Multilingual Transfer Learning for Hate Speech Detection

Aarushi Gupta
Georgia Tech / Atlanta, GA
agupta857@gatech.edu

Garimendra Verma
Georgia Tech / Atlanta, GA
garimendra@gatech.edu

Abstract

This paper works through pre-trained BERT model and its application in social media hate speech detection for multiple languages. The paper aims to leverage the lexical similarity between two different languages to solve the problem of unavailability of annotated low resource data. The experiment is performed on 4 languages - English, French, Spanish and German. Lexical Similarity Index (SI) is based on literature. In linguistics, lexical similarity is a measure of the degree to which the word sets of two given languages are similar. We here propose that English German with SI 0.6 will perform better than English French with SI 0.27. Similarly, French Spanish with SI 0.75 should perform better than French-German 0.29 SI. We chose transfer learning as our method to train a pre-trained multilingual BERT from one language to another language. The results are then compared for how lexical similarity affects the performance for Hate Speech Detection. The paper will discuss the benefits of our approach and shortcomings that needs to be rectified in future scope. We were able to find that English supported Other languages in sync with the lexical similarity, but French could not. The possible reasons are discussed in later sections of report. *Please click for github URL for our project can be found here.*

1 Introduction & Related Work

With the increase in social media interactions, there are high chances that a person can see some tweet which is either offensive to them or their loved ones. This raises a serious concern when such hateful communications leads to bigger impacts like violence, racism, sexism, etc. Social media is a powerful tool in building a stronger community, thus there is a need to develop tools to detect aggressive behaviour in general and hate speech in particular, that could lead to not just in verbal crime but physical violence and serious harms. There are lot of efforts in this area to automatically recognize such hateful tweets leveraging the Natural Language Processing Methods.Especially BERT has proven to be an efficient model to solve this problem, as shown by Pinkesh Badjatiya (2017). Recently, a problem that has surfaced in such solutions is the presence of different languages on social media.

With globalisation of many social media apps and a quest to support more and more native languages, the data being generated is very different from the models so far developed. Many work have been done on multilingual natural language processing like Data Augmentation Mihaela Bornea (2020) and Transfer Learning (Shifan Mao, 2017). After, TAs comment in midway report we differentiated in Data Augmentation and Transfer Learning, to only proceed with Transfer Learning for our project. Earlier, HASOC-FIRE (2020) launched a challenge to develop a model which can detect hate speech in tweets for three different languages - English, German and Hindi. This challenge lead to many experiments with multilingual BERT model that showed good performance [Marzieh Mozafari (2019)]. We chose one of the published researches from the challenge, by Suman Dowlagar (2021), where English was solved using 'bert-base-cased' and German and Hindi were solved using 'bert-base-multilingual-uncased' pre-trained models. We chose transfer learning to further improve upon the baselines established in this paper by using the concept of lexical similarity between languages.

The transfer learning method has been explored but for one language, inspired by the work of Mozhi Zhang (2018), we decided to take leverage of lexical similarity in two languages for training. We aim to propose transfer learning of hate speech trained model for one language to another language with high similarity index(SI) as described in Wikipedia, Figure 1. This paper will discuss four languages - English, German, French and Span-

	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

Figure 1: Image showing lexical similarity for language combinations.[1: exactly same, 0: nothing similar] Source: Wikipedia

ish, hypothesizing that BERT trained for English will perform well for German with high SI 0.6 and not good for French with low SI 0.27. Similarly, French should work better with Spanish (SI 0.75) but not with German (SI 0.29). Since, similarity index is not given for Spanish we do not consider as base language. The study could help with low resource documents such as German and French.

2 Methods

The following section will discuss the data collected, data pipeline, model pipeline developed and baselines established to compare the performance of our proposed strategy to.

