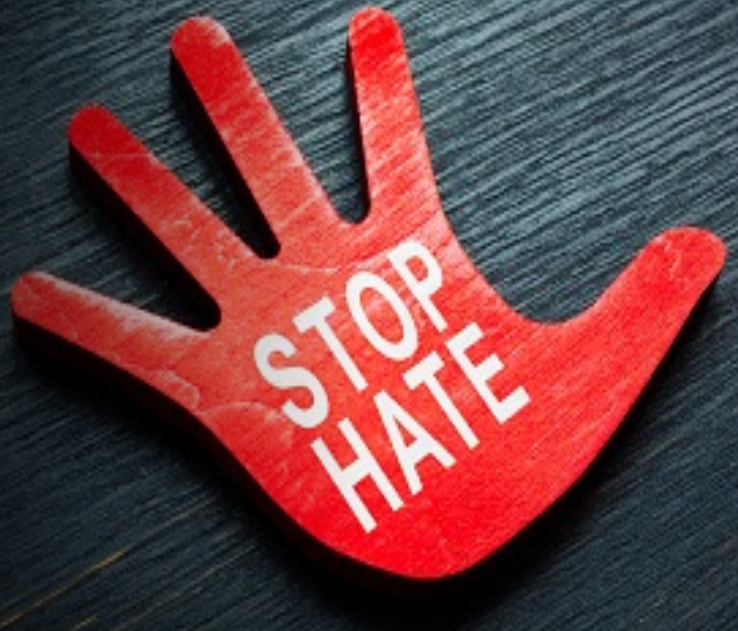
Multilingual Transfer Learning Hate Speech Detection



Group 25:

Aarushi Gupta
Garimendra Verma

Final Presentation(G-1): 6th December 2021
CS 4650: Natural Language (Prof. Diyi Yang)

Problem Definition

What is the problem?
- Hate Speech Detection - Multilingual

Is the proposed project unique?
- Scarcity of annotated low resource data like French, Spanish, German.

Why Transfer Learning?
- English - German **SI 0.6**
- French - Spanish **SI 0.75**
- English - French **SI 0.27**

Is the problem solvable?
- Combining - Hate Speech and Transfer Learning
- Many research studies already exists

Our proposition
- English + German > English + French / Spanish
- French + Spanish > French + German

Lexical Similar Language

	Catalan	English	French	German	Italian
Catalan	1	-	0.85	-	0.87
English	-	1	0.27	0.60	-
French	0.85	0.27	1	0.29	0.89
German	-	0.60	0.29	1	-
Italian	0.87	-	0.89	-	1
Portuguese	0.85	-	0.75	-	-
Romanian	0.73	-	0.75	-	0.77
Romansh	0.76	-	0.78	-	0.78
Russian	-	0.24	-	-	-
Sardinian	0.75	-	0.80	-	0.85
Spanish	0.85	-	0.75	-	0.82
	Catalan	English	French	German	Italian

Source: Wikipedia https://en.wikipedia.org/wiki/Lexical_similarity

0. Earlier Work

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8

Tweets - **Binary Classified**
{Hate Offensive, Not Hate Offensive}

Three Languages - **English, Spanish and German**

Baselines established by running pre-trained BERT on each language individually.

Use of Pre-Trained **BERT Model**
'bert-base-multilingual-uncased'.

1. Add Language

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8

Tweets - Binary Classified
{Hate Offensive, Not Hate Offensive}

Four Languages - English, **French** Spanish and German

Baselines established by running pre-trained BERT on each language individually.

Use of Pre-Trained **BERT Model**
'bert-base-multilingual-uncased'.

2. Update Baselines

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	**95.47%**	**83%**	**86%**	**0.87**

Tweets - Binary Classified
{Hate Offensive, Not Hate Offensive}

Four Languages - English, **French** Spanish and German

Baselines updated by running pre-trained BERT on each language individually.

Use of Pre-Trained **BERT Model**
'bert-base-multilingual-uncased'.

