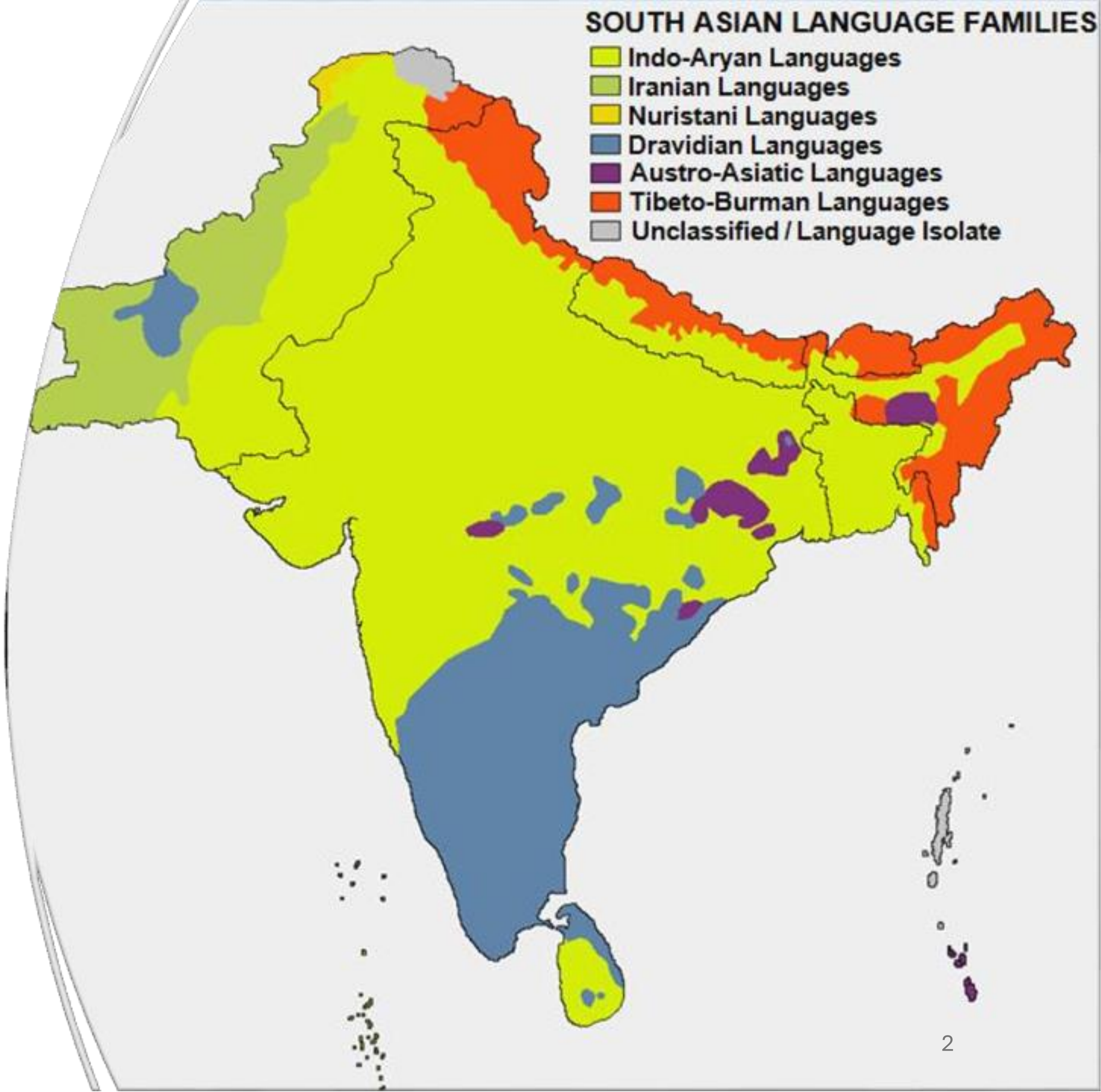
Sentiment Analysis for Chungli Ao

BY: Lena, Nicholas, Ravi &
Nellia

Language Families in south Asia

- People in India speak languages from four language families
- Indo-Aryan
- Dravidian
- Austro-Asiatic
- Tibeto-Burman

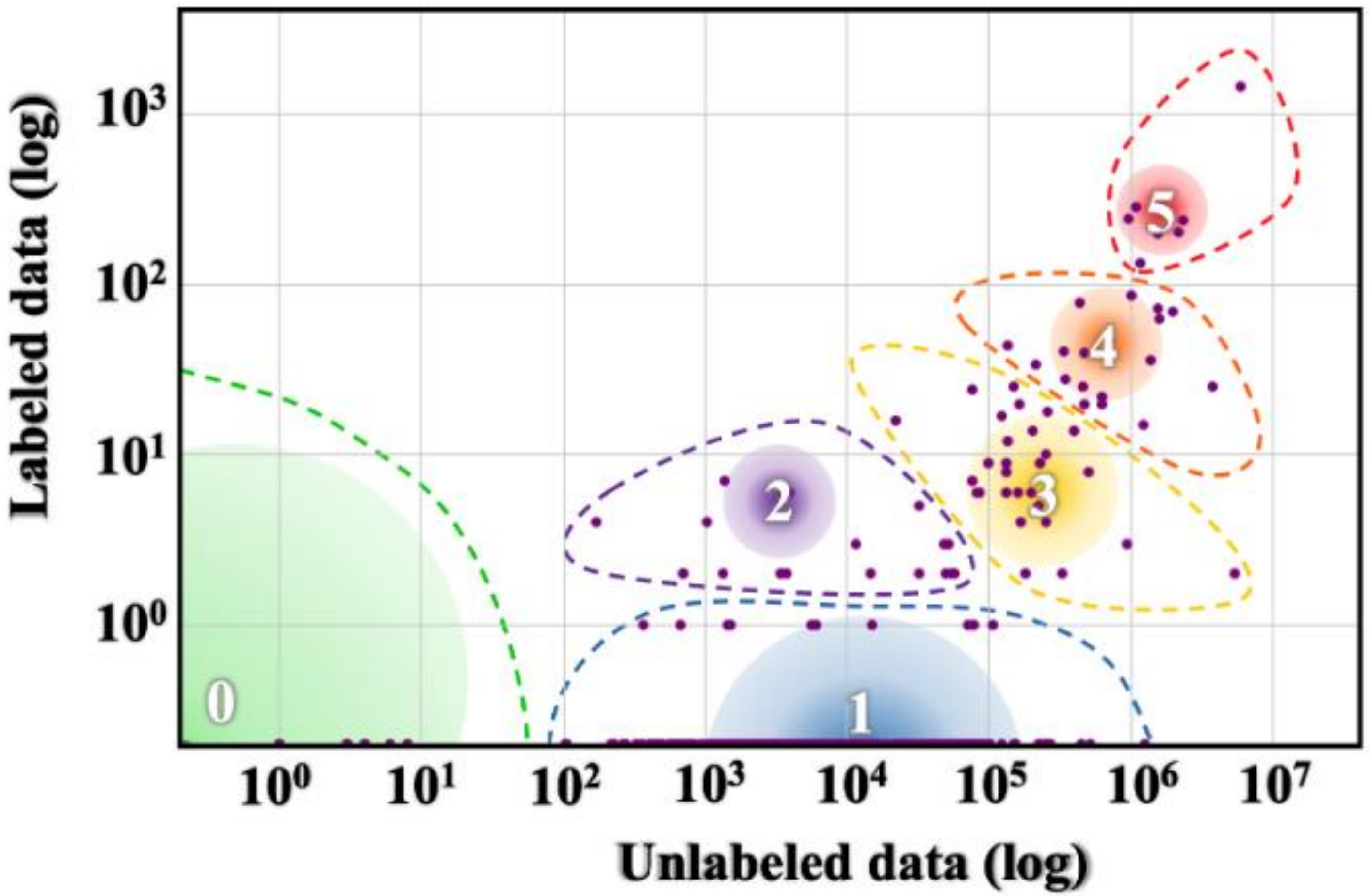


North east India

- Also known as seven sisters.
- Chungli Ao is a dialect of Ao language.
- It is an administrative language.
- Spoken in Nagaland.
- Mizo is official language of Mizoram.
- Mizo and Chungli Ao fall under Sino-Tibetan language family.

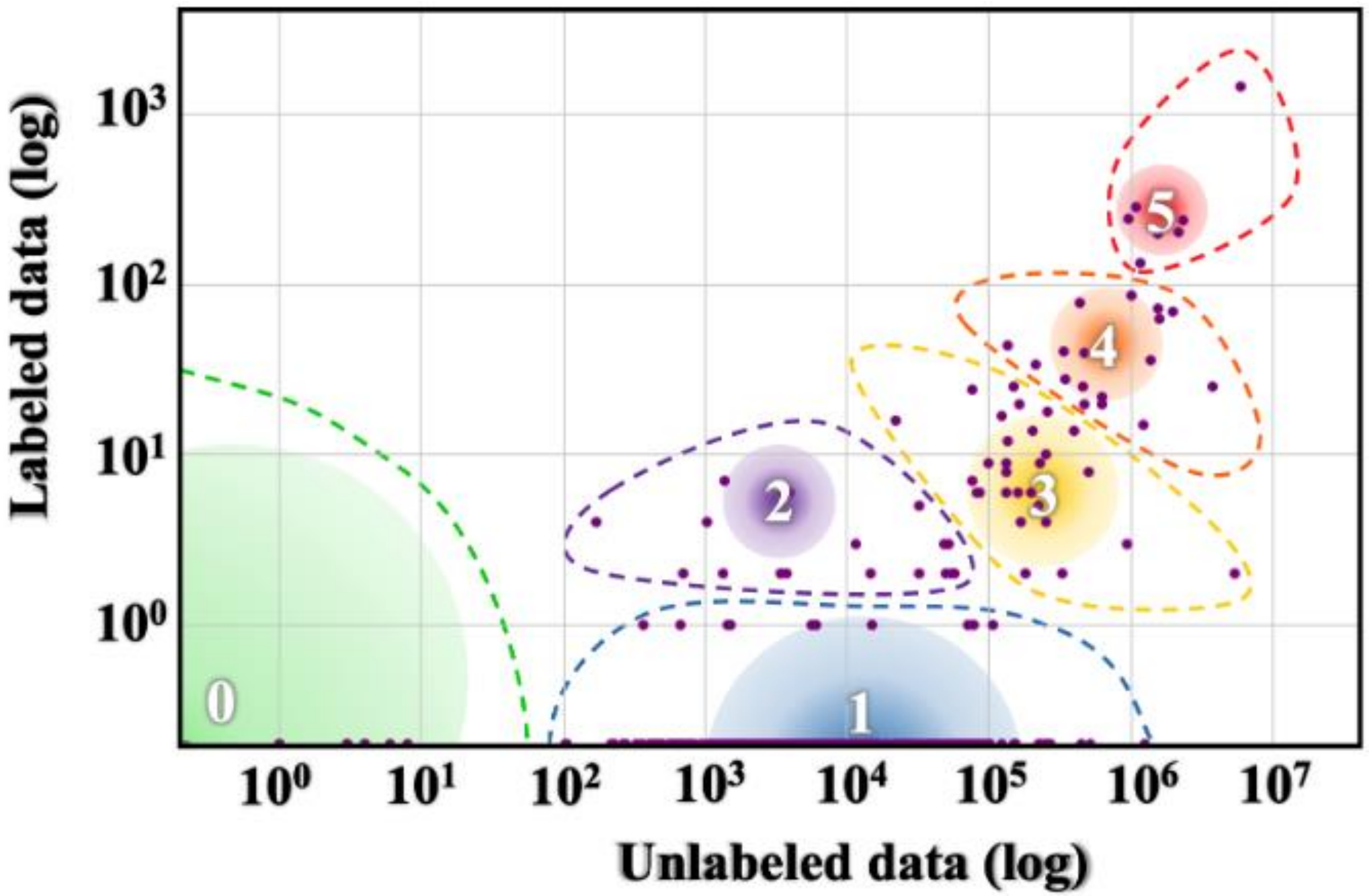
Language Classification by Resource Availability

- categories according to Joshi et al 2020.
- The left behinds (0): Ignored in language tech, No unlabeled data
- The scraping By's (1) : Some unlabeled data, Potential with organized efforts
- The Hopefuls(2) : Small labeled datasets.