2.1 Dataset

Dataset Collection & Analysis The dataset required for this experiment was supposed to have a combination of language which has a high similarity index and a low similarity index for comparing the effects of SI on multilingual hate speech detection. Earlier we were able to gather twitter hate speech data for English, German and Hindi from the challenge HASOC-FIRE (2020), which has a subtask with focus on Hate Speech and Offensive language identification offered for English, German and Spanish. The annotated data is available for the tweets classified into 2 classes, namely : Hate and Offensive (HOF) and Non-Hate and Offensive (NOT). We disregarded Hindi as it has a very low lexical similarity with other languages.

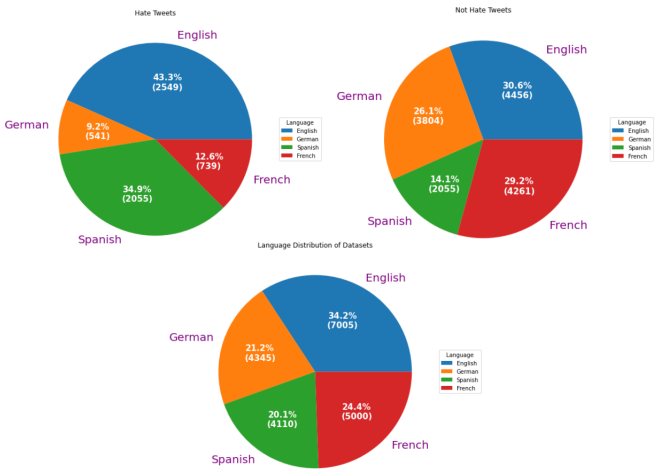

Figure 2: Distribution of Tweets

Later, we gathered French and Spanish twitter data with same class labels from (Kaggle) and SemEval Task 2019 (Mihai Manolescu, 2019). Figure 2 shows that English has most number of tweets and other languages are almost same. This is natural as English is most spoken language on social media platforms. However, the class distribution is not equal for German and French, unlike Spanish and English. This might create a difference in the results that we are expecting. This can be validated in Figure 3 showing the train and test split of the tweets which data consisted. Train data is split in ratio 80-20 later for model training. There is very high skewness in French and German data, but not removing this skewness becuase in real life also, hate speech tweets might be available in low number.

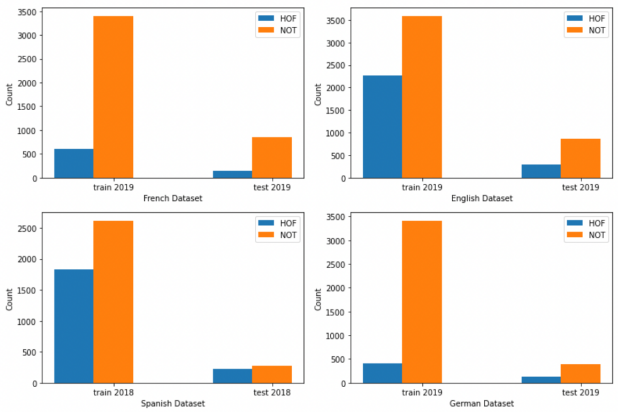

Figure 3: Data Split for French , English, Spanish & German (top-bottom , left-right)

Some examples from datasets are as follows-

- **French Dataset**:

(1) Hateful - *L'ensemble de sied van reil devrait être putain*
(2) Not Hateful - *J'ai oublié un de mes tweet peeps ... nooooo! Tweets drôles !!! X*

- **English Dataset**:
  (1) Hateful - *He's had the worst numbers of any president ever. Theres not much "Lower" to go. TrumpIsATraitor*
  (2) Not Hateful - *Kathy Zhu, a supporter of President Trump, was stripped of her title because of comments about African-Americans and Muslims.*

- **Spanish Dataset**:
  (1) Hateful - *Tú eres la perra, no te dejes engañar.*
  (2) Not Hateful - *Ciberactivismo por la Revolución y los DerechosHumanos en Túnez. Mujeres en primera línea de la revolución árabe.*

- **German Dataset**:
  (1) Hateful - *Halts Maul du hurensohn deine Mutter ist stolz auf mich*
  (2) Not Hateful - *zdf was ist denn da bei euch los????*

For **data preprocessing**, we removed (1) stop words (using nltk library), (2) special characters like &,,%,$, etc. (3) twitter handles and (4) non word characters.

