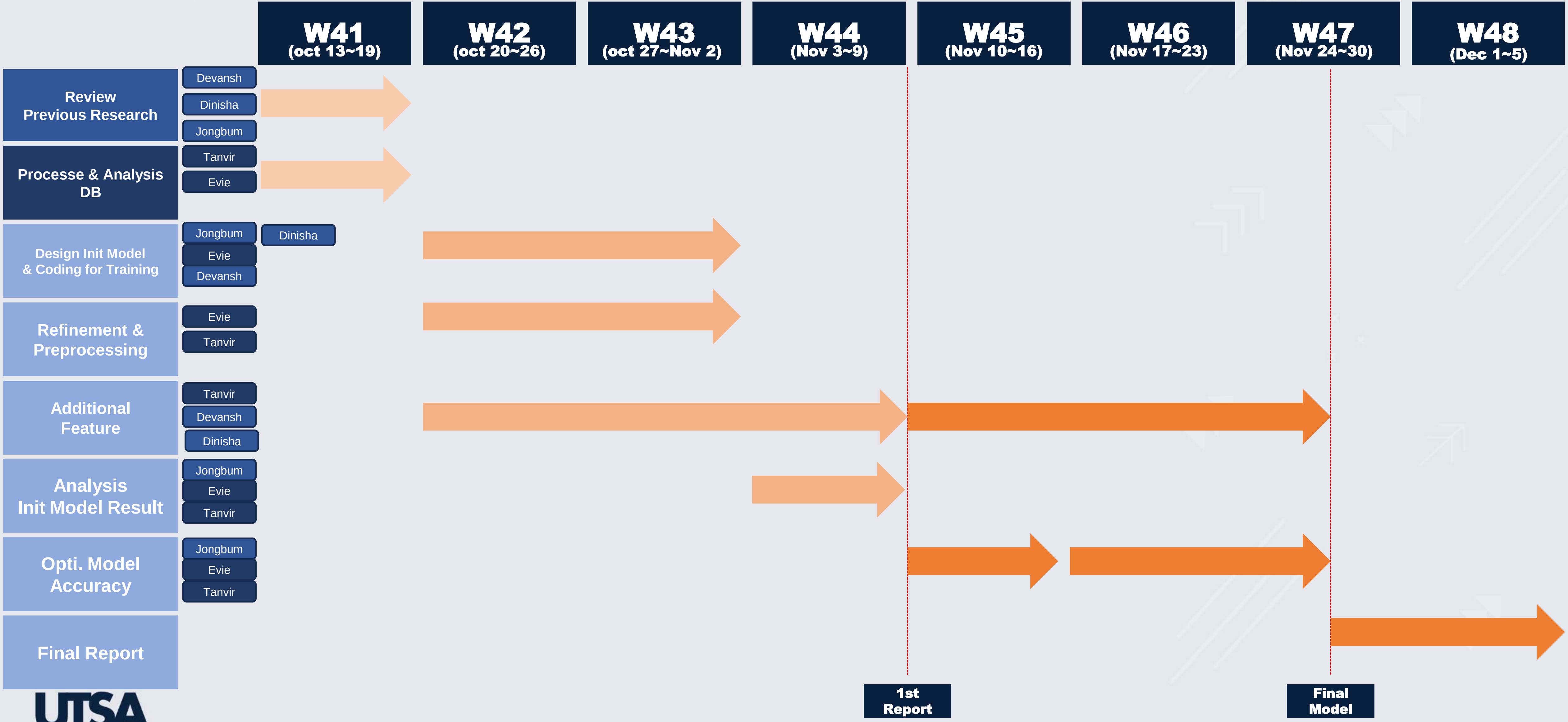
Project Timeline



		W41 (oct 13~19)	W42 (oct 20~26)	W43 (oct 27~Nov 2)	W44 (Nov 3~9)	W45 (Nov 10~16)	W46 (Nov 17~23)	W47 (Nov 24~30)	W48 (Dec 1~5)
Review Previous Research	Devansh Dinisha Jongbum	→							
Processe & Analysis DB	Tanvir Evie	→							
Design Init Model & Coding for Training	Jongbum Evie Devansh	Dinisha	→	→					
Refinement & Preprocessing	Evie Tanvir		→	→					
Additional Feature	Tanvir Devansh Dinisha		→	→	→	→	→	→	
Analysis Init Model Result	Jongbum Evie Tanvir				→				
Opti. Model Accuracy	Jongbum Evie Tanvir					→	→	→	
Final Report								→	→

1st Report

Final Model

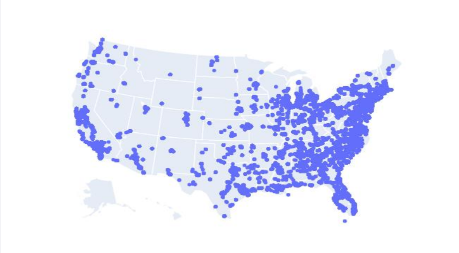
Predict Geolocation based on Social Media Text

Fall 2024 – AI Practicum

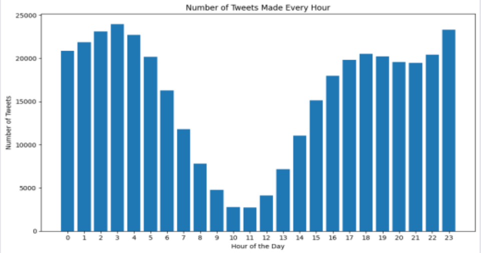
Social Media Data

- **Statistical DB analysis**
  - Geographical, Tweet Time Distribution

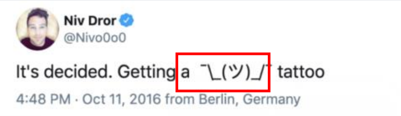
- **Noise and Outlier analysis & Cleaning**
  - Remove Emoji, punctuations
  - Word Lemmatization
  - Non-US DB Translation



< Geographical Analysis>



< Tweet time Analysis>



Niv Dror
@Nivo0o0

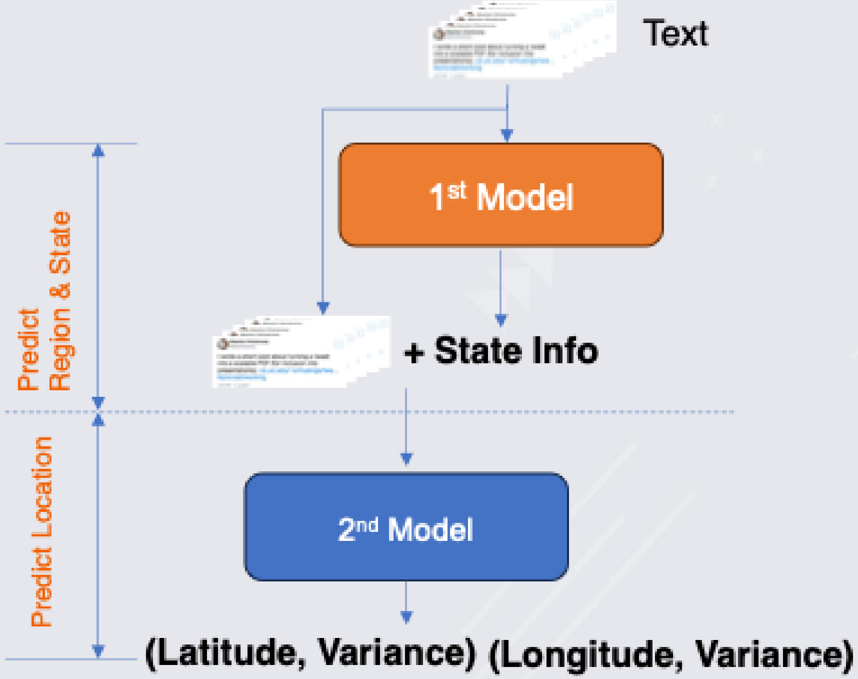
It's decided. Getting a ¯_(ツ)_/¯ tattoo

4:48 PM · Oct 11, 2016 from Berlin, Germany

It's decided. Getting a tattoo.