Language Classification by Resource Availability

- The raising stars (3): Benefit from pre-training, Strong online presence.
- The Underdogs(4): Lots of resources and unlabeled data, Less labeled data.
- The winners(5): Leading in language tech, Major investments.



Motivation for this project

- **High Resource Languages**:
  - Extensive research and resources available
- **Chungli Ao**:
  - Limited research and resources
  - Identified research gap
- **Exploring Multilingual Models**:
  - Investigate the use of current multilingual models like XLM-R and m-BERT
- **Comparison of Approaches**:
  - Compare Machine Learning(ML) approaches with Deep Learning approaches (DL)

Experiments for this project

- ML experiments (using SVM & Naïve bayes)
- Zero-shot (Base line)
- Experiments with Chungli Ao Bert
- Multilingual Data Augmentation
- Back Translation Data Augmentation
- Pretraining
- Accuracy is the metric for evaluation

Languages used for this project

- English, German, Russian, Telugu, Mizo & Chungli Ao
- English, German, Russian, Telugu: Because the team knows these languages
- Mizo : Closely related to Chungli Ao

Open source datasets

English : Kaggle dataset based on Twitter data

German : The dataset Broad-Coverage German Sentiment Classification Model for Dialog Systems

Russian : An automatically collected dataset for sentiment analysis of product reviews

Mizo: Sentiment data created from various domains

Telugu Data

- Telugu sentiment data
- Web scraping data from YouTube
- YouTube API
- 2 speakers cleaned and checked the quality
- Domains: Movies, Music, News

Chungli Ao Data

- Chungli Ao sentiment data
- Translation of Amazon Reviews from English to Chungli Ao
- Newspapers data converted to sentiment data
- Domains: Product reviews, News

Data information for this project

Language	Positive	Negative
Chungli Ao	4505	4074
Telugu	3006	3237
German	3000	3000
English	3000	3000
Mizo	3000	3000
Russian	3000	3000

Train set

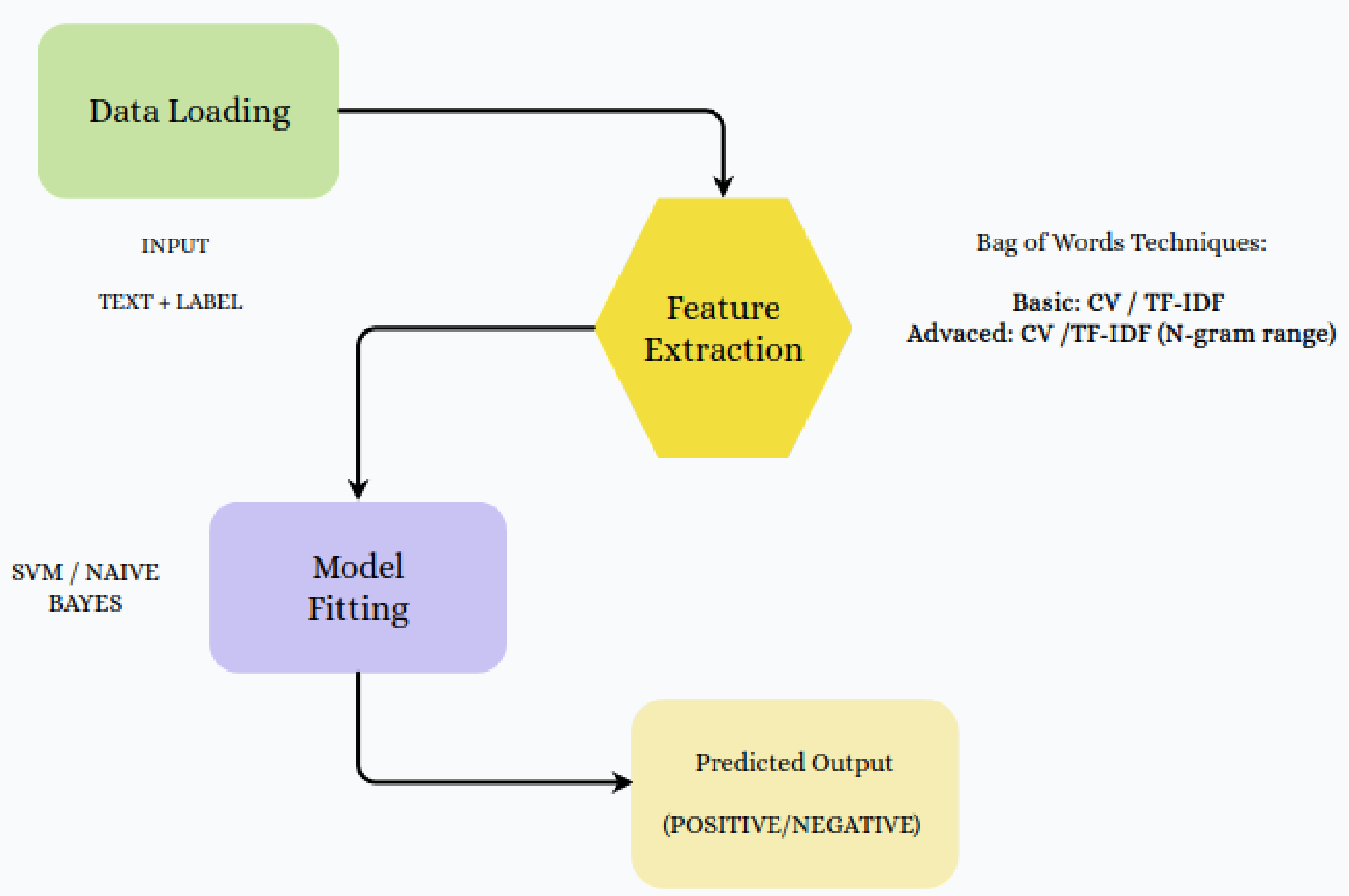
Language	Positive	Negative
Chungli Ao	1000	1000

Test set

Chungli Ao Bert

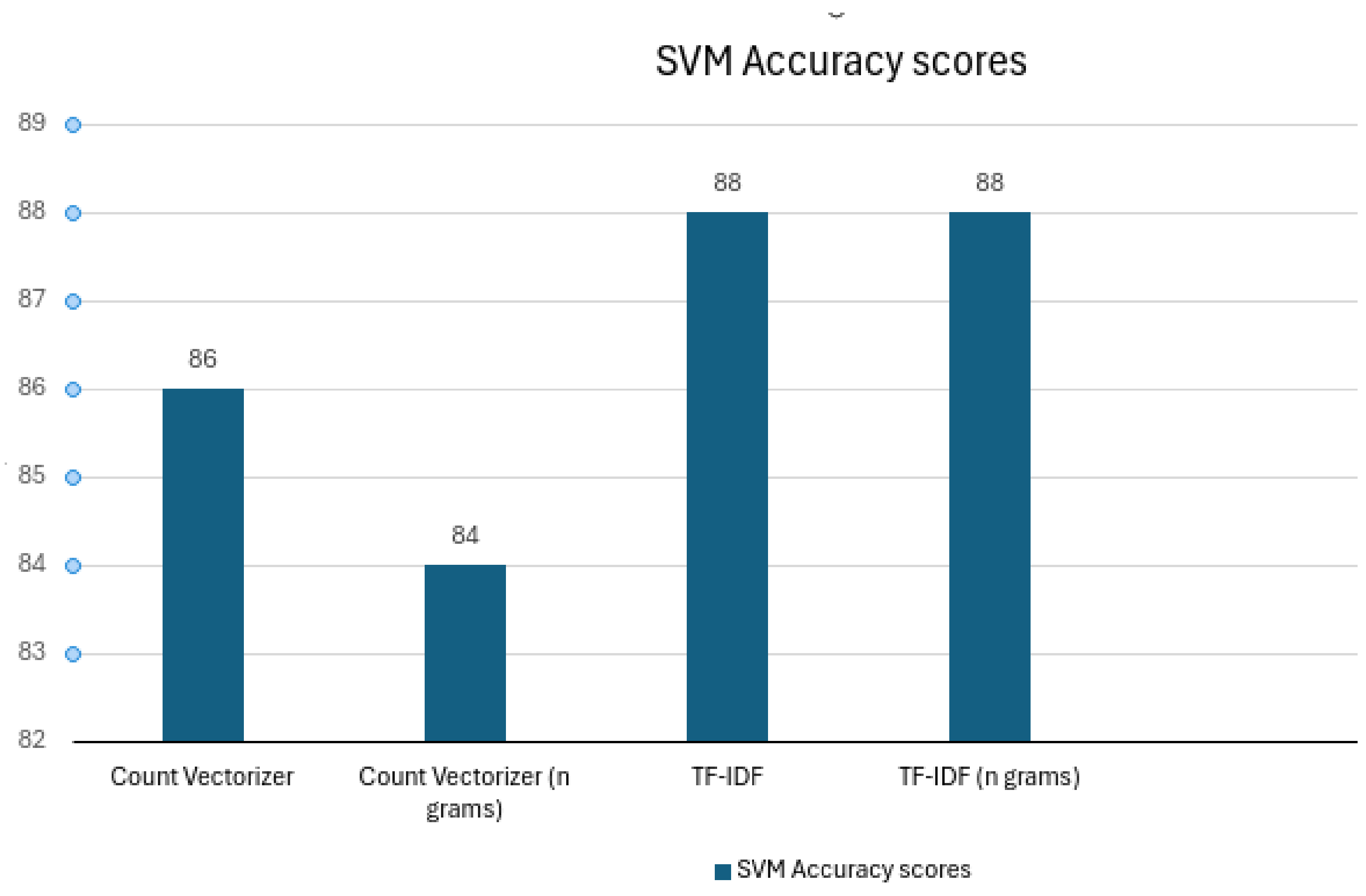
- Un labelled of Chungli Ao is used to create tokenizer
- Created a tokenizer using "*word piece*"
- **['[CLS]', 'sikkim', 'kubok', 'namchi', 'central', 'jail', 'nung', 'puoka', 'alir', 'aser', 'staff', 'sentepa', 'nisung', '[SEP]']**
- Finetuned a Bert model using sequence to sequence classification
- Pushed the model to hugging face with the tokenizer