2.2 Models and Analysis

This section will explain the model we chose to experiment with. It will explain how the model pipeline is created and baselines established.

BERT is State of the art language model widely used for making machines understand NLP. It is pre-trained by Google AI on 102 languages. The model is a bidirectional transformer which uses masked language modeling (MLM) for the next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. We intend make use of *'bert-base-multilingual-uncased'* transformer provided by hugging face.

Analysis Pipeline includes fine-tuning the pre-trained BERT models for enhancing the performance of hate speech detection on low resource languages. We first selected combinations of languages that we want to analyse the performance

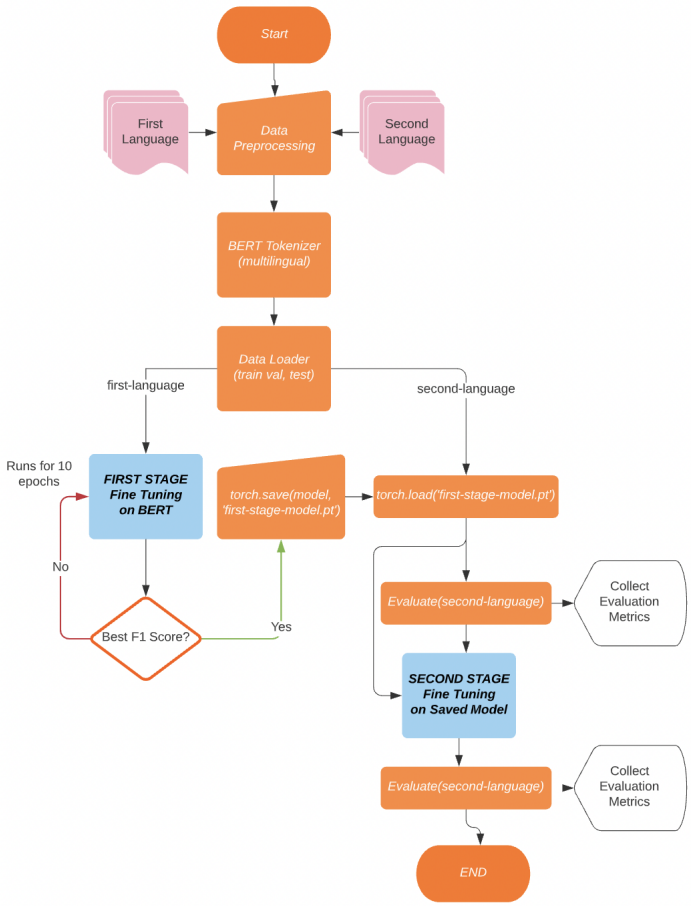

Figure 4: Analysis Pipeline for Multilingual Transfer Learning.

on. These combinations are summarized in the Figure 6 with SI indicating how well each one should support the language. Please note, that these combinations are selected on the basis of availability of data and Lexical Similarity Index.

Figure 4 shows the complete flow of the process, to apply the transfer learning on combination of two languages. There are two stages of fine tuning in this flow. **First stage**, fine tuning on pre-trained BERT is performed for one language, this fine tuned model is efficient in hate speech detection for one language. **Second stage**, fine tuning is performed on the output of previous step. This second fine tuning is the transfer learning which we apply for hate speech detection from one language to another. Since same BERT tokenizer (multilingual based) is used for both the tuning, we ensure that two language to share embedding space. And being lexical similar language must have similarity in the word embeddings. This way we expect lexical similar language to perform better when one fine tuned model is again fine tuned on second

language.

Evaluation is done in two steps. Firstly, we collect test accuracy on test dataset of second language by passing it as input for first stage fine tuned model. Secondly, we again collect evaluation metrics on the test dataset passed as input to second stage fine tuned model. We evaluate performance in terms of (1) **F1 Score** - combination of precision and recall, taking into account both false positive and false negative and (2) **Test Accuracy** - accuracy is calculated on held out test dataset from second language after both stages. First stage results, show how well can model detect the second language when it is fine tuned on first one. Second stage results are the values that we want the new model to beat the baseline model.

