘de’)


    # 7
    # Save Sentence on Column
    df_en.at[index, 'sentence_de'] = sentence

# 8
#Save the file with all the translations!
df_en.to_csv('./data/pd_en_translated.csv', index=False)
```

MIDDLE TENNESSEE STATE UNIVERSITY

Methodology
Dataset Extension

```python
# Split the English and German Dataframes for Training and Testing
# We are using a 20/80 Split
df_en_train, df_en_test = df_en.randomSplit([0.85, 0.15], seed=2023)
df_de_train, df_de_test = df_de.randomSplit([0.85, 0.15], seed=2023)
print(f"English Train Row Count: {df_en_train.count()}")
print(f"German Train Row Count: {df_de_train.count()}")


# Create the Extended Dataframe wiht Translated Data
df_en_train_extended = df_en_train.union(df_de.select(*df_en_train.columns))
df_de_train_extended = df_de_train.union(df_en.select(*df_de_train.columns))
print(f"English Extended Train Row Count: {df_en_train_extended.count()}")
print(f"German Extended Train Row Count: {df_de_train_extended.count()}")
```



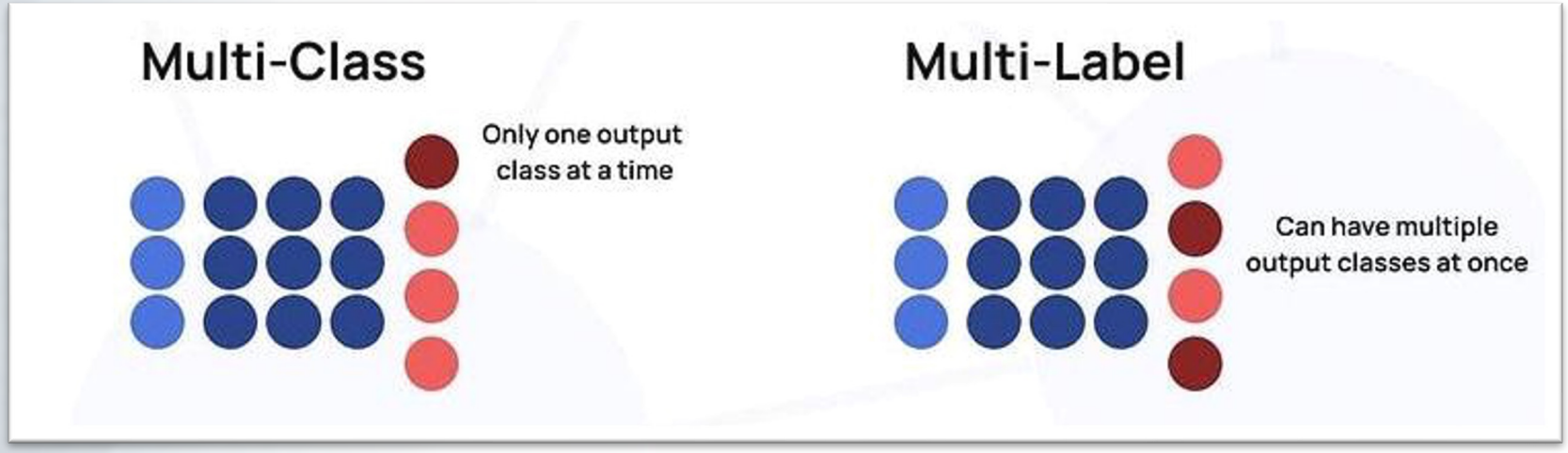
ML Models			
Model	Data	Rows for Training (85%)	Rows for Testing (15%)
A	English Original	1,275	225
B	English Extended By German	2,775	225 same as Model "A"