3. Transfer Learning

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	**95.47%**	**83%**	**86%**	**0.87**

Tweets - Binary Classified
{Hate Offensive, Not Hate Offensive}

Four Languages - English, **French** Spanish and German

Baselines updated by running pre-trained BERT on each language individually.

Performed Multilingual Transfer Learning

English + French / Spanish / German [3 combination]
French + Spanish / German [2 combination]

4. Result Accumulation

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	**95.47%**	**83%**	**86%**	**0.87**

Tweets - Binary Classified
{Hate Offensive, Not Hate Offensive}

Four Languages - English, **French** Spanish and German

Baselines updated by running pre-trained BERT on each language individually.

Dataset Collection | **Model Development** | **Collect Metrics**

Performed Multilingual Transfer Learning

English + French / Spanish / German [3 combination]
French + Spanish / German [2 combination]

Fine Tune on 1st Language → Evaluate on 2nd Lang → Fine Tune on 2nd Lang → Evaluate on 2nd Lang

Train and Val Accuracy Output

Test Accuracy Output

Train and Val Accuracy Output

Test Accuracy Output

Trained for 1st Language Evaluated on 2nd Language

Test Accuracy

Model trained on → 2nd Language ↓	Baseline	English	French
German	0.75	0.7	0.35
Spanish	0.8	0.45	0.54
French	0.86	0.35	-

BASELINES

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	95.47%	83%	86%	0.87

Inferences Drawn

English - German SI 0.6, Best performer

French - Spanish SI 0.75, best performer.

English - French SI 0.27 and French - German SI 0.27 , WORST performer

In Sync with lexical similar coefficient

Trained for 1st Language Evaluated on 2nd Language

Test Accuracy

Model trained on → / 2nd Language ↓	Baseline	English	French
German	0.75	0.7	0.35
Spanish	0.8	0.45	0.54
French	0.86	0.35	-

Trained on ENGLISH, further Fine tune for 2nd Language

Metrics → / 2nd Language ↓	Train	Val	Test	F1 Score
German	0.89	0.90	0.75	0.8
Spanish	0.92	0.82	0.83	0.83
French	0.95	0.84	0.85	0.86

BASELINES

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	**95.47%**	**83%**	**86%**	**0.87**

Inferences Drawn

English - French SI 0.27, Accuracy DEGRADED from baseline.

English - German SI 0.6, F1 slightly INCREASED.

Spanish English does not have similarity index, but number shows that English SUPPORTS Spanish.

In Sync with lexical similar coefficient

Trained for 1st Language Evaluated on 2nd Language

Test Accuracy

Model trained on → 2nd Language ↓	Baseline	English	French
German	0.75	0.7	0.35
Spanish	0.8	0.45	0.54
French	0.86	0.35	-

Trained on ENGLISH, further Fine tune for 2nd Language

Metrics → 2nd Language ↓	Train	Val	Test	F1 Score
German	0.89	0.90	0.75	0.8
Spanish	0.92	0.82	0.83	0.83
French	0.95	0.84	0.85	0.86

Trained on FRENCH, further Fine tune for 2nd Language

Metrics → 2nd Language ↓	Train	Val	Test	F1 Score
German	0.91	0.88	0.74	0.76
Spanish	0.92	0.8	0.79	0.79

BASELINES

Language	Train Accuracy	Validation Accuracy	Test Accuracy	F1 Score
English	95%	66%	78%	0.77
German	96%	85%	75%	0.76
Spanish	94%	80%	80%	0.8
French	95.47%	83%	86%	0.87

Inferences Drawn

French - Spanish SI 0.75, does not guarantee improvement

French - German SI 0.27, does not guarantee performance degrade

NOT In Sync with lexical similar coefficient

Data Augmentation?

Verify Similarity Index?

Conclusions

Future Scope

Problem Solved ?