To compare between high similar and low similar language combinations, we use English as first language - to be used for fine tuning in first stage; and French, Spanish, and German to be used for fine tuning in second stage. Similarly, another combination of fine tuning stages is French in first stage, with Spanish and German in second stage.

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

Figure 5: Baselines established for Multilingual Transfer Learning.

Baselines are established from our reference paper Suman Dowlagar (2021). Here, we fine tuned the pretrained BERT models on individual language and record the training, validation and test accuracy along with F1 score on test dataset. Figure 5 show these metrics. We aim to use English and French fine tuned model and again fine tune them on other languages to understand how lexical similar index could be related to performance changes observed.

Basic fine tuning shows that BERT provides maximum accuracy for French dataset, followed by Spanish, English and German. But there is obviously scope for improvement due to low F1 Score.

First Language	Second Language	Lexical Similarity Index	Hypothesis
English	French	0.27	Should degrade performance
English	Spanish	Not Available	-
English	German	0.6	Should improve performance
French	Spanish	0.75	Should improve performance
French	German	0.29	Should degrade performance

Figure 6: Hypothesis formed based on Lexical Similarity Index.

3 Results

This section will discuss the experimental setup and results collected from first and second stage fine tuning of BERT multilingual model. This section will also consider work division.

3.1 Experiment Setup

The experiment was setup using Google Colab Pro. It requires GPU and High RAM to be computationally efficient. The experiment is performed using pretrained BERT models imported from transformers library [Face]. For this transformers should be installed. We have used 'bert-base-multilingual-uncased' from transformers so that all languages share same tokenizer and share the embedding space. The experiment also requires NLTK library which is equiped to for text manipulations, we have used it for data preprocessing. The model training uses Optimizer AdamW. AdamW is a stochastic optimization method that modifies the typical implementation of weight decay in Adam to combat Adam's known convergence problems by decoupling the weight

3.2 Result Comparison

First-Stage Tuning Evaluation The first set of results are shown in Figure 7. Here, we can clearly see that the results obtained are in sync with the lex-

Model trained on → 2nd Language ↓	Baseline	English	French
German	0.75	0.7	0.36
Spanish	0.8	0.45	0.54
French	0.86	0.35	-

Figure 7: Results from First Stage Fine-Tuning.

ical similarity index mentioned in Figure 6. These results are obtained by testing the first stage fine tuned model on second languages mentioned as row names. These results are an indication of how well the lexical similarity index depicts that two languages are similar and not. We see that German gave 70% accuracy when the model was not even tuned on German text at all. Similarly, Spanish gave a 54% accuracy on French tuned first stage model. These results suggests that even data augmentation can be a valid method for multilingual hate speech detection, provided two languages combined are lexical similar. Also, these results show that we can consider these accuracy as a symbol of lexical similarity index. We need this because of the non availability of SI for many combinations of languages (here, Spanish).

Metrics → / 2nd Language ↓	Train	Val	Test	F1 Score
German	0.89	0.90	0.75	0.8
Spanish	0.92	0.79	0.76	0.83
French	0.95	0.84	0.84	0.86

Figure 8: Results from Second Stage Fine-Tuning on English trained Hate Speech BERT.

Second Stage Tuning Evaluation for English Tuned Hate Speech Bert Figure 8 shows results for how the second languages (mentioned as row indices) performed on model fine tuned for English. Overall it is again in sync with the lexical similarity index. The test accuracy for German remained same as compared to baseline but F1 score improved from 0.76 to 0.8, that means there is reduction in false positives and false negatives. Similarly, for French the test accuracy is degraded from 86% to 84%, which is expected as they have a lower SI. Since, French and Spanish have a higher lexical similarity by transition property, English and Spanish should have low lexical similarity index as well, and it is depicted by results as well when test accuracy reduces from 80% to 76%. Hence, this experiment proves our hypothesis for English based lexical similarity.