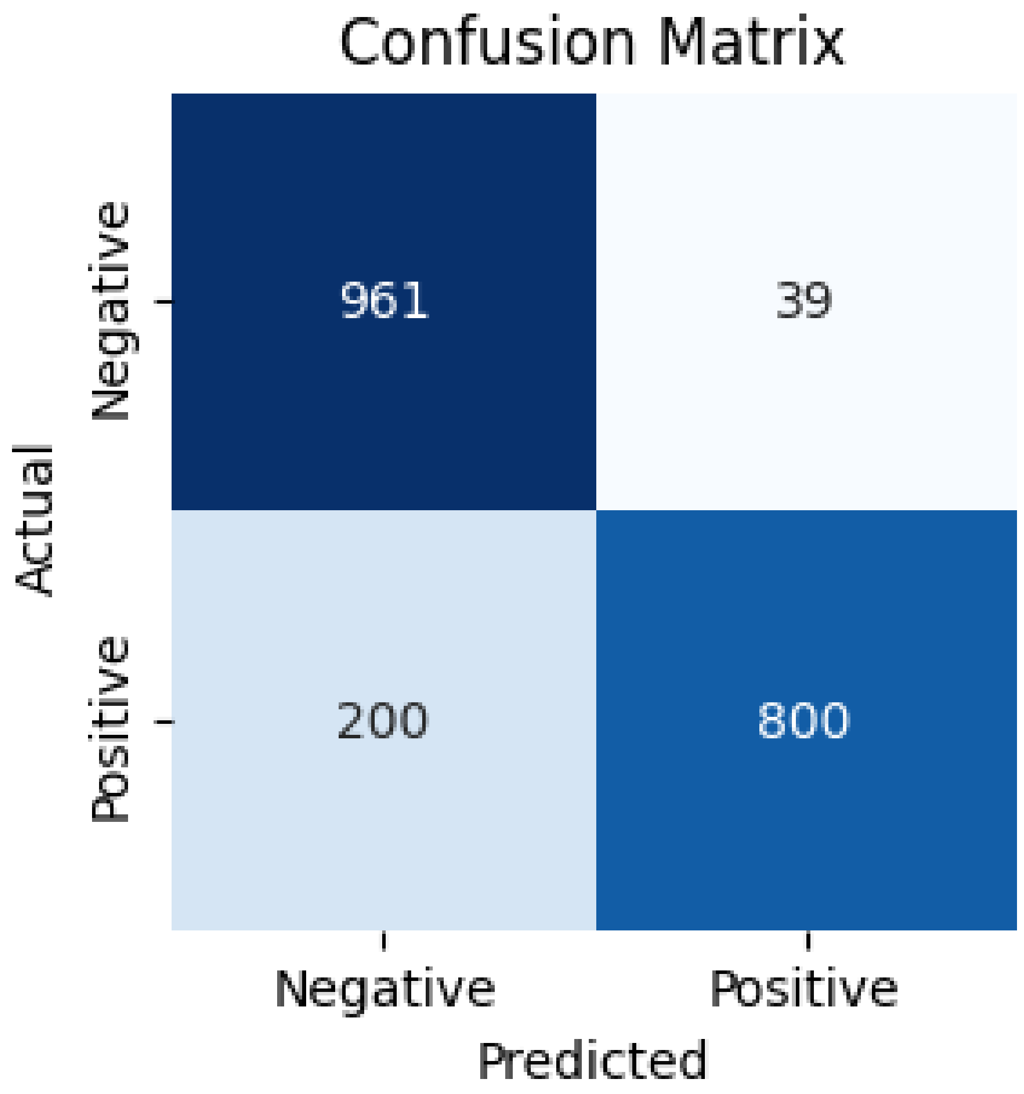
Flow chart for ML approaches



SVM

- Support vector machines (SVM)
- X-axis different feature engineering techniques
- Y-axis Accuracy scores

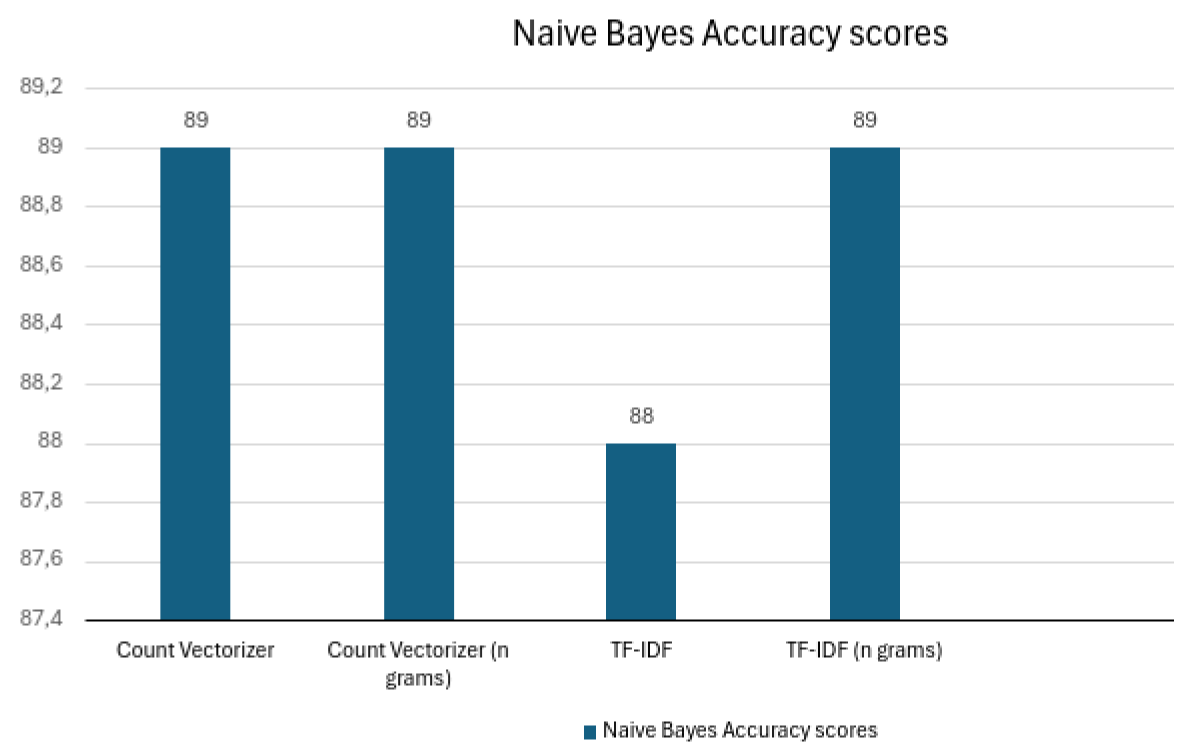




Analysis of SVM

- True Negatives : 961
- True Positive: 800
- False Positives: 39
- False Negatives: 200
- SVM has less False Positives