Solution Design

- **Pipeline of Two Cascading Models**
  - **1st Model: TD-IDF and Linear SVC**
  - **2nd Model: Predict Location**



Text

1st Model

+ State Info

2nd Model

(Latitude, Variance) (Longitude, Variance)

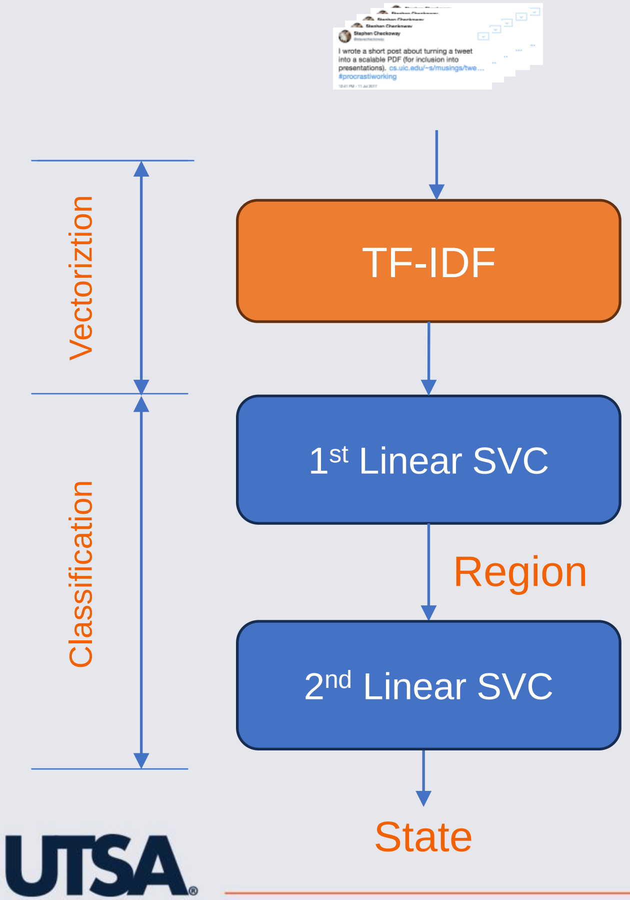
Predict Region & State

Predict Location

Model Design & Training – Region and States

Fall 2024 – AI Practicum

- Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency)
- Classification: Two Linear SVC Model (Region and State)



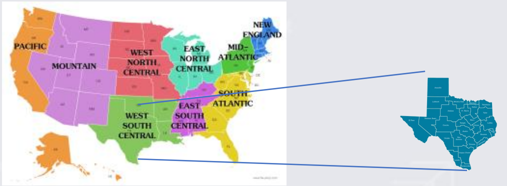
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe... #procrastiworking
Vectoriztion
Classification
TF-IDF
1st Linear SVC
Region
2nd Linear SVC
State

- **TF-IDF Machine Learning Model**
  - Statistical formula to Convert text doc into vectors

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right)$$

- $w_{ij} = weight\ of\ word\ i\ for\ documnet\ j$
- $tf_{i,j} = \frac{occurence\ of\ i\ in\ document\ j}{Total\ num\ of\ word\ in\ j}$
- $df_i = num\ of\ documnet\ containing\ word\ i$
- $N = total\ num\ of\ doctumnet$

- **Linear SVC (Support Vector Classifier)**
  - 1st Model: Text → 9 Regions
  - 2nd Model: 9 Regions → State



< 1st SVC Result: Region> < 2nd SVC Result: State >

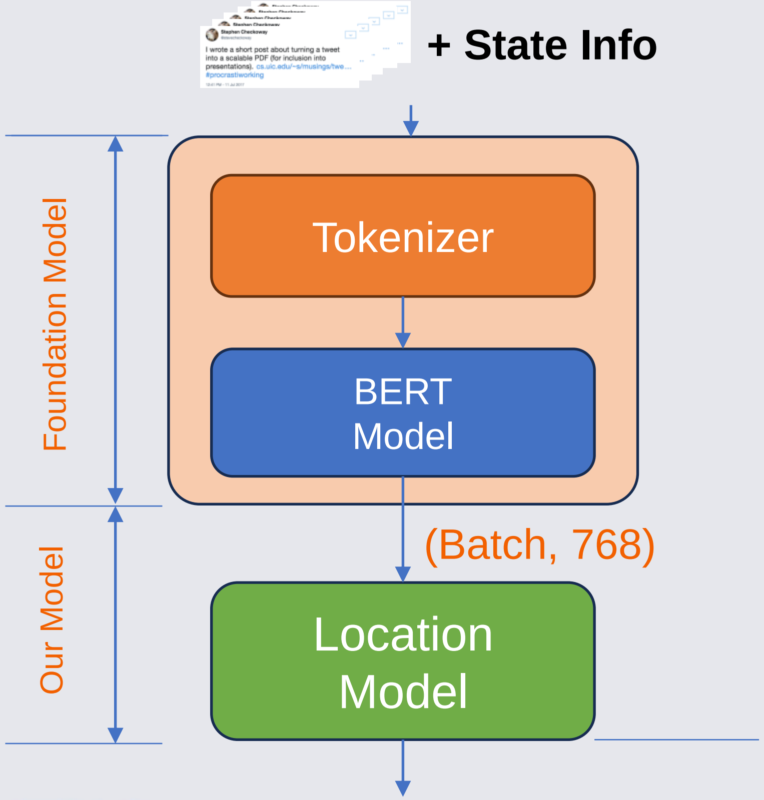
- **Test Result**
  - Precision: 98%, Recall: 98% f1-score: 98%
  - Extract Common Words by Region

Region	Common words			
Mountain	asu	phx	arizona	kubball
NW Central	geeksquad	hum	sherroncollins	kubball
SW Central	texas	dallas	sse	houston

Model Design & Training – Location Model

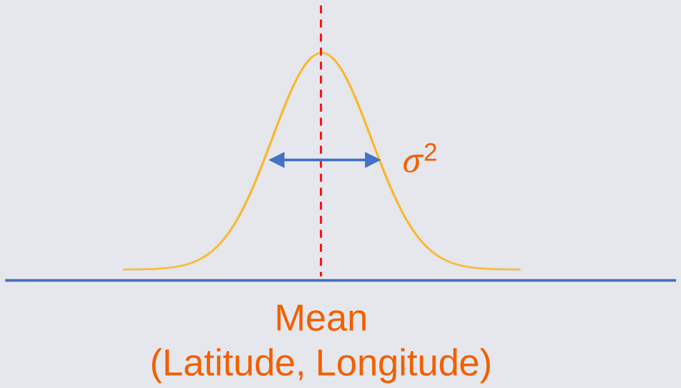
Fall 2024 – AI Practicum

- Foundation Model: Simplified BERT Model (HuggingFace)
- Location Model: Convert Vector to Location with Uncertainty