From these learning curves 9 we can see that how training on German actually helping the model to learn more, but for French 10 the model is overfitting as training accuracy increases but validation accuracy decreases and fluctuates since the start.

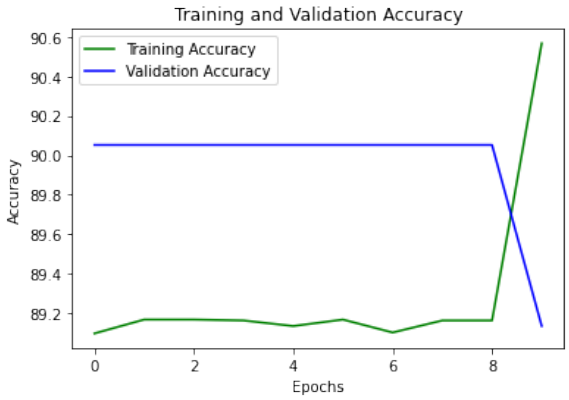

Figure 9: Learning Curve for English on Second Stage Fine-Tuning on German trained Hate Speech BERT.

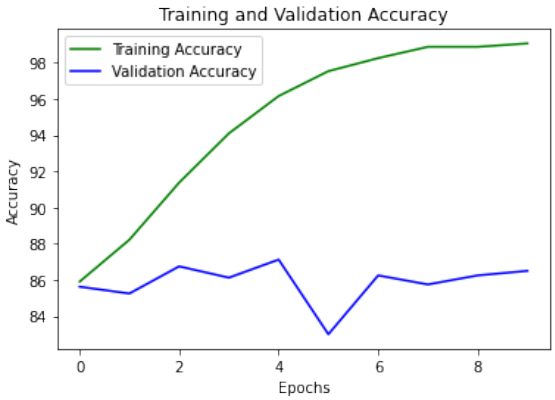

Figure 10: Learning Curve for English on Second Stage Fine-Tuning on French trained Hate Speech BERT.

Second Stage Tuning Evaluation for French Tuned Hate Speech Bert Figure 12 shows results for how the second languages (mentioned as row indices) performed on model fine tuned for French. Here, we do not test French-English combination as we already have abundance of English annotated data. Quite contrary to our hypothesis the results are not in sync with our hypothesis. Spanish test accuracy and F1 Score reduced from baselines by 1% after second stage fine tuning. Also, German and French have a very low SI of 0.29 but the performance is at par with baselines and not much affected. Therefore, these results do not show very promising confidence to use lexical similarity index. However, we did more analysis for this.

Modification to Nullify negative results in French based Transfer Learning When we look at French and German dataset, they are highly skewed with Not Hateful tweets much more in number hence, we can assume that the skewness might have affected these results. Similarly, there can

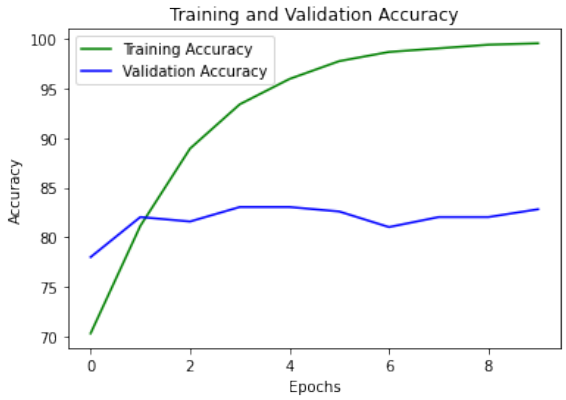

Figure 11: Learning Curve for English on Second Stage Fine-Tuning on Spanish trained Hate Speech BERT.

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

Figure 12: Results from Second Stage Fine-Tuning on French trained Hate Speech BERT.

be wrong annotations in the data provided (as in ground truth). We need to verify the datasets as manually there were some false positives observed in french dataset on random analysis. This could be a future scope for this project.