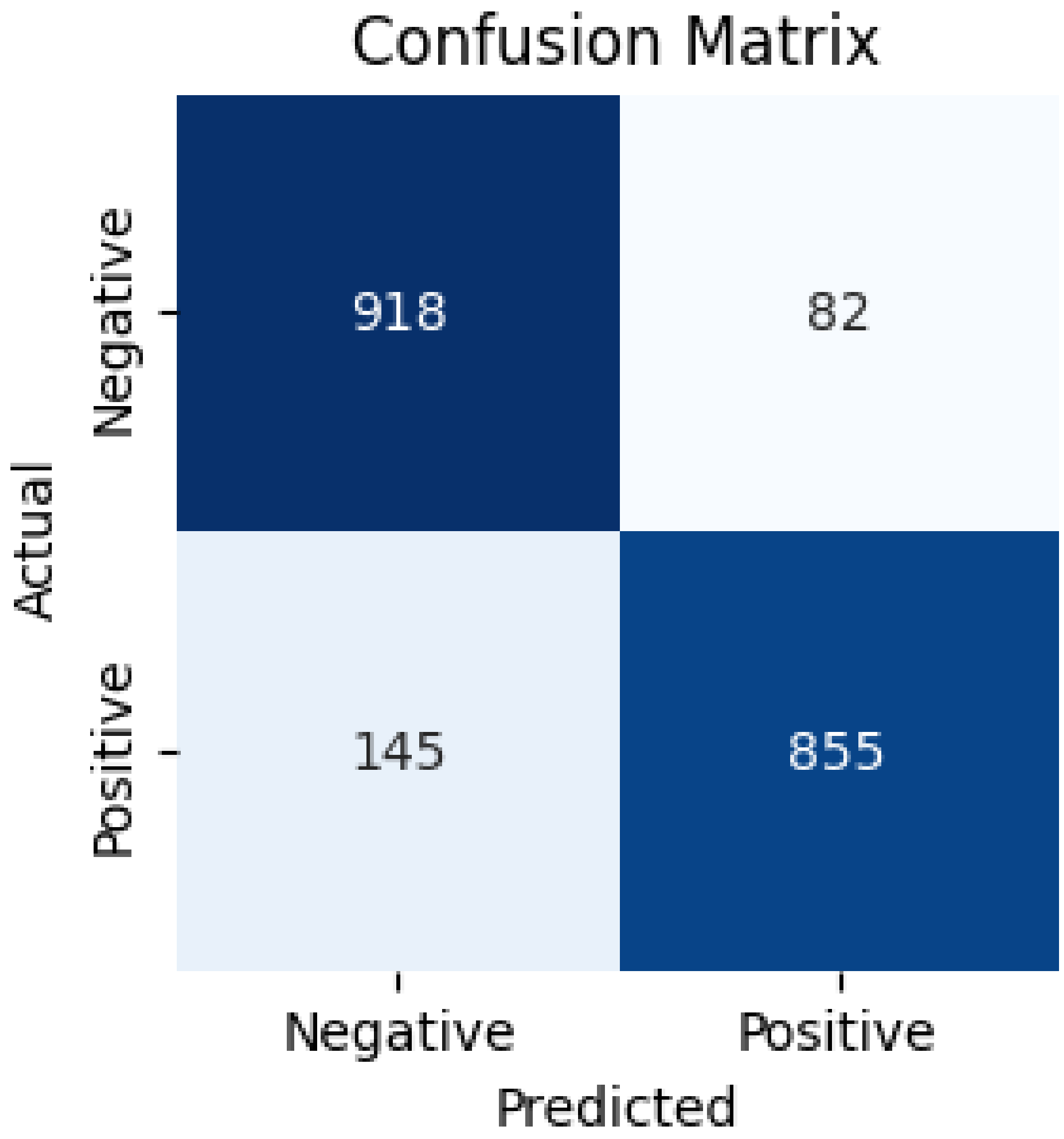
Naïve Bayes



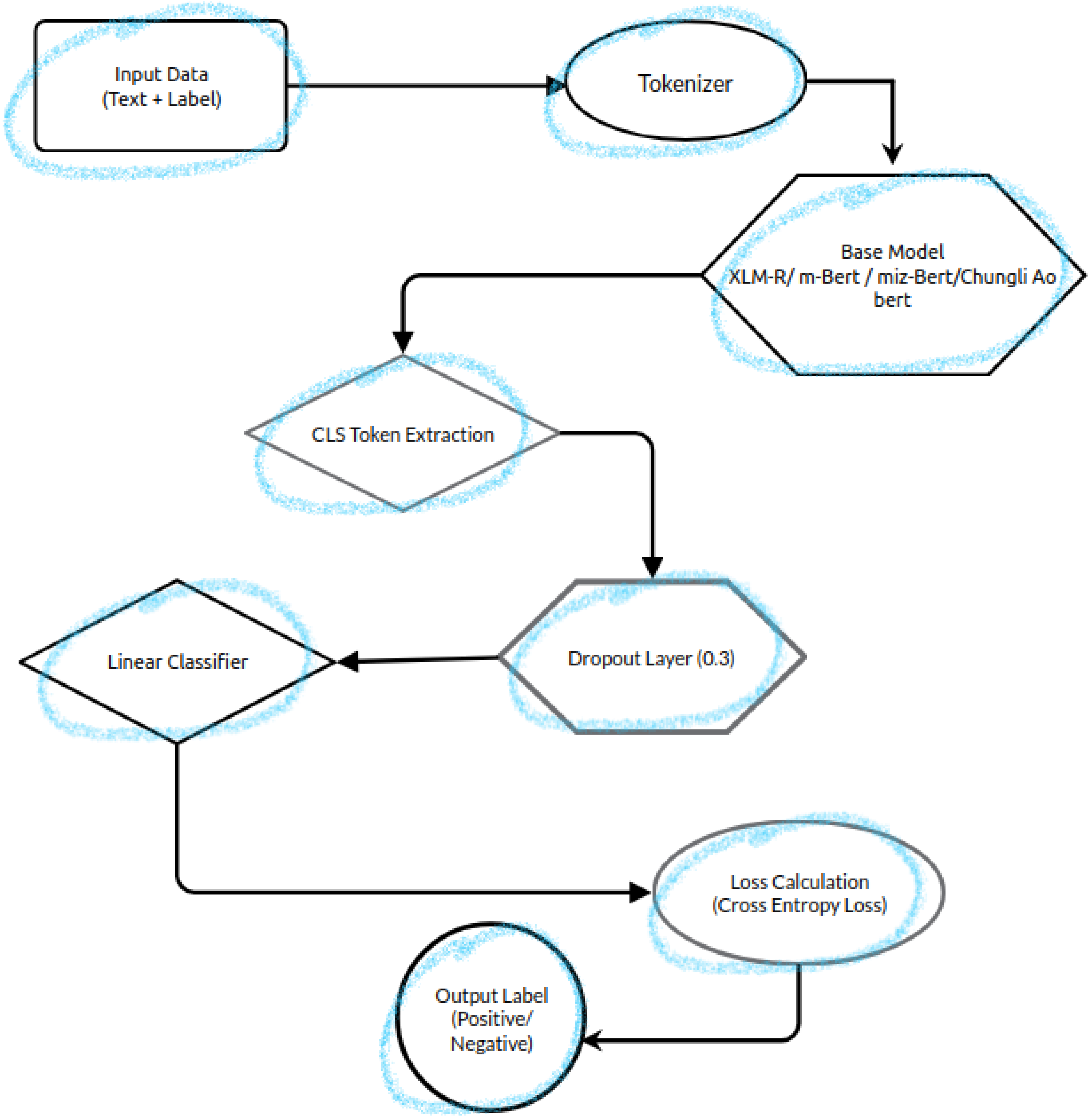
- Naïve Bayes, multinomial
- X-axis different feature engineering techniques
- Y-axis Accuracy scores

Analysis of Naïve bayes

- True Negatives :918
- True Positives: 855
- False Positives : 82
- False Negatives : 145
- Naïve Bayes has less False Negatives



Architecture & Pipeline



Fine-Tuning pipeline

Hyperparameter Settings

Data split	80:20 train:val
Learning rate	0.0001
# training epochs	3
Batch size	16
Evaluation	every 10 steps

Zero-Shot Experiments

+ fine-tuning on Chungli Ao sentiment data & experiments with Chungli Ao BERT

Experiments

- Fine-tuning all models with sentiment data & testing on Chungli Ao testset
  - Zero-shot for XLMR, MizBERT & m-BERT: models have no knowledge of Chungli Ao
  - Not Zero-shot for Chungli Ao BERT: model has knowledge of Chungli Ao
- Data used for fine-tuning:
  - Chungli Ao; English, German, Russian, Telugu, Mizo; All (minus Chungli Ao)

Experiments

- Mutiple runs:
  1. Early stopping & callback set to 3
  2. Early stopping & callback set to 10 → does extended training improve model performance?
  3. Run for error analysis (repeat of most successful experiment)

Results

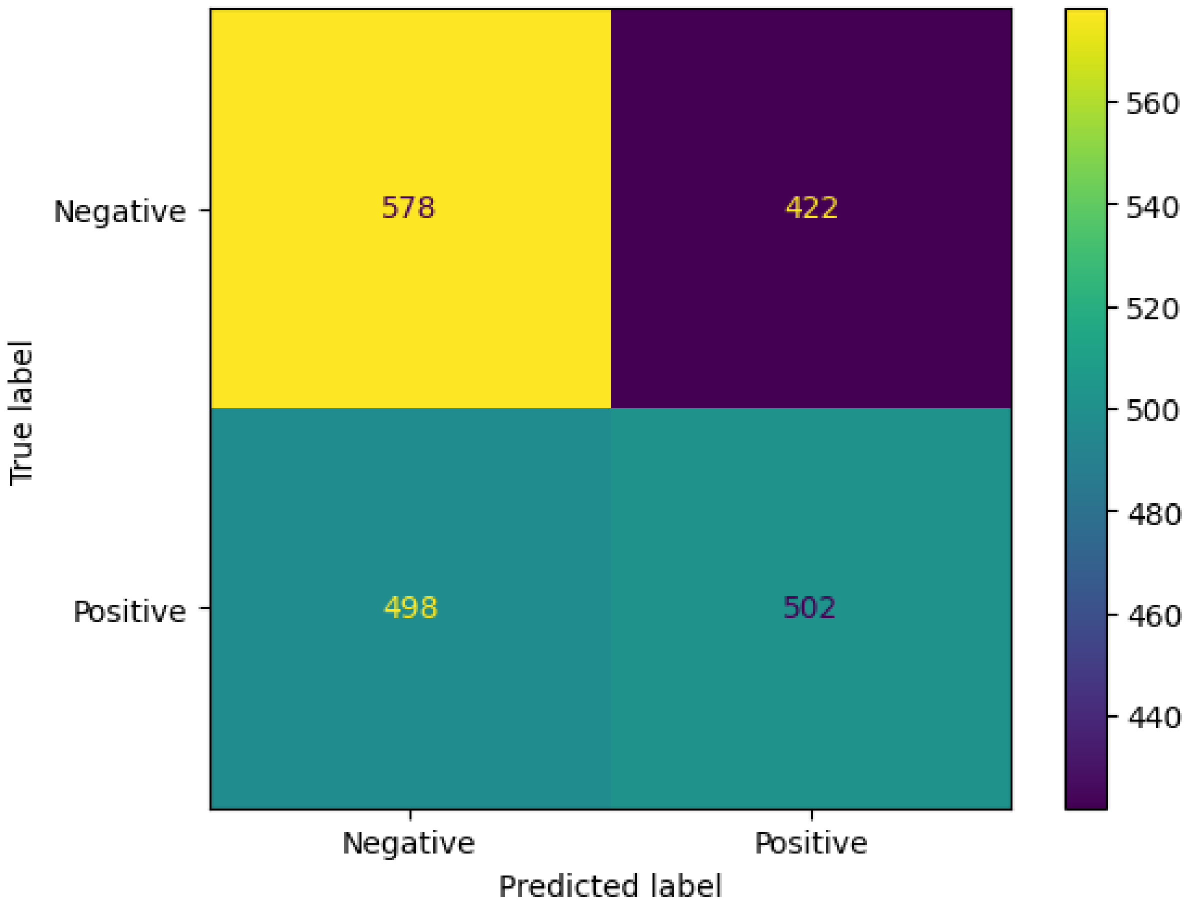
Model	Chungli Ao Accuracy (test, avg.)	Best accuracy (zero-shot, test, avg.)	Best run (zero-shot, test)	Accuracy of “all languages” (zero-shot, test)
XLMR				
M-BERT				
Miz-BERT				

Results

Model	Chungli Ao Accuracy (test, avg.)	Best accuracy (zero-shot, test, avg.)	Best run (zero-shot, test)	Accuracy of “all languages” (zero-shot, test)
XLMR	0.79	Telugu (0.55)	Telugu (0.57)	0.48
M-BERT				
Miz-BERT				

XLMR – Error Analysis

- Telugu train set with call-back set to 3
- Relatively even distribution of classes
- Not significantly better than chance

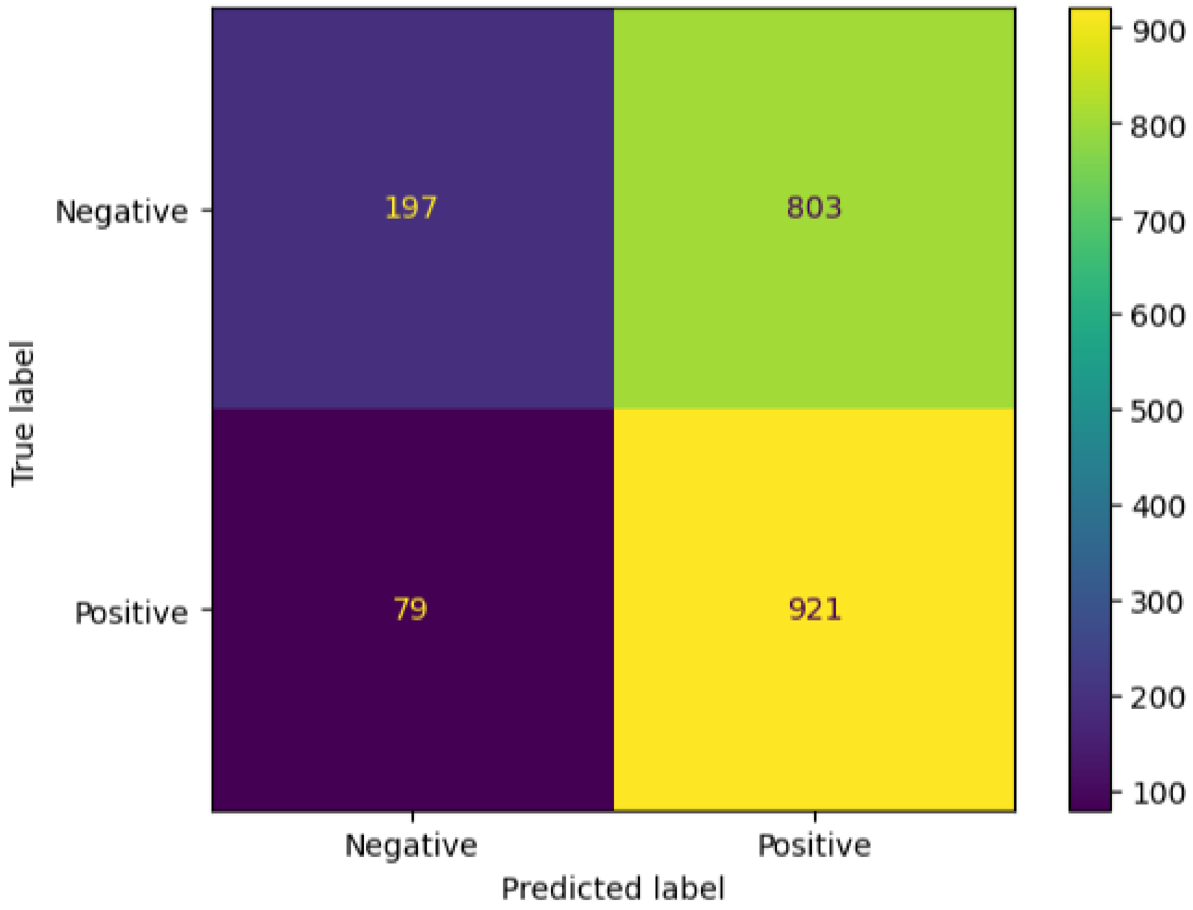


Results

Model	Chungli Ao Accuracy (test, avg.)	Best accuracy (zero-shot, test, avg.)	Best run (zero-shot, test)	Accuracy of “all languages” (zero-shot, test)
XLMR	0.79	Telugu (0.55)	Telugu (0.57)	0.48
M-BERT	0.77	Telugu (0.57) Mizo (0.57)	Telugu (0.62)	0.53
Miz-BERT				