(Mean, Variation) of Location

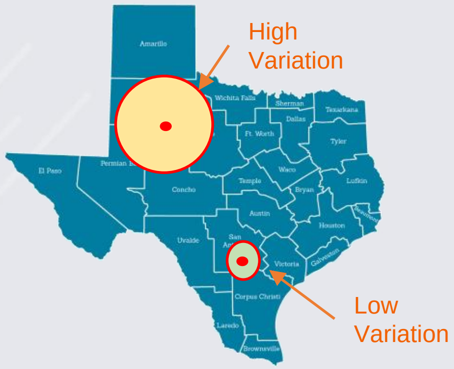
- Model Output: **Mean & Uncertainty(Variation)[1]** of Location



$$Loss = \frac{1}{2\sigma^2}\|y - f(x)\|^2 + \log\sigma$$

- $y: GT$
- $f(x): Predicted\ Value$
- $\sigma^2: Uncertainty\ (variation)$

- **Test Result**
  - Avg Error: 168.42 km
  - **High Variation Avg Error: 401.66 km**
    (e.g "I don't feel so good today!", "I am home now.")
  - **Low Variation Avg Error: 34.88km**
    (e.g "I am in the river walk", "New york, New york!!")

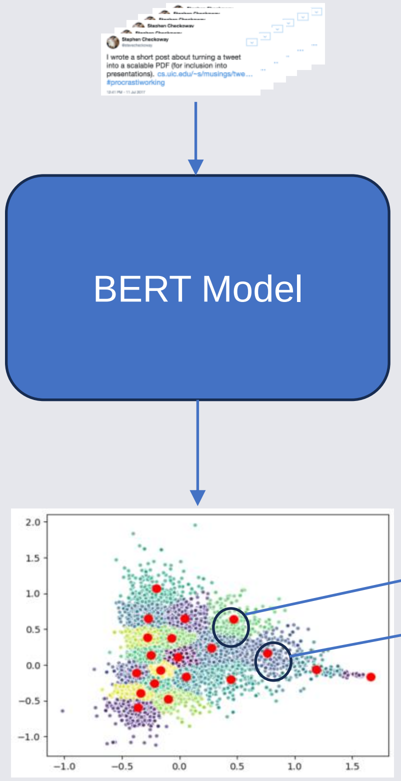


[1] **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics**

Additional Feature – Recommendation

Fall 2024 – AI Practicum

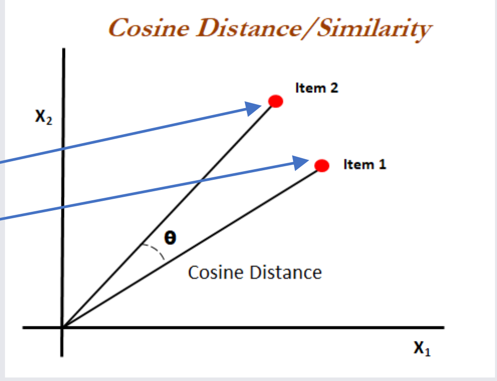
- Vectorized tweet(text) Clustering
  - Generate sentence embeddings by using pre-trained BERT model
  - Sentence Embedding with Cosine Distance
  - Find the cosine distance with the other users post and **recommend top 5 related posts**

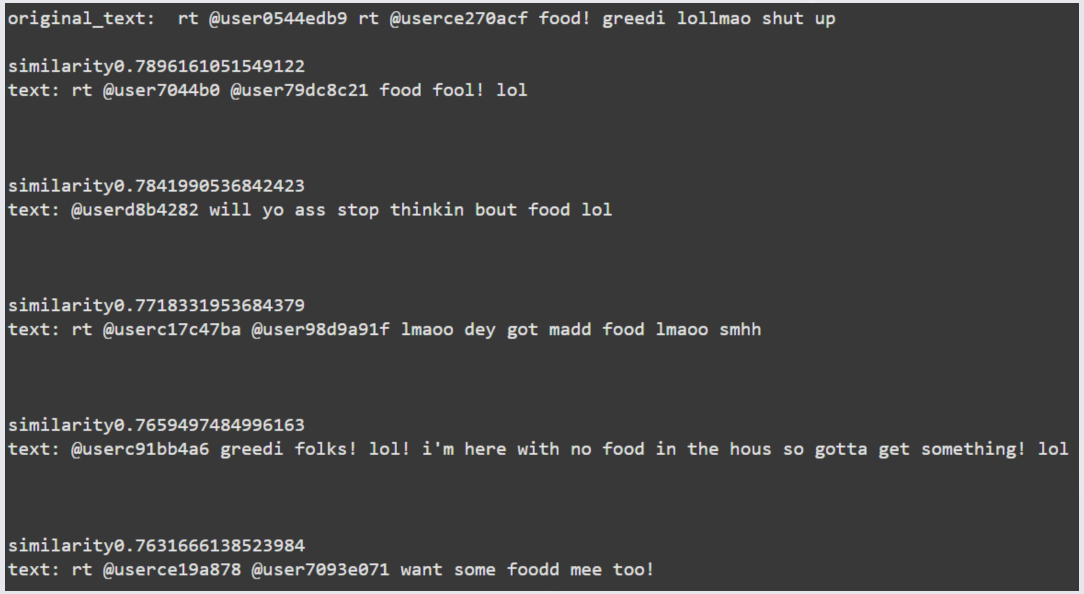

Stephen Checkoway
@stevecheckoway
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe...
#procrastiworking
10:41 PM - 11 Jul 2017
BERT Model
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$



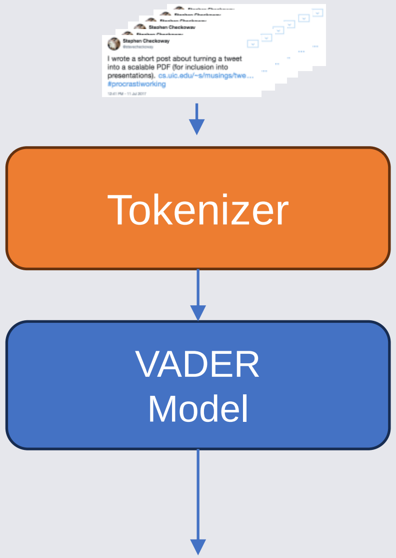


```
original_text:  rt @user0544edb9 rt @userce270acf food! greedi lollmao shut up

similarity0.7896161051549122
text: rt @user7044b0 @user79dc8c21 food fool! lol



similarity0.7841990536842423
text: @userd8b4282 will yo ass stop thinkin bout food lol



similarity0.7718331953684379
text: rt @userc17c47ba @user98d9a91f lmaoo dey got madd food lmaoo smhh



similarity0.7659497484996163
text: @userc91bb4a6 greedi folks! lol! i'm here with no food in the hous so gotta get something! lol



similarity0.7631666138523984
text: rt @userce19a878 @user7093e071 want some foodd mee too!
```