We also analyzed that the learning curve 14 is not learning at all. Hence the model is overfitted and does not provide any good results for French based Transfer Learning. However, Figure 16 shows how model keeps learning and hence, is better at predicting the French hate speech when combined with Spanish based Transfer Learning. We had to tune the number of epochs to reduce overfitting in second learning curve. But same was not the case with first one. We could not tune it to reduce the overfitting.

To prove that the French dataset do no fail because of wrong SI or wrong hypothesis assumed earlier, we use Spanish too for fine tuning in first

Metrics → 2nd Language ↓	Train	Val	Test	First Stage Test Acc
German	0.927	0.88	0.74	0.29
French	0.91	0.86	0.88	0.79

Figure 13: Results from Second Stage Fine-Tuning on Spanish trained Hate Speech BERT.

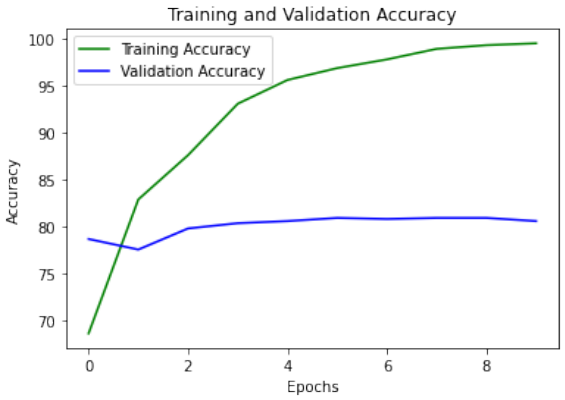

Figure 14: Learning Curve for French on Second Stage Fine-Tuning on Spanish trained Hate Speech BERT.

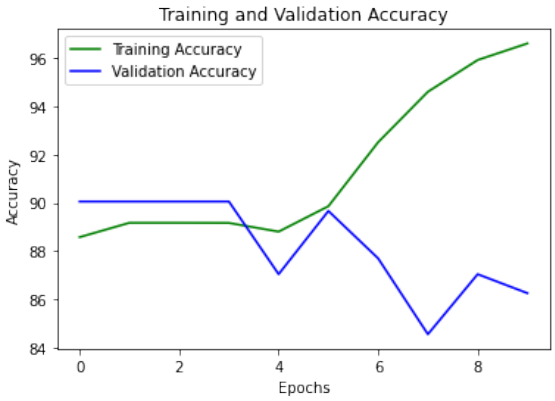

Figure 15: Learning Curve for French on Second Stage Fine-Tuning on German trained Hate Speech BERT.

stage and then fine tune French and German on it. Spanish dataset is not skewed and can provide better analysis at this. Figure 13 shows that French has performed very well on Spanish fine tuned first stage BERT model. The improvement is more than that of English and German, which means that SI is relative to the performance that can be improved. French test accuracy jumped from 86% to 88% after we used Spanish finetuned hatespeech BERT. Even, for Spanish based transfer learning, we achieved firsts stage test accuracy for French as 79%. Hence, our hypothesis is still true. Considering French and German have low SI, so will Spanish and German, we see that performance for German has reduced by 1% (Same as French based German results).

Learning curves for each model trained could be found in Jupyter Notebooks saved in github URL provided with this report (abstract).

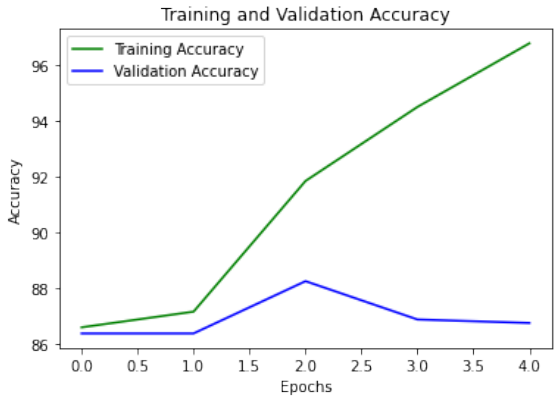

Figure 16: Learning Curve for Spanish on Second Stage Fine-Tuning on French trained Hate Speech BERT.