M-BERT – Error Analysis

- Mizo train set with call-back set to 3
- Tendency for “positive” label leading to low accuracy (close to chance)

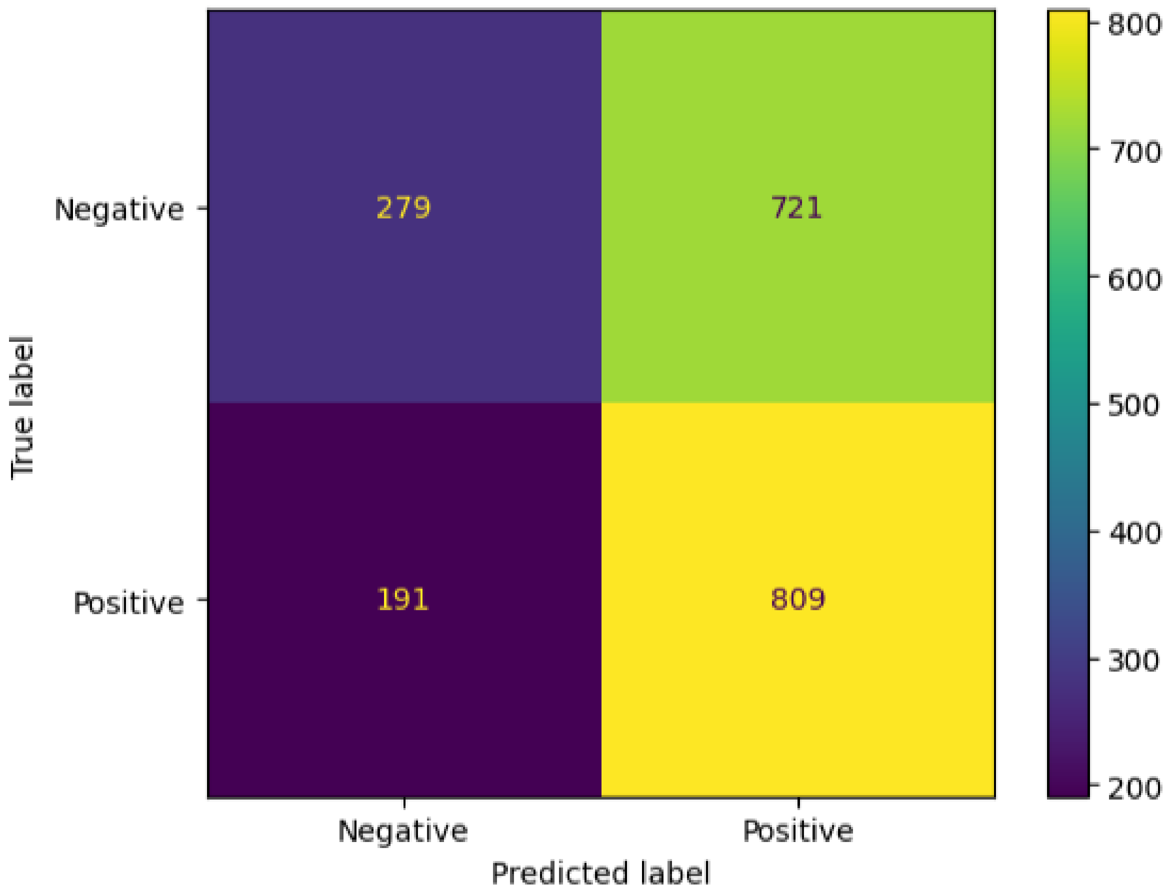


Results

Model	Chungli Ao Accuracy (test, avg.)	Best accuracy (zero-shot, test, avg.)	Best run (zero-shot, test)	Accuracy of "all languages" (zero-shot, test)
XLMR	0.79	Telugu (0.55)	Telugu (0.57)	0.48
M-BERT	0.77	Telugu (0.57) Mizo (0.57)	Telugu (0.62)	0.53
Miz-BERT	0.70	Mizo (0.56)	Telugu (0.57) Mizo (0.57)	0.45

Miz-BERT – Error Analysis

- Mizo train set with call-back set to 10
- Tendency for "positive" label leading to low accuracy (close to chance)



Zero-shot results

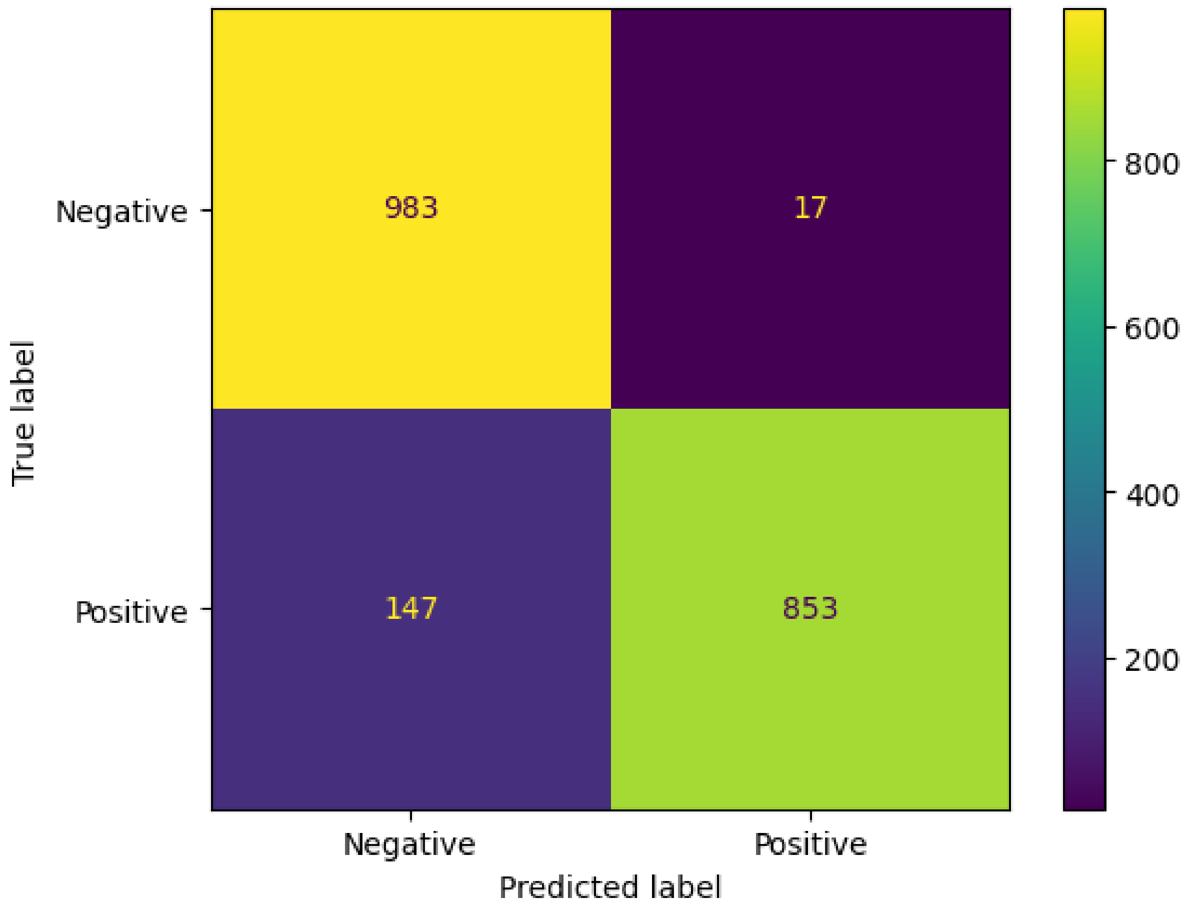
- Models benefit from fine-tuning on Chungli Ao → not zero-shot
- Within zero-shot experiments: mixed results
  - Models benefit most from fine-tuning on Telugu and / or Mizo
  - Models don’t generalize well; little to no cross-lingual transfer
  - Adding all languages does not help
- MizBERT doesn’t generalize well to Chungli Ao
- Significantly lower performance on zero-shot experiments compared to traditional ML methods

Chungli Ao BERT

Train set	Validation		Test Accuracy Scores		Mean Accuracy
	1st Run	2nd Run	1st Run	2nd Run	
Chungli Ao	0.99	0.94	0.68	0.92	**0.80**
Telugu	0.79	0.51	0.51	0.73	0.62
German	0.73	0.68	0.49	0.48	0.48
English	0.81	0.69	0.70	0.85	0.77
Mizo	0.93	0.98	0.73	0.87	**0.80**
Russian	0.50	0.50	0.61	0.86	0.73
All	0.50	0.49	0.53	0.50	0.51

Chungli Ao BERT – Error Analysis

- Chungli Ao train set with call-back set to 10
- Correctly classifies over 90% of all instances



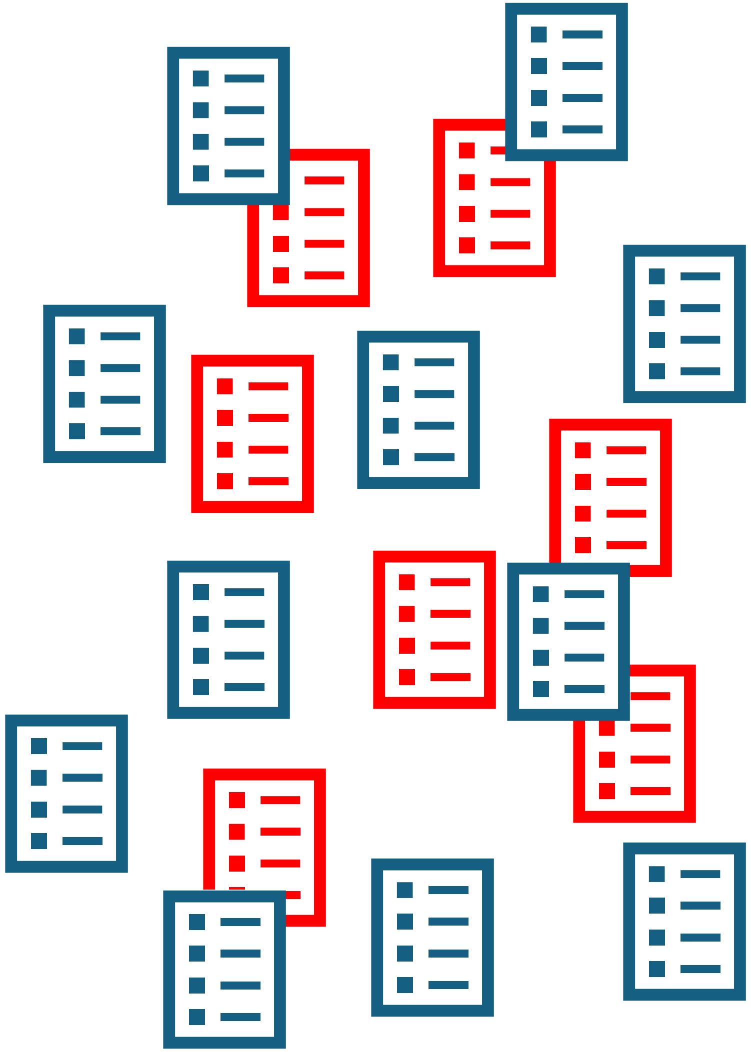
Chungli Ao BERT results

- Model fine-tuned on Chungli Ao significantly outperforms other models
- Highest score out of all models when using Chungli Ao train set
- Outperforms traditional ML methods on some runs
- Generalizes well on Mizo data
- Effectively reduces all classification errors → most effective method for Chungli Ao sentiment analysis

Multilingual Data Augmentation

Data Augmentation

- Methods to artificially increase training data
- “Monolingual” Data Augmentation (synonym replacement, subtree swapping, ...)
- Multilingual Data Augmentation: increase training data by adding data from a different language



Experiments

- Used Multilingual Data Augmentation for all models

1. Chungli Ao + 1 language
2. Chungli Ao + all languages

Research Questions

1. Does adding a related language (Mizo) help?
2. Does adding high-resource languages (e.g. English, German) help?
3. Do the models benefit from more data in general?