Additional Feature – Sentiment Analysis

Fall 2024 – AI Practicum

- Foundation Model: VADER-Sentiment-Analysis



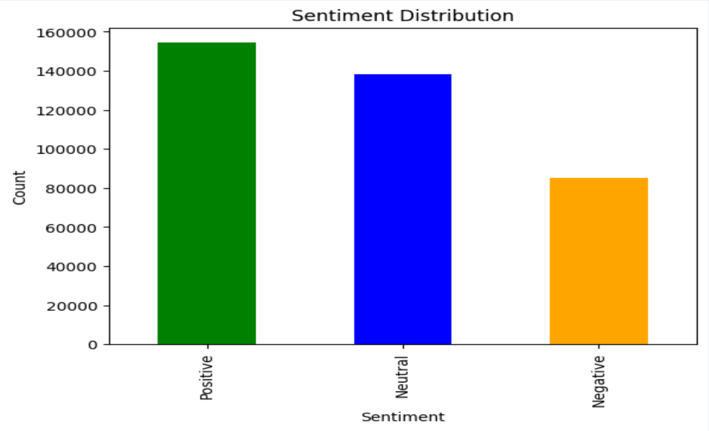
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe… #procrastiworking

Tokenizer

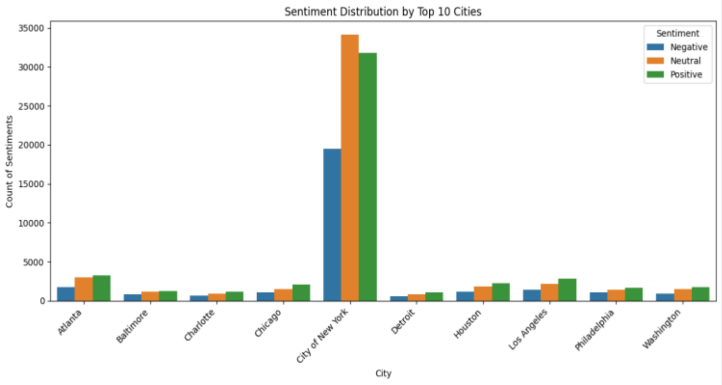
VADER Model

(Batch, 3)

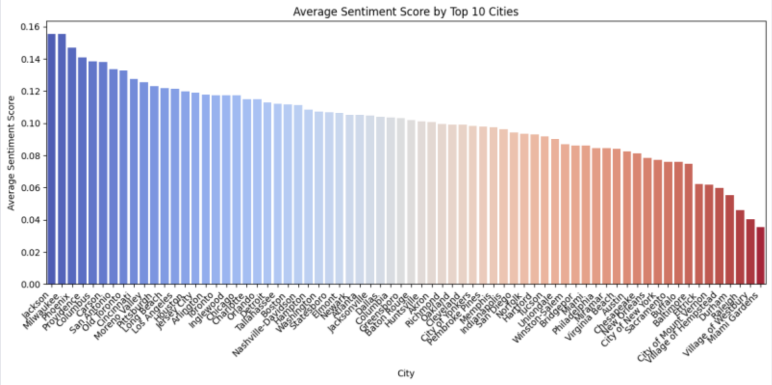
'neg': Negative sentiment score
'neu': Neutral sentiment score
'pos': Positive sentiment score



< Sentiment Distribution >



< Sentiment Distribution: Top 10 Cities >



< Avg Sentiment Score: Top 10 Cities >

- Conclusion : VADER works well for short texts, while BERT is better for longer texts with more context.

Conclusion

Fall 2024 – AI Practicum

- ProProcessing
  - Provide valuable insights
  - Cleaning is Related model performance.

- Model
  - Machin Learning Model is good option for Text Classification
  - BERT Model: Provides vector information applicable to various applications.
    (e.g., Geolocation Prediction, Sentiment and Recommendation)

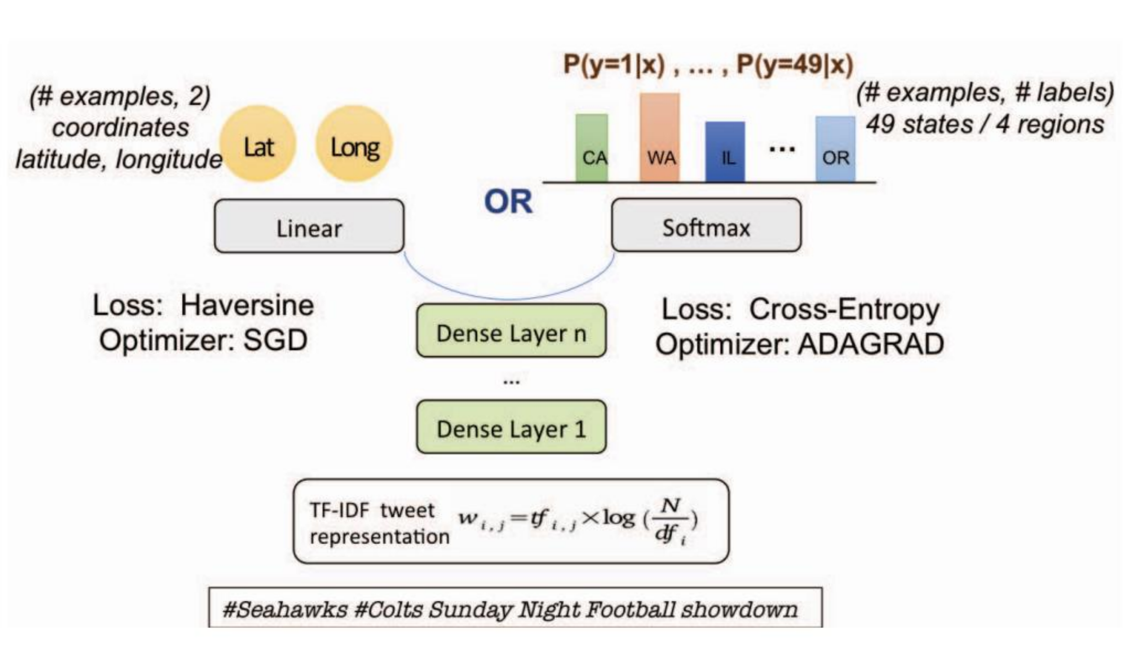
  - Mean and Uncertainty: provide a range of values instead of a single figure for the model.

Appendix

Review Papers - RNN

Fall 2024 – AI Practicum

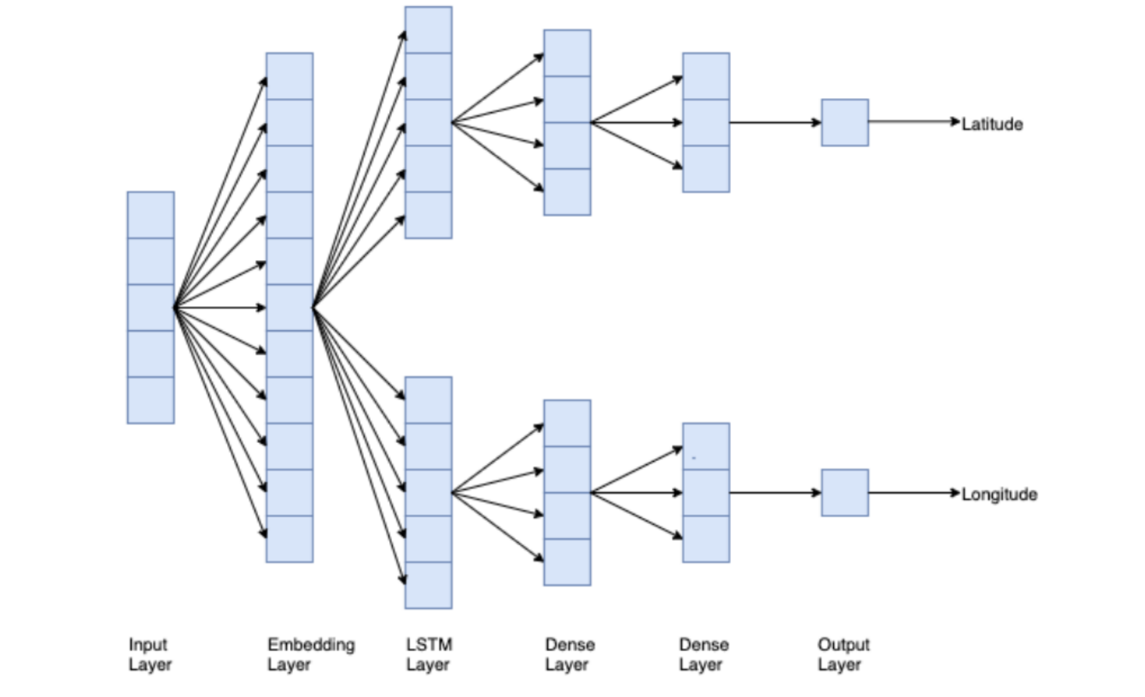
- Text-based Geolocation Prediction of Social Media Users with Neural Network
  - https://isminoula.github.io/files/geoNN.pdf
  - TF-IDF for Text Embeding
  - MLP (Dense Layer)
  - Cross-Entropy Loss