3.3 Work Division

The following is the detailed work division for each team member. **Please consider this.**

Aarushi Gupta has participated in brainstorming for the idea for project and modifications performed, implementation of code to conduct experiments, data pipeline, model pipeline, store the required results, drawing conclusions and inferences from results, complete final report, complete midterm and final presentation PowerPoint slides, partial midterm report.

Garimendra Verma has written project proposal and partial midterm report. His contribution is also there for project idea brainstorming.

4 Conclusions

High Level Conclusions (1) We can use First Stage Evaluation Metrics as BERT based lexical similarity index, we found that they were properly in sync with our source SI. We were able to find a similarity coefficient for even Spanish which was absent from the source (Wikipedia). (2) Second Stage Evaluation Results show how Spanish and English were able to support French and German. English and Spanish are among the most widely spoken languages and with these results we can positively affirm that High lexical similar languages can aid low resource documents in multilingual NLP problems (specifically Hate Speech Detection).

Low Level Conclusions (1) First stage tuning evaluation metrics show that even without second stage fine tuning we were able to achieve good score, which means that there is a high possibility that Data Augmentation can help when two high lexical similar index languages are combined. (2) Skewness and less annotated data cannot be used as first langauge for First Stage Fine tuning of BERT model.

Comparison with Baselines Our proposed strategy was able to **beat the baseline models**. Our models were able to imporve the baseline performances. French performance improved from 86% to 88% when fine tuned over Spanish, German F1 Score improved from 0.77 to 0.8 when fine tuned on English.

Future Scope Define Lexical Similarity Index mathematically, based on first stage tuning results. This ensures that similarity is in accordance with Vocabulary of BERT and a shared embedding space.

Data Augmentation, results show promising outcomes if we augment the data instead of transfer learning, as computationally less expensive and less time to train the models.

5 Ethical/Broader Impacts Statement

Positive – Social media is a multilingual platform. It could be helped with hate speech prohibition. Right now many false positive exists, which can lead to prohibiting a language specific group of people. Specially, when they are in low resource. Also, language translation models can use our hate speech detection to modify source language and do not provide translation for hated words.

Negative – The model if in wrong hands can be used to detect hate speech communications. People can use this to target specific tweets to specific group of audience. As it helps in detection hate speech, it can be used spread hatred even faster.

References

Hugging Face. bert-base-multilingual-uncased.

HASOC-FIRE. 2020. Hasoc-fire 2020.

Kaggle. French tweets.

Noel Crespi Marzieh Mozafari, Reza Farahbakhsh. 2019. A bert-based transfer learning approach for hate speech detection in online social media. *Social and Information Networks*, arXiv:1910.12574.

Sara Rosenthal Radu Florian Avirup Sil Mihaela Bornea, Lin Pan. 2020. Multilingual transfer learning for qa using translation as data augmentation. *Computation and Language*, arXiv:2012.05958.

Adham Nasser Mohamed Saber Masoumeh Moradipour Tari Mihai Manolescu, Denise Löfflad. 2019. Lstm approach to hate speech detection in english and spanish.

Jordan Boyd-Graber Mozhi Zhang, Yoshinari Fujinuma. 2018. Exploiting cross-lingual subword similarities in low-resource document classification. *Computation and Language*, arXiv:1812.09617.

Manish Gupta-Vasudeva Varma Pinkesh Badjatiya, Shashank Gupta. 2017. Deep learning for hate speech detection in tweets. *Computation and Language*, arXiv:1706.00188.

Jake Dong Shifan Mao, Weiqiang Zhu. 2017. Transfer learning on stack exchange tags. *CS224n: Natural Language Processing with Deep Learning*.

Radhika Mamidi Suman Dowlagar. 2021. Using bert and multilingual bert models for hate speech detection.

Wikipedia. Lexical similarity.