Results

Model	Language	Accuracy (Test)
XLMR	Mizo	0.68
M-BERT	English	0.72
Miz-BERT	German	0.69
Chungli Ao BERT	German	0.71

Results

- No clear pattern of which language(s) aided most
- Mizo helped in training, but not as much as expected
- Adding all languages did not help model performance
- Not possible to make definite statements

→ **statistical analysis**

There is an improvement in all metrics between the model trained on Chungli Ao and any one language (Telugu, German, etc.) and the model training on Chungli Ao and all languages

Is it statistically significant?

First we did the Friedman test to see if any model is significantly better than the others.

Friedman chi-squared = 6, df = 2, p-value = 0.04979

Anyway, we find that there is a significant difference

	Telugu	Mizo
Mizo	0.438	-
All	0.038	0.438

Nemenyi tells us which model is significantly different. We find that All is significantly better than Telugu, but not significantly better than Mizo. Mizo is not significantly better than Telugu.

We also conducted the Mann-Whitney U test to perform pairwise comparisons between Mizo and other training data

≥ Mann-Whitney is a non-parameter test to compare two independent groups

≥ We used Bonferroni correction to the p-values

We found no significant differences in performance (alpha = 0.05)

Comparison (Mizo vs.)	Metric	Mann-Whitney p-value	Bonferroni corrected p-value
Telugu	F1	0.18315	0.91575
Telugu	Loss	0.18315	0.91575
English	F1	0.20000	1.00000
English	Loss	0.34286	1.00000
German	F1	0.68571	1.00000
German	Loss	0.88571	1.00000
Russian	F1	0.48571	1.00000
Russian	Loss	0.88571	1.00000
All	F1	0.20000	1.00000
All	Loss	0.34286	1.00000

Pairwise Mann-Whitney U Test Results and Bonferroni Corrected p-values for F1 Score and Loss on Evaluation between Mizo and Other Languages

Back translation data augmentation

Back translation

- A two-step translation process:

Language 1 -> Language 2 -> Language 1

- Deep-translator library; Google Translate API

- 12 datasets based on three languages (English, Mizo, Telugu)

Experiments

- Yes, we cannot generate new data for Chungli Ao. But perhaps adding back translated data from other languages can improve performance?

- XLMR

Experiments

Train set	Accuracy (Train)	Accuracy (Test)
Mizo_Telugu_Mizo	0.9425	**0.5742**
Mizo_English_Mizo	0.9637	0.5284
Mizo_German_Mizo	0.9654	0.5389
Mizo_Russian_Mizo	0.9845	0.5360
Telugu_Mizo_Telugu	0.9824	0.6381
Telugu_English_Telugu	0.9904	**0.6405**
Telugu_German_Telugu	0.9955	0.6362
Telugu_Russian_Telugu	0.9203	0.5728
English_Mizo_English	0.9266	**0.6057**
English_Telugu_English	0.8904	0.5088
English_German_English	0.9570	0.5718
English_Russian_English	0.93166	0.5308

Accuracy scores for XLMR models with back translated data

Experiments

- What if we combine Chungli Ao with all the best results from previous experiments?

Test set	Accuracy (Train)	Accuracy (Test)
Chungli Ao + Telugu + Telugu_English_Telugu	0.9806	**0.6477**
Chungli Ao + Mizo + Telugu_English_Telugu	0.9563	0.6362

Results

- Using Telugu helps.
- Adding Mizo and English data also helped but wasn't as effective as just using Telugu.
- Adding different back translated datasets makes the model more variable, which makes performance worse sometimes

Pre-Training

Pre-Training

- Data
- Method
- Experiments

Data

- Sourced Chungli Ao Newspaper Tir Yimyim
  - From April 2021 to Febuary 2022
- Raw text extracted from PDF files

Tir Yimyim

www.tiryimyim.in

Regd. No. RNI. NAGAAO/ 2004 / 13113. Postal-NE/RN-717. e-mail : tiryimyim@gmail.com

TAPAK 7 Yangia anidakangma : Tir Yimyim tiryimyim@aolima tir yimyim TAPAK 10

Oxygen agi anüngba mapang PM-i CM tem den senden amen

Tokyo Olympics nung Neymar asayadaktsüner: Andre

VOL. XVIII NO. 189 (ADOK 189) DIMAPUR KÜPTOKNÜ (SATURDAY) RONGCHII (APRIL) 24, 2021 ₹ 5.00

Dimapur nung Holotoli School shibangtsür

Dimapur, April 23 (TYO): Dimapur nung Holotoli School, Padumpukhuri shibangtsür ta Hokolbarnü Chief Medical Officer, Dimapur Dr Mereninla Senlem-i metetdaktsüogo.

Iba osang ya CMO, Dimapur-i Deputy Commissioner, Commissioner of Police aser Holotoli School kibur dangi züluba shidi ka nung metetdaktsü.

India nung COVID-19 azüoktsü tasak: WHO

New Delhi, April 23 (Agencies): India nung COVID-19 menatepba azüoktsü kanga tasak asütsü ta Hokolbarnü World Health Organization (WHO)-i metetdaktsüogo.

India nung COVID-19 kanga putetba atema tebilemtsü WHO Emergencies director Mike Ryan-i "India nung COVID-19 menatepba azüoktsü atema senzüsenbongba noktangtsüla" ta ashi.

"India nung menatepa aoba azüoktsübaji kanga tasak lir. Asenoki wara azüoktsü atema nisungtem meyoktepba ajema kümdaktsütsüla. India sorkari iba mapaji inyaktsüla" ta Ryan-isa ashi.

Hokolbarnü India nung nisung 3,32,730 dak COVID-19 putet. Tang linük nung nisung 1,62,63,695 dak putetogo, ta Ministry of Health and Family Welfare-i metetdaktsü.

COVID-19 Nagaland: Nisung 89 dak putet; 95 süogo

Kohima, April 23 (TYO): Hokolbarnü Nagaland nung nisung 89 dak COVID-19 aliba putet aser parnok züngsema nisung ajak agi 12,889 kümogo. Külen, tanü COVID-19 agi shiranger ka asüba züngsema tasür ajak agi nisung 75 kümogo.

"Tanü nisung 89 dak COVID-19 putet. Parnokji Dimapur nung nisung 85 aser Kohima nung 5 lir. Ano, Kohima nung shiranger ka taneptsü angu" ta Health & Family Welfare Minister, S Pangnyu Phom-i metetdaktsü.

Nisung 12,889 rongnungi 12,117 tashi taneptsü nguogo aser nübo aliba osang jangjatepogo aser nüngdakba ajiteta inyaktsü, ta paisa ashi.

Sorkar aser department-i iba wara azüoktsü atema akokba tashi mapa inyaktsü anungji nübortemi tebilemtsü abenba senso kaka onok den yariteptsü ayongzüker, ta Pangnyu-isa ashi.

"Nagaland ung COVID-19 nungi taneptsü angur 94.01% dang kümogo," ta State Nodal Officer for Integrated Disease Surveillance Programme, Dr Nyanthung Kikon-i ashi.

"Tuensang district nung COVID-19 agi shiranger ka tera timtema oxygen agidar aser 4 kanga mejungi shiranga oxygen agüja anepaludar" ta paisa shisem.

Nagaland nung COVID-19 alitsü akok ta temolung melemi bilemba sample 1,42,528 tendangogo. RT-PCR ajanga 77,149, TrueNat ajanga 37,877 aser Rapid Antigen Test ajanga 27,503 tendang, ta Dr Kikon-isa ashi.

Külen, Brihostibar tashi nung Nagaland nung nisung 1,41,406 nem covishield indang dose 1,77,549 agütsüogo ta State Immunization Officer Dr Ritu Thurr-ia metetdaktsü.

Vaccine agirtemji frontline

Pre-Processing

1

Split raw text into sentences using regular expressions

2

Remove sentences containing less than 4 words

3

Remove sentences containing noise (URLs, etc.)

Final Dataset

Our Dataset

- Training set
  - 38k sentences
  - 827k words
- Validation set
  - 9.5k sentences
  - 206k words

Other Datasets

- BookCorpus
  - Used by original BERT
  - 800M words
- Mizo News corpus
  - Used by MizBERT
  - 72M words
  - 2M sentences

Method

- Language Adaptive Pre-Training
  - o Additional training to adapt model to new language
- Models
  - o mBERT
  - o XLM-RoBERTa
  - o MizBERT
  - o (ChungliAo-BERT)

Method

- Masked Language Modelling (MLM)
- Basic Idea
    - Mask words (tokens) in sequence with some probability (usually 0.15)
    - Model predicts original words
    - Classification Task
- Gain language understanding without need for labelled data
- Use Pre-Trained Models for downstream tasks e.g. sentiment analysis

Masked Language Modelling

Original Sentence
Pa ya Mon district nungi liasü.
Tokenized Sentence
'Pa', 'ya', 'Mon', 'district', 'nun', '##gi', 'li', '##as', '##ü', '.'
Masked Sentence
'Pa', 'ya', [MASK], 'district', 'nun', '##gi', 'li', [MASK], '##ü', '.'

Hyperparameter Search

Model	Batch Size	Learning Rate
mBERT	16	1e-4
MizBERT	8	1e-4
XLM-RoBERTa	16	1e-4

Best Hyperparamerters for minimizing validation Loss

Pre-Training the Models

New Models

- Chungli-Ao-mBERT
- Chungli-Ao-MizBERT
- Chungli-Ao-XLM-RoBERTa

- → All models made available on HuggingFace

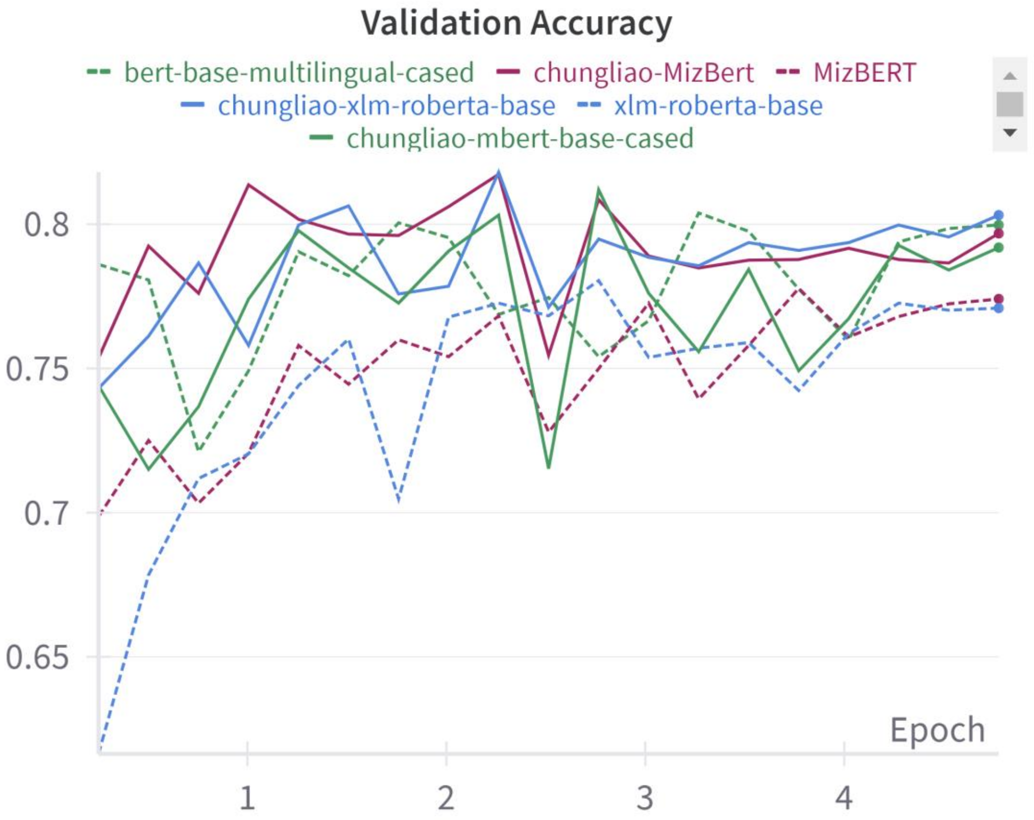
What effect did Pre-Training have?