C	German Original	1,275	225
D	German Extended By English	2,775	225 same as Model "C"

MIDDLE TENNESSEE STATE UNIVERSITY

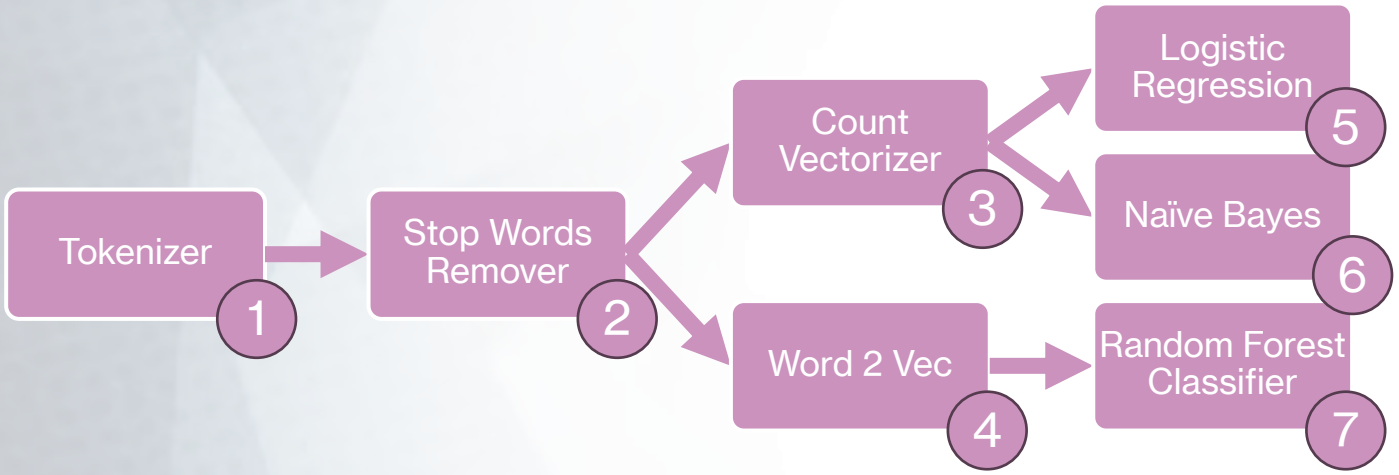
Methodology

ML Training & Testing with PySpark Multi-Class Classification

- Due to the eight (8) emotions present our model classification we decided to build of type **Multi-Class**
- **This is due to the research scope to be in search of predictive improvement and measuring a single class of emotion is more distinct than using Multi-Label**



Methodology
ML Training & Testing with PySpark Pipeline: 1



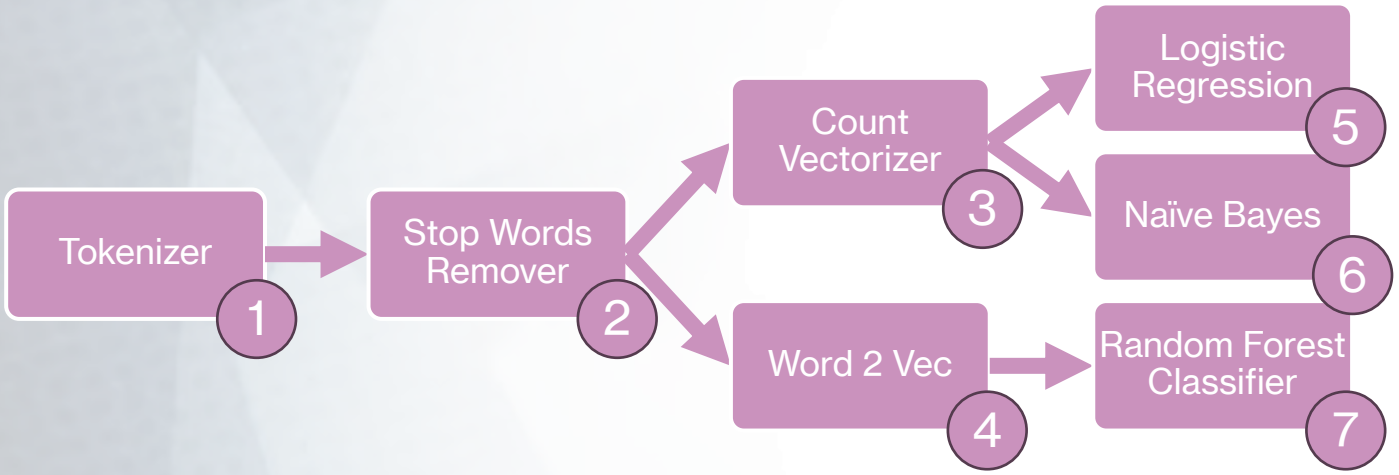
1 - Tokenization

- Used DistilBERT (multi-language) for word tokenization.
- We used DistilBERT from Hugginface's Transformers[9] due to it's download size and multi-language features.
- We opted to not use case sensitive tokenization due to the nature of tweets generally not following capitalizations.

MIDDLE TENNESSEE STATE UNIVERSITY

Methodology

ML Training & Testing with PySpark Pipeline: 2, 3, & 4



2 - Stop Words Remover

Removing words that occur commonly across the dataset.

3 - Count Vectorizer

Is a method to convert text to numerical data.