Review Papers – Deep Learning

Fall 2024 – AI Practicum

- Geolocation of Tweets with a BiLSTM Regression Model
  - https://aclanthology.org/2020.vardial-1.27.pdf
  - Bidirectional LSTM
  - FastText Embeding (Subword)
  - Distance based Loss Function



Review Papers - BERT

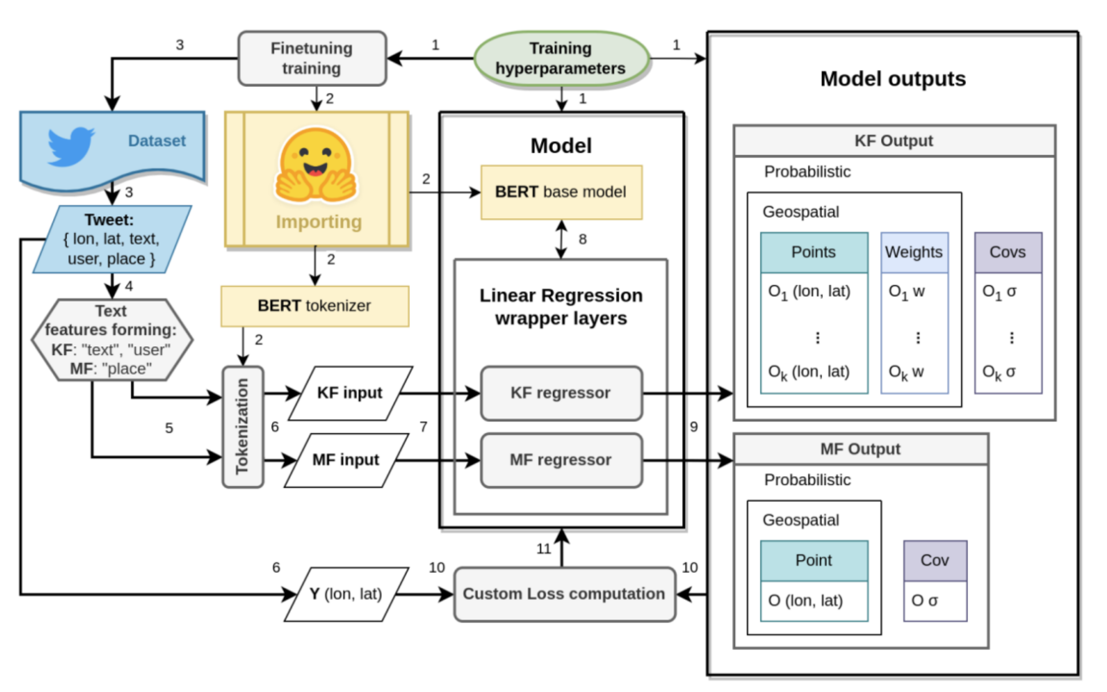
Fall 2024 – AI Practicum

- HeLju@VarDial 2020
  - https://aclanthology.org/2020.vardial-1.19.pdf
  - **ML Approach**
    - SVR with the TF-IDF weighted Character n-gram(n=3~6)
  - **DL Approach**
    - Pre-trained BERT
    - Added FC for geolocation outputs (Longitude, Latitude)

Review Papers - BERT

Fall 2024 – AI Practicum

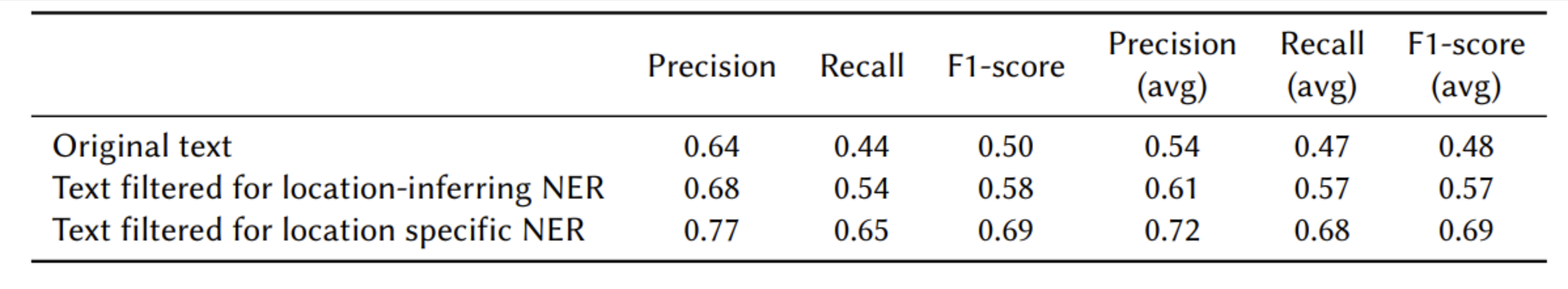
- Predicting the Geolocation of Tweets Using transformer models on Customized Data
  - https://arxiv.org/html/2303.07865v3
  - Use Text(Tweet) and Meta (Timestamp, GeoTag, TimeZone, etc)
  - GMM (Gaussian Mixture Model) based Output.