Pre-Training Vs No Pre-Training

- Fine-tune models on sentiment analysis
- Use lower and higher learning rate
  - $5e^{-5}$ (lower)
  - $1e^{-4}$ (higher)
- 5 epochs
- Train on entire training set
- Test set as validation set
- Compute validation metrics 4 times per epoch

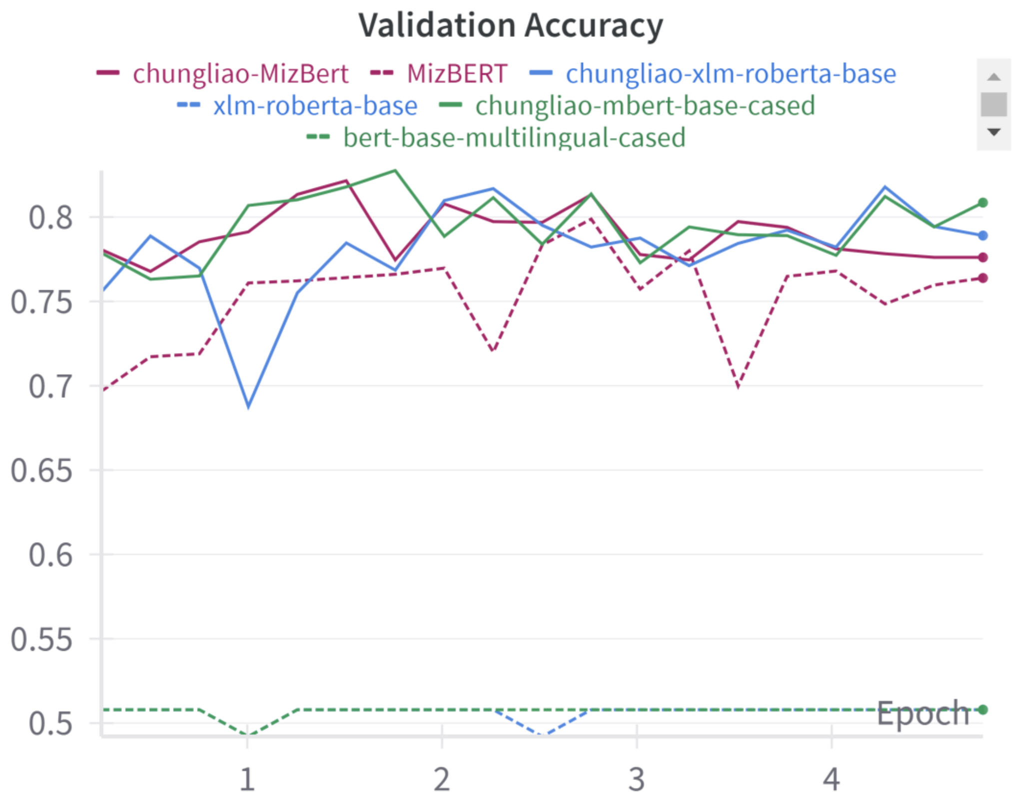
Pre-Training Vs No Pre-Training

- Lower learning rate
- Overall stronger by Chungli-Ao models
- Only mBERT base is competitive with Chungli-Ao models

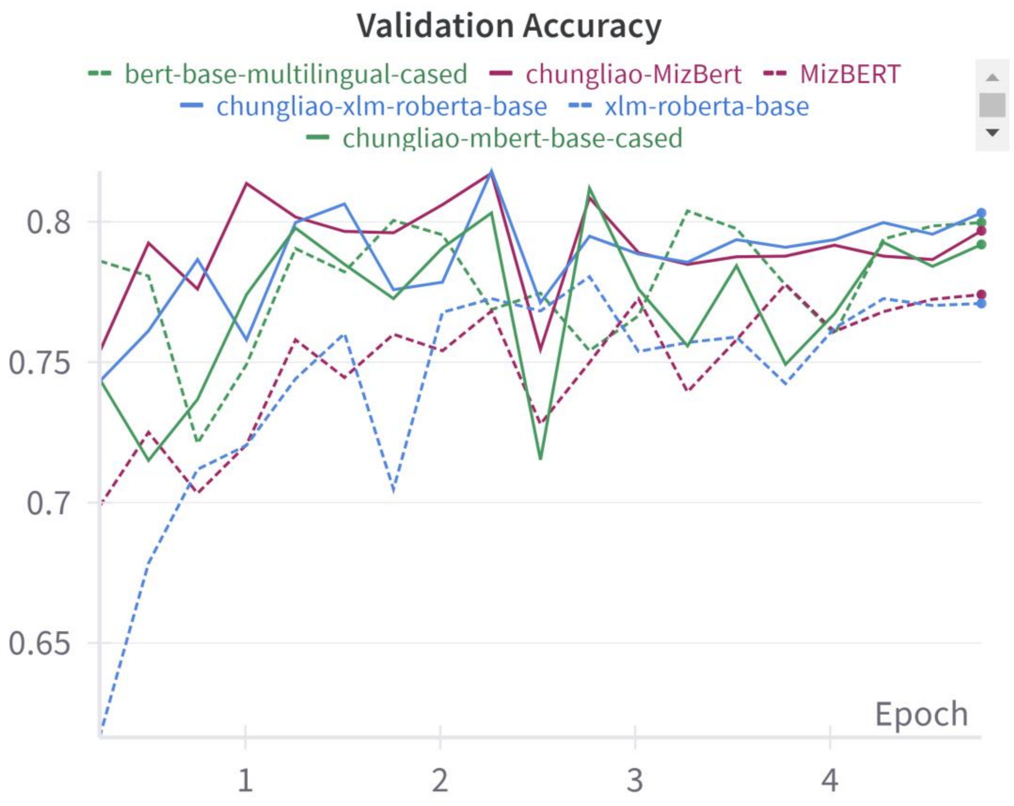


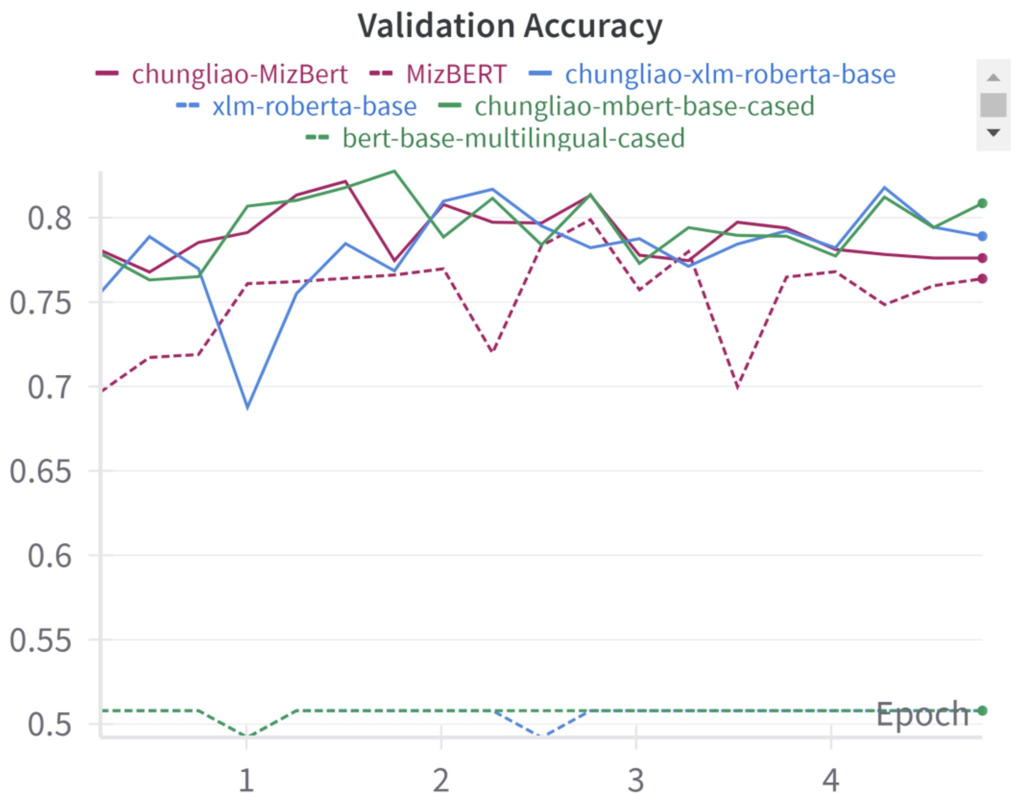
Pre-Training Vs No Pre-Training

- Higher learning rate
- Stronger performance by Chungli-Ao models
- Multilingual models fail to learn
- MizBERT can still learn



Pre-Training Vs No Pre-Training





Pre-Training Vs No Pre-Training

Model	$5e^{-5}$	$1e^{-4}$
Chungli-Ao-mBERT	0.81	**0.83**
mBERT	0.80	0.51
Chungli-Ao-MizBERT	0.82	0.82
MizBERT	0.77	0.79
Chungli-Ao-XLM-RoBERTa	0.82	0.82
XLM-RoBERTa	0.78	0.51