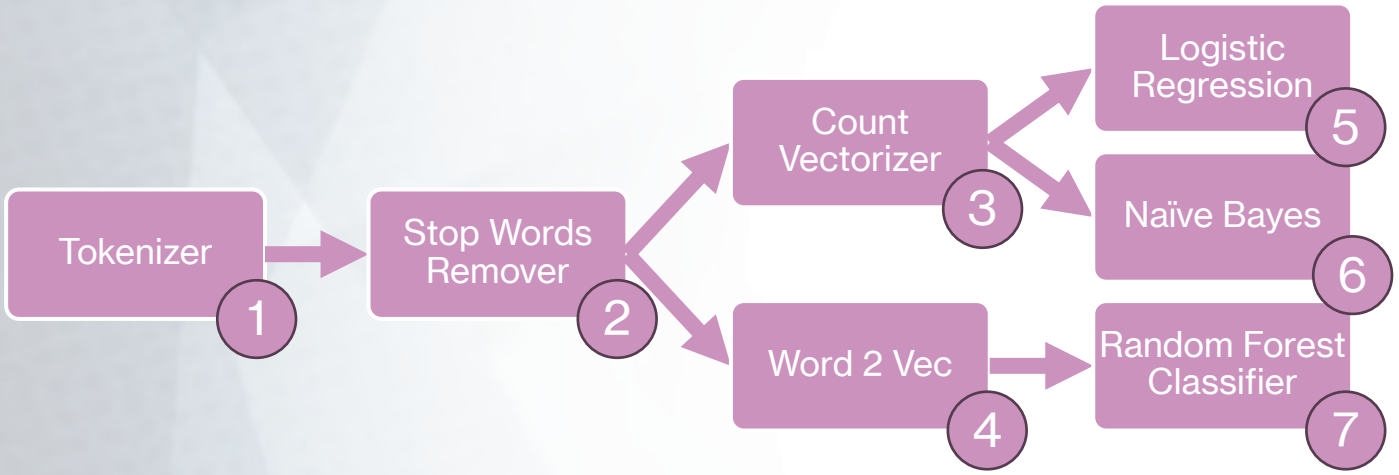
4 - Word 2 Vec

Is a way to group the vectors of similar words together in order to detect similarities mathematically. This was only used on Random Forest Classifier

MIDDLE TENNESSEE STATE UNIVERSITY

Methodology

ML Training & Testing with PySpark Pipeline: 5, 6, & 7



Learning Models – All Supervised

5 – Logistic Regression

It uses a logistic function to model the dependent variable.

6 – Naïve Bayes

This model is a probabilistic machine learning model that's used for classification task.

7 – Random Forest Classifier

RFC consists of a large number of individual decision trees that operate as a group. Each individual tree spits out a class prediction and the class with the most votes becomes our model's prediction.

Methodology

Creating an API for Real-Time Testing

Results & Analysis

Results & Analysis

Prediction Results of Original & Extended Datasets

Results & Analysis

Analysis of Impact of Extending Datasets

MIDDLE TENNESSEE STATE UNIVERSITY

Results & Analysis

Real-Time API Translation and Prediction

Conclusions & Future Work

Thank you

Citations

1. Andry Chowanda et al. “Exploring Text-based Emotions Recognition Machine Learning Techniques on Social Media Conversation”. In: Procedia Computer Science 179 (2021). 5th International Conference on Computer Science and Computational Intelligence 2020, pp. 821–828. issn: 1877-0509. doi: https : / / doi . org / 10 . 1016 / j . procs . 2021 . 01 . 099. url: https://www.sciencedirect.com/science/article/pii/S1877050921001320.

2. *Jay Stanley. ”THE DAWN OF ROBOT SURVEILLANCE”. 2019. url: https://www.aclu. org/report/dawn-robot-surveillance*

3. Sajani Ranasinghe et al. “An Artificial Intelligence Framework for the Detection of Emotion Transitions in Telehealth Services”. In: July 2022, pp. 1–5. doi: 10.1109/HSI55341.2022.9869503.

4. Laura Mascarell et al. “Stance Detection in German News Articles”. In: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER). Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 66–77. doi: 10.18653/v1/2021.fever-1.8. url: https://aclanthology.org/2021.fever-1.8.

5. Valentina Colonnello, Katia Mattarozzi, and Paolo M Russo. “Emotion recognition in medical students: effects of facial appearance and care schema activation”. In: Medical Education 53.2 (2019), pp. 195–205. doi: https://doi.org/10.1111/medu.13760. eprint: https:// onlinelibrary.wiley.com/doi/pdf/10.1111/medu.13760. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/medu.13760.

6. Shalini Kapoor and Tarun Kumar. “Detecting emotion change instant in speech signal using spectral patterns in pitch coherent single frequency filtering spectrogram”. In: Expert Systems with Applications 232 (2023), p. 120882. issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120882. url: https://www.sciencedirect.com/science/article/pii/S0957417423013842.

7. Sheetal Kusal et al. “AI Based Emotion Detection for Textual Big Data: Techniques and Contribution”. In: Big Data and Cognitive Computing 5.3 (2021). issn: 2504-2289. doi: 10.3390/bdcc5030043 . url: https://www.mdpi.com/2504-2289/5/3/43.

8. Kaggle. Url: https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

9. Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: ArXiv abs/1910.01108 (2019).