Review Papers - BERT

Fall 2024 – AI Practicum

- Geolocation Extraction From Reddit Text Data
  - https://ceur-ws.org/Vol-3683/paper2.pdf
  - PreProcessing: NER (Named Entity Recognition)
    - Geolocation Extraction from location specific Reddit

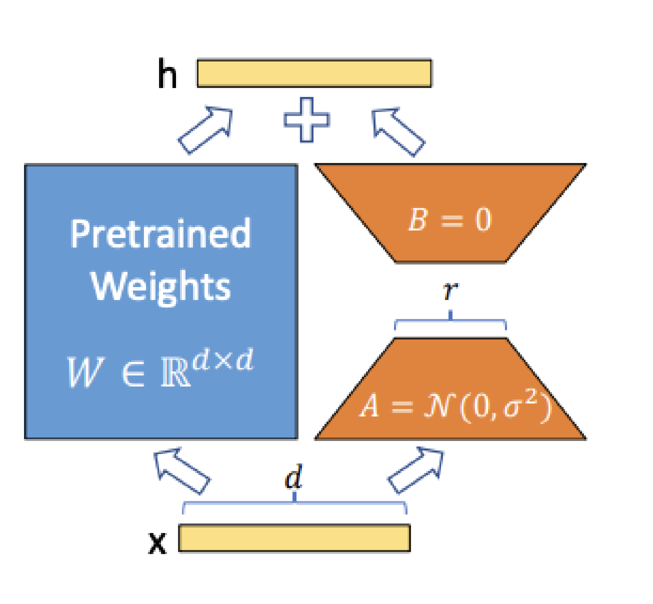


	Precision	Recall	F1-score	Precision (avg)	Recall (avg)	F1-score (avg)
Original text	0.64	0.44	0.50	0.54	0.47	0.48
Text filtered for location-inferring NER	0.68	0.54	0.58	0.61	0.57	0.57
Text filtered for location specific NER	0.77	0.65	0.69	0.72	0.68	0.69

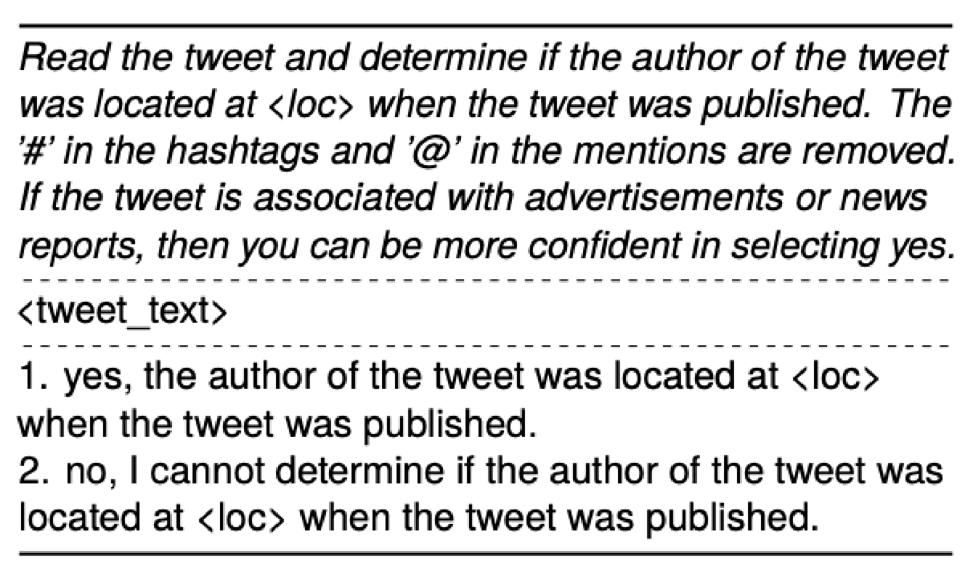
Review Papers - LLM

Fall 2024 – AI Practicum

- Analyzing Large Language Models’ Capability in Location Prediction
  - https://aclanthology.org/2024.lrec-main.85.pdf
  - LoRa(Low-Rank Adaptation) for Fine-Tuning



h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x



Read the tweet and determine if the author of the tweet was located at <loc> when the tweet was published. The ’#’ in the hashtags and ’@’ in the mentions are removed. If the tweet is associated with advertisements or news reports, then you can be more confident in selecting yes.

<tweet_text>

1. yes, the author of the tweet was located at <loc> when the tweet was published.
2. no, I cannot determine if the author of the tweet was located at <loc> when the tweet was published.

Progress Report – Nov 10th

Fall 2024 – AI Practicum

Action Items	Progress
Review Previous Research	• Review the papers of location prediction Model
Processing & Analysis DB	• Cleaning and Organizing DB • Translation • Statistic Analysis
Initial Model Design & Training	• BERT based Model + Header for Location → Fine-Tuning • Predict State, Location (Latitude, Longitude) • LLM: Inappropriate for Limited Resources • Cluster: Vector Clustering based Location Prediction
Additional Feature	• Sentiment Analysis: VADER model, 3 Emotions • Recommendation:

Review Previous Research

Fall 2024 – AI Practicum

- Machine Learning
  - TF-IDE + SVR: Accuracy is worse than the Deep Learning based model

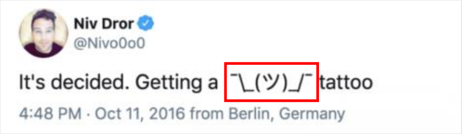
- Deep Learning Model
  - RNN
    - LSTM based Model
    - CNN+SVM
  - **BERT**
    - Transformer Encoder: Foundation model → Model convert text to Vectorized data
    - BERT + Additional Layer: Fine-Tuning only Additional Layer.
  - **LLM (GPT, LLMA)**
    - Adopted Transformer Decoder
    - Fine-Tune with LoRa: require many GPUs.

Process & Analysis DB

Fall 2024 – AI Practicum

Data Preprocessing

- Cleaning DB
  - emojis
  - Missing text fields
  - Inaccurate location data
  - Non-US locations
- Translation
  - Translate Non-English Database Using Google Translator



Niv Dror
@Nivo0o0

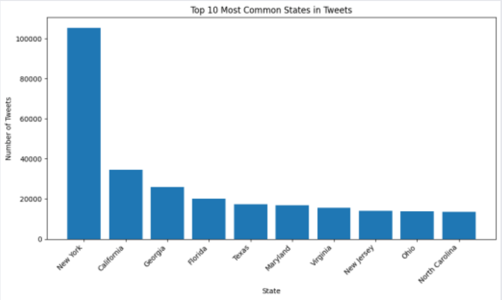
It's decided. Getting a ¯_(ツ)_/¯ tattoo

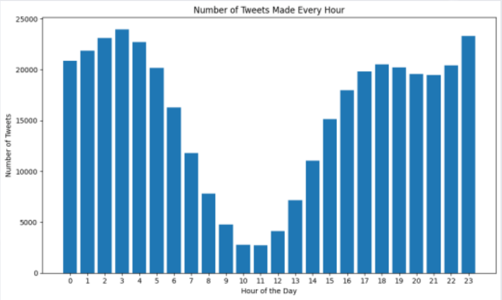
4:48 PM · Oct 11, 2016 from Berlin, Germany

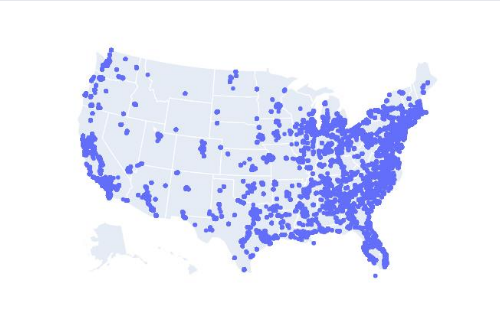
It’s decided. Getting a tattoo.