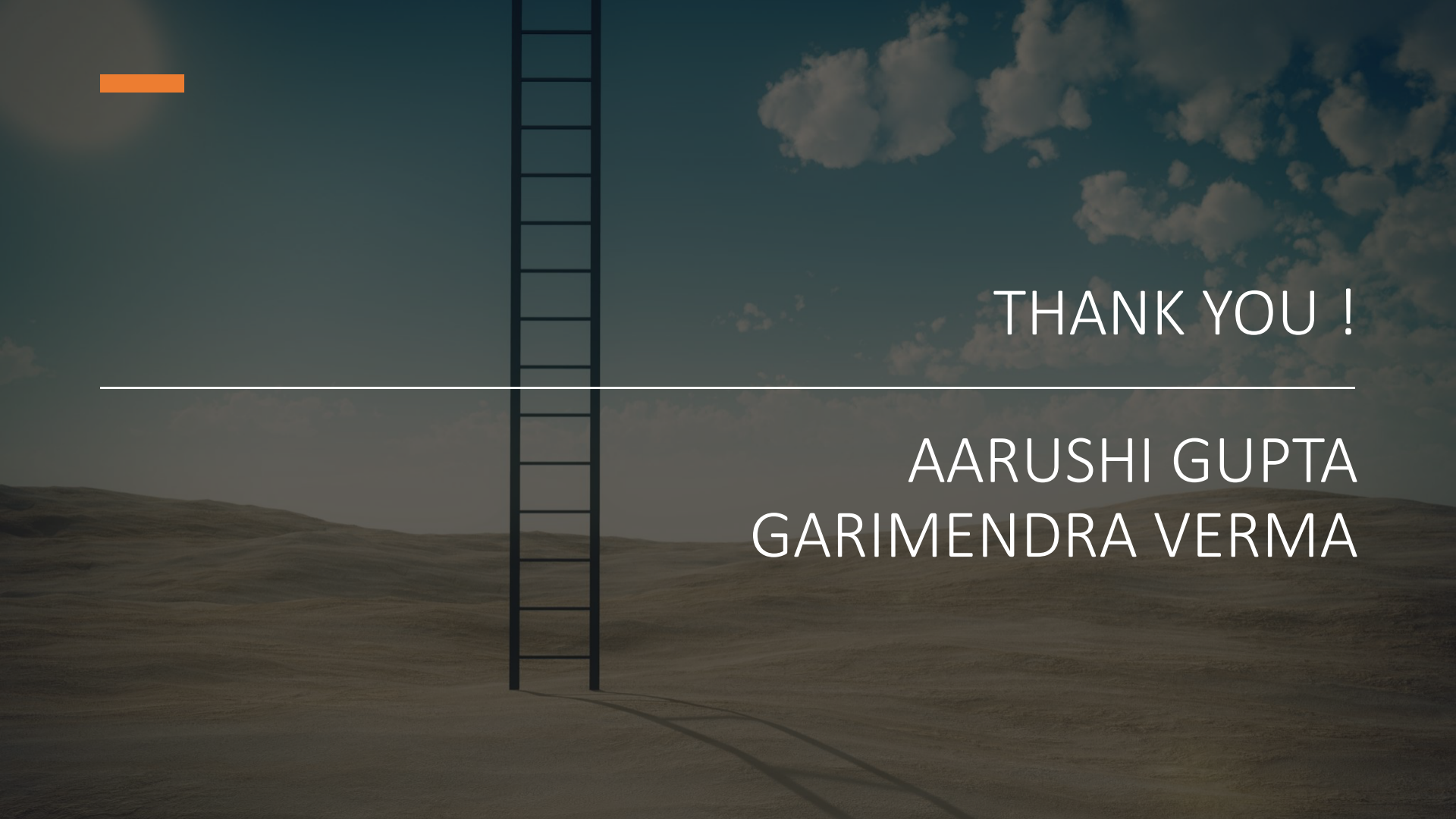
- **YES**, to the very least, similar languages sustain the performance , helpful for low resource documents.

Our Proposition is True?

- **MOSTLY**, we need to develop new model to verify what similarity index for language combinations.

Transfer Learning worked?

- **ALMOST**, Data Augmentation can be an alternative.



THANK YOU !

AARUSHI GUPTA
GARIMENDRA VERMA