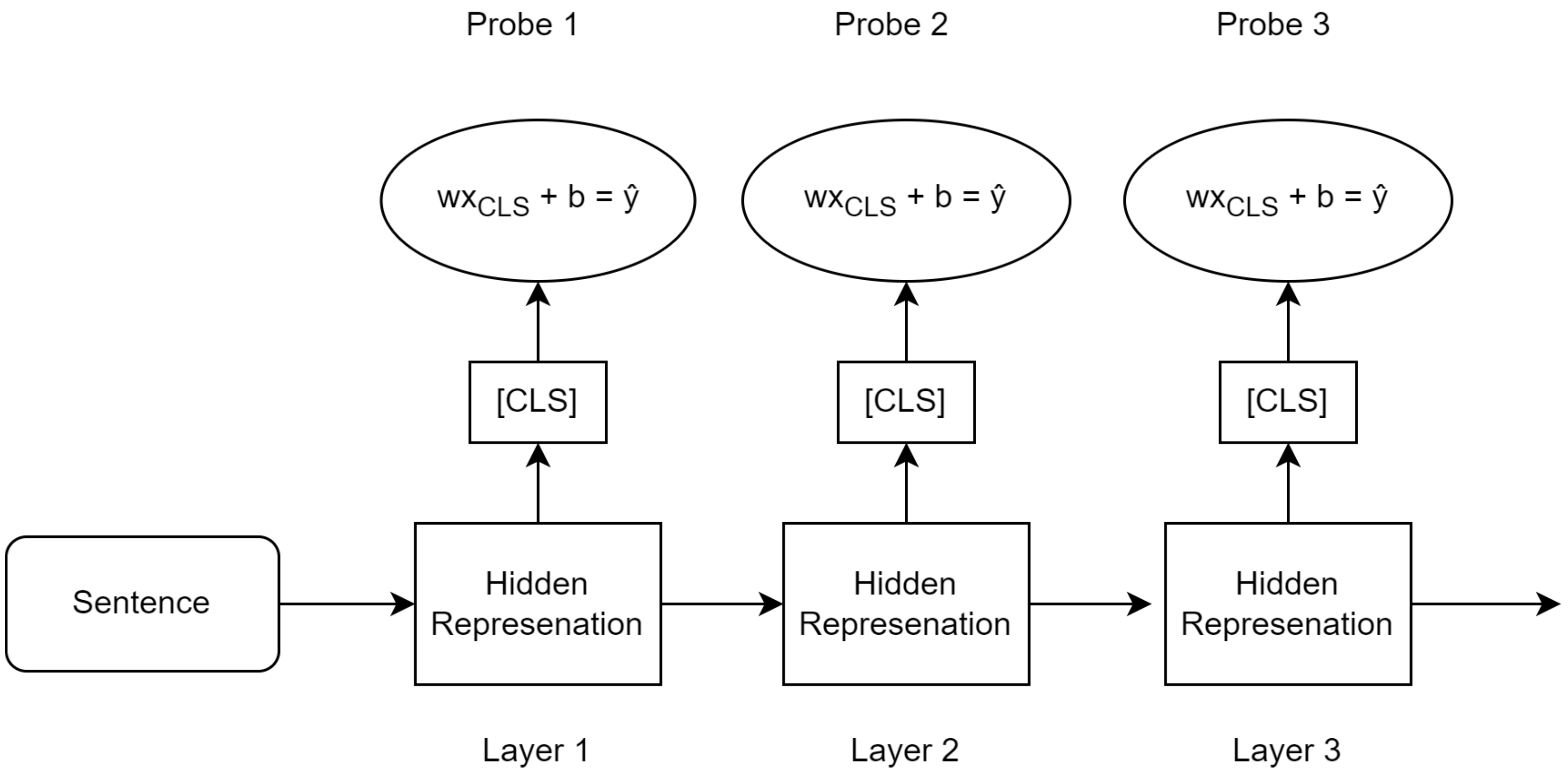
Best Validation Accuracy with a lower and higher learning rate

What caused this result?

Probing

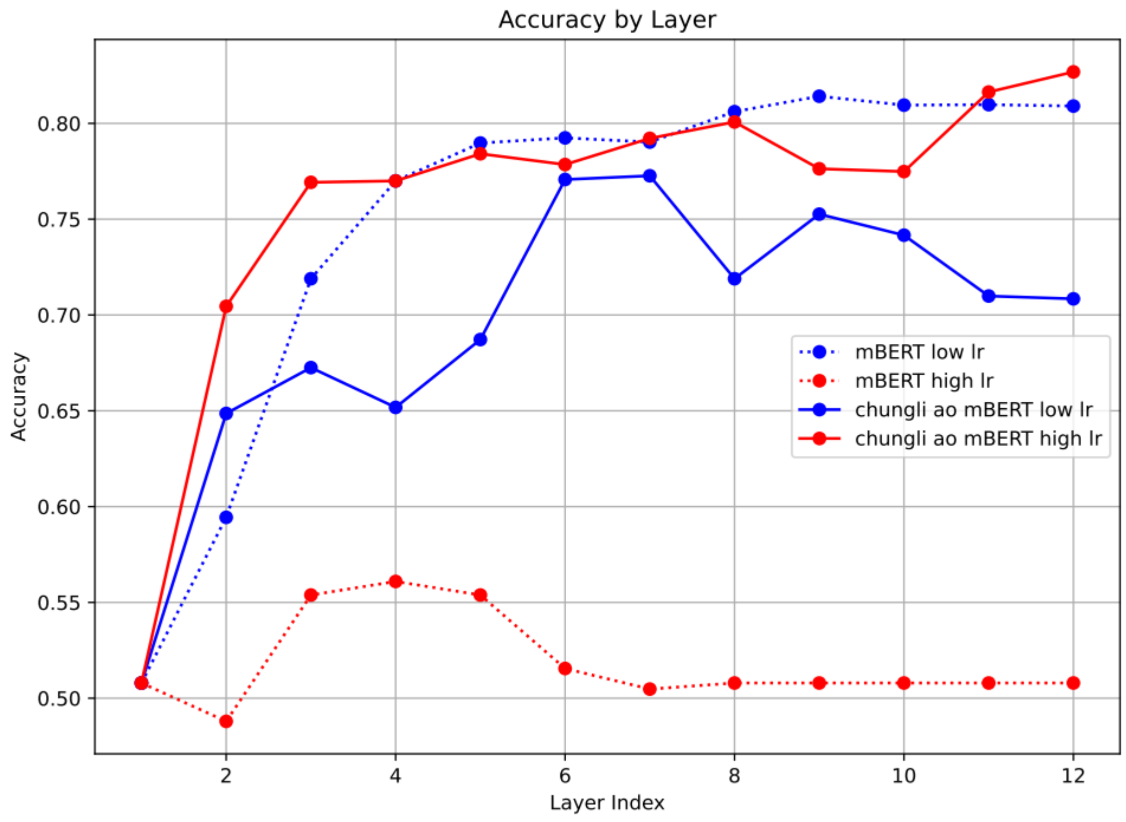
- Probe Chungli-Ao-mBERT and base mBERT
  - Lower and higher learning rate
  - Use parameters with highest accuracy for each model
- Freeze Parameters
- Train linear probes on each layer
  - Input: [CLS] token representation
  - Output: Sentiment Prediction (Probability)

Probing

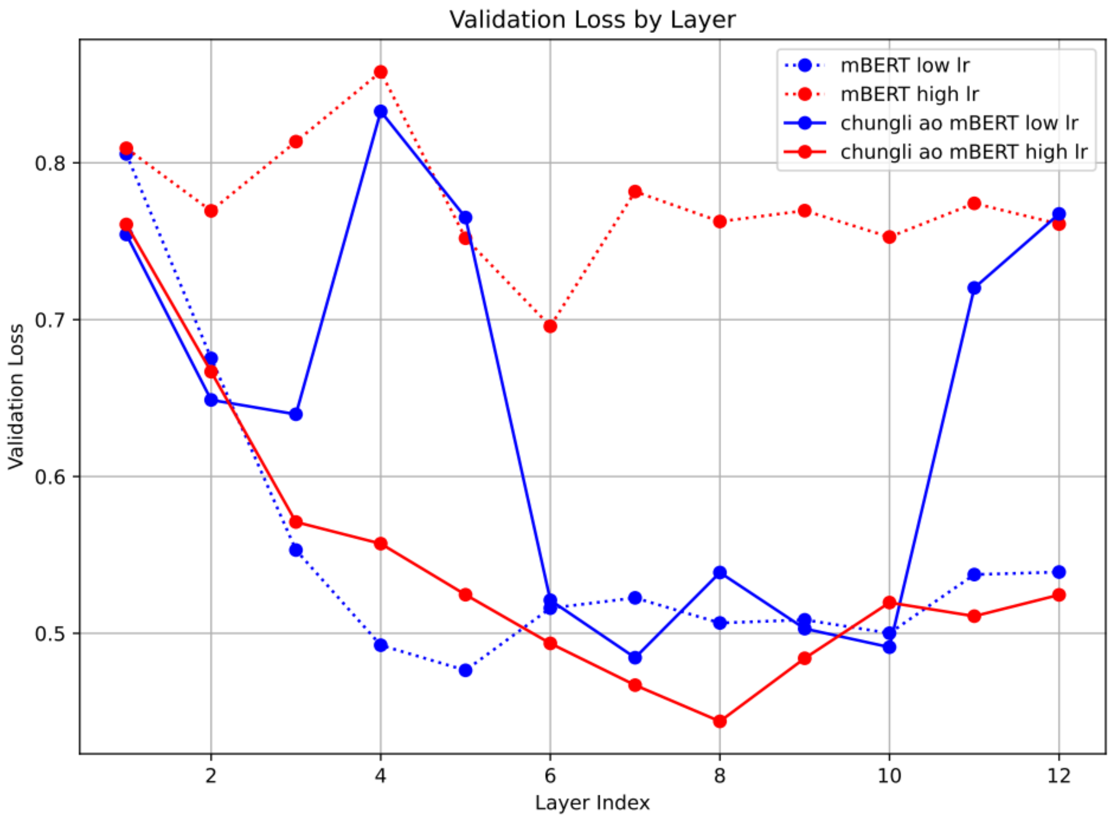


Probe 1
Probe 2
Probe 3
$wx_{CLS} + b = \hat{y}$
$wx_{CLS} + b = \hat{y}$
$wx_{CLS} + b = \hat{y}$
[CLS]
[CLS]
[CLS]
Sentence
Hidden Represenation
Hidden Represenation
Hidden Represenation
Layer 1
Layer 2
Layer 3

Probing



Accuracy by Layer
Accuracy
Layer Index
0.80
0.75
0.70
0.65
0.60
0.55
0.50
2
4
6
8
10
12
mBERT low lr
mBERT high lr
chungli ao mBERT low lr
chungli ao mBERT high lr



Validation Loss by Layer
Validation Loss
Layer Index
0.8
0.7
0.6
0.5
2
4
6
8
10
12
mBERT low lr
mBERT high lr
chungli ao mBERT low lr
chungli ao mBERT high lr

Probing

- Accuracy by layer
  - Similar graph for low lr mBERT and high lr Chungli-Ao-mBERT
  - Less smooth graph for low lr Chungli-Ao-mBERT
  - Suprising jump in accurcacy in earlier layers of high lr mBERT
- Loss by layer
  - Similar graph for low lr mBERT and high lr Chungli-Ao-mBERT
  - Unstable graph for low lr Chungli-Ao-mBERT

Pre-Training final scores

- Use training, validation and test set
- 3 Callbacks
- Worse performance than highly optimized base models

Models	Val Acc	Test Acc
Chungli-Ao-MBERT	0.95	**0.81**
Chungli-Ao-MizBERT	**0.96**	0.77
Chungli-Ao-XLM-RoBERTa	**0.96**	0.8

Pre-Training Takeaways

- Not as strong performance as highly optimized base models and Chungli-Ao-BERT
- Evidence that additional pre-training makes models more robust to learning rate
    - Higher learning rate may be more optimal (for Chungli-Ao-mBERT)

→ Better results with more hyperparameter tuning may be possible

Limitations

- Biases in the datasets
- Limited availability of datasets
- Translationese in the Chungli Ao
- We explored only two ML approaches
    - o Random forest, KNN ...?
- Small data set

Future Work

- Adapter Fusion
  - Freeze base model
  - Fine-tune smaller adapter model placed on top of base model
- Create more diverse datasets in Chungli Ao
  - More domains for sentiment analysis
  - More task e.g. POS-tagging or NER

Future Work

- Pre-training on more data
  - Chungli Ao bible
- Exploring the tokenizers
  - Adapting multilingual model tokenizer vocabulary
  - Efficient token embedding intialization
- Generating more data through Machine Translation
  - Train on MT system on parallel bible corpus
  - Less costly than manual translations

Conclusion

- Current multilingual models perform poorly in zero-shot experiments for Chungli Ao
- Fine-tuning multilingual models helps in generalization.