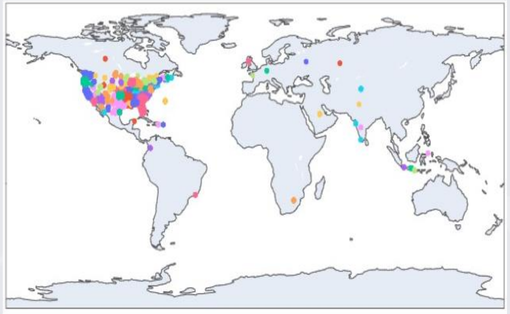
Analysis DB

- Statistical Analyses
  - Geographical Tweet Distribution
  - Tweet Time Distribution Analysis
  - Visualize Geographical Distribution









Review Previous Research

Fall 2024 – AI Practicum

- Machine Learning
  - TF-IDE + SVR: Accuracy is worse than the Deep Learning based model

- Deep Learning Model
  - RNN
    - LSTM based Model
    - CNN+SVM

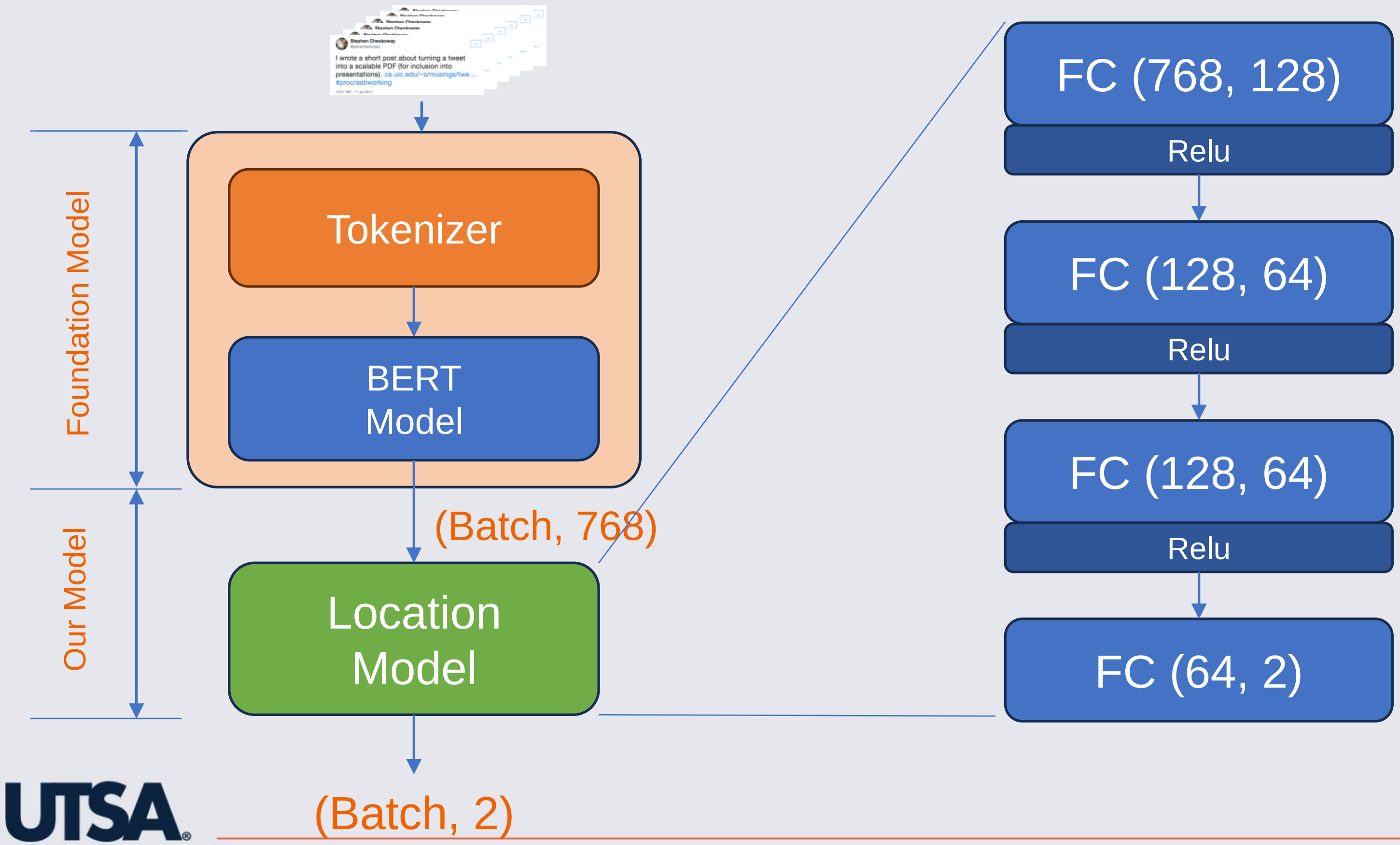
  - **BERT**
    - Transformer Encoder: Foundation model → Model convert text to Vectorized data
    - BERT + Additional Layer: Fine-Tuning only Additional Layer.

  - **LLM (GPT, LLMA)**
    - Adopted Transformer Decoder
    - Fine-Tune with LoRa: require many GPUs.

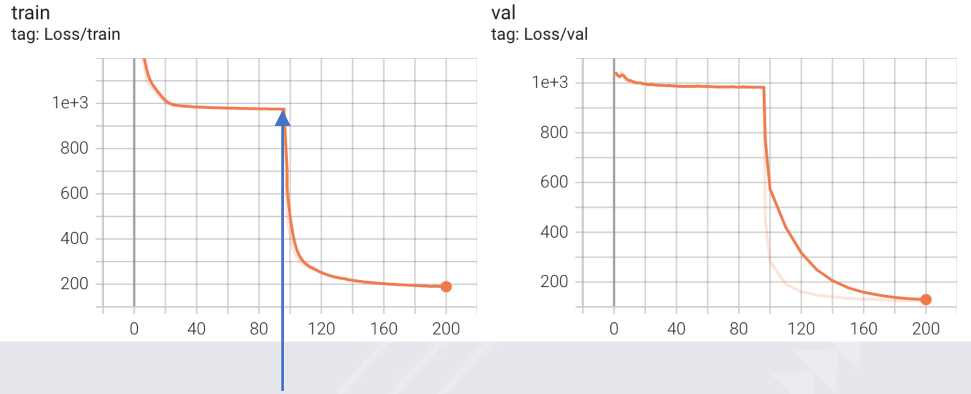
Initial Model Design & Training – BERT for Location

Fall 2024 – AI Practicum

- Foundation Model: Simplified BERT Model (HuggingFace)
- Location Model: Convert Vector to Location



- Training Result
  - Fine-Tune 96 epochs
  - Accuracy: Avg 974km
- Future Work
  - **More complex Model**
  - **Additional Information**

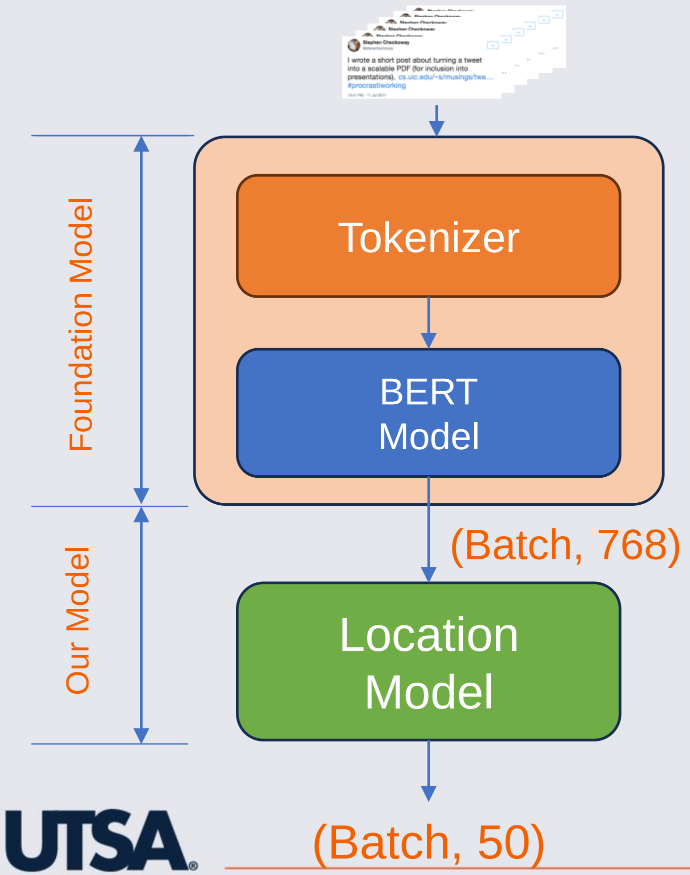


Adding state information to the input significantly improves accuracy. Accuracy improvement is limited with tweet text alone; additional information is needed for Location Prediction.

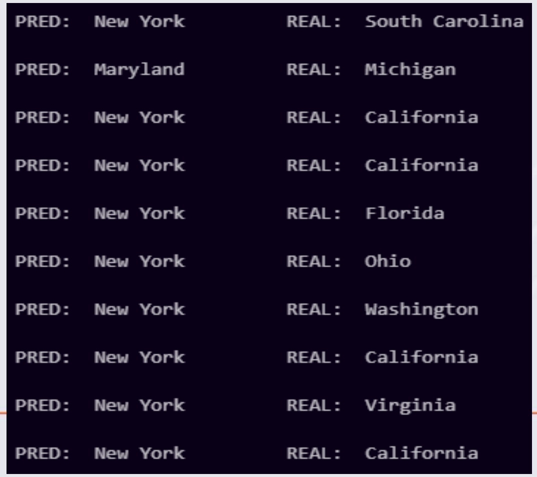
Initial Model Design & Training – BERT State

Fall 2024 – AI Practicum

- Foundation Model: BERT Model (HuggingFace)
- Location Model: Convert Vector to State


Stephen Checkoway
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe... #procrastiworking
Foundation Model
Tokenizer
BERT Model
(Batch, 768)
Our Model
Location Model
(Batch, 50)

- Training Result
  - Accuracy: 30%
  - Very time cost train: 1epoch per hour
  - Suspected to be overfitting since 1/3 of all data is from New York
- Future work
  - **Use ARC for more GPU power**
  - **Truncate the data**



```
PRED:  New York          REAL:  South Carolina
PRED:  Maryland          REAL:  Michigan
PRED:  New York          REAL:  California
PRED:  New York          REAL:  California
PRED:  New York          REAL:  Florida
PRED:  New York          REAL:  Ohio
PRED:  New York          REAL:  Washington
PRED:  New York          REAL:  California
PRED:  New York          REAL:  Virginia
PRED:  New York          REAL:  California
```

Initial Model Design & Training – LLM Approach

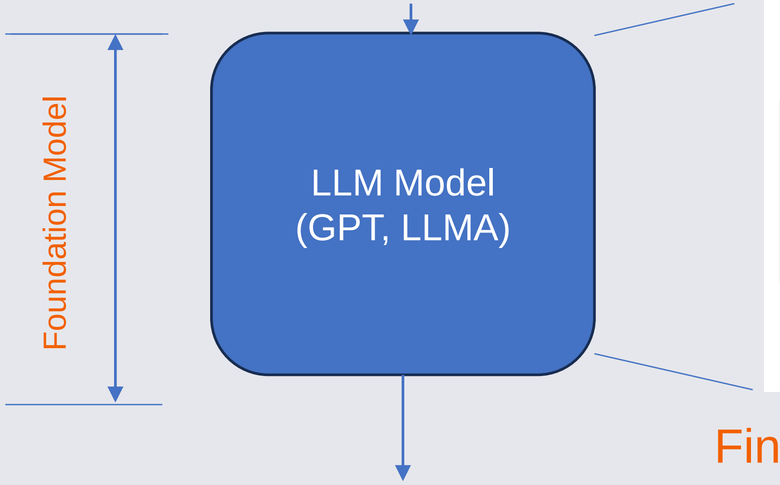
Fall 2024 – AI Practicum

- Foundation Model: LLM Model with LoRA

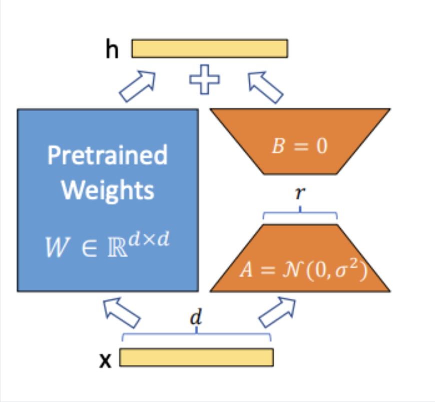
Stephen Checkoway
@stevecheckoway

I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe… #procrastiworking

12:41 PM - 11 Jul 2017



(Batch, Location Text)


h
Pretrained
Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

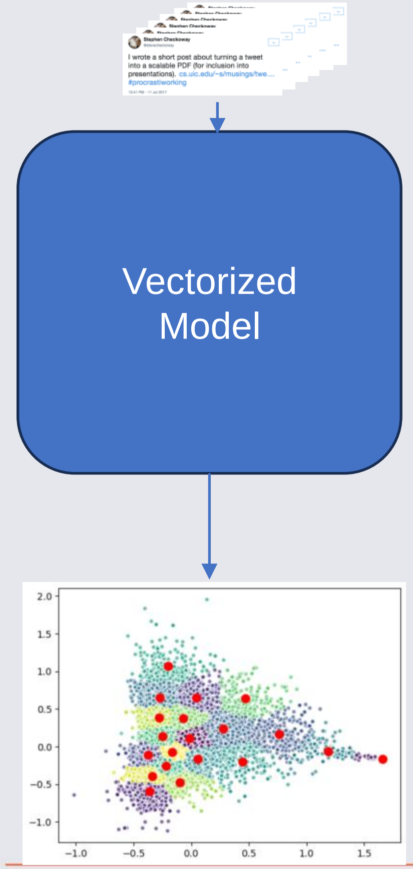
Fine-Tuning with LoRa

- Progress
  - Failed to get permission for LLMA weight Access
  - not feasible due to the high number of GPUs required

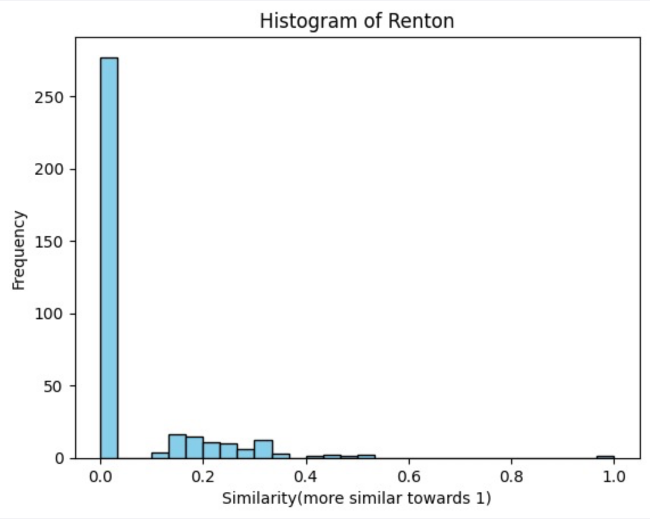
Initial Model Design & Training – Cluster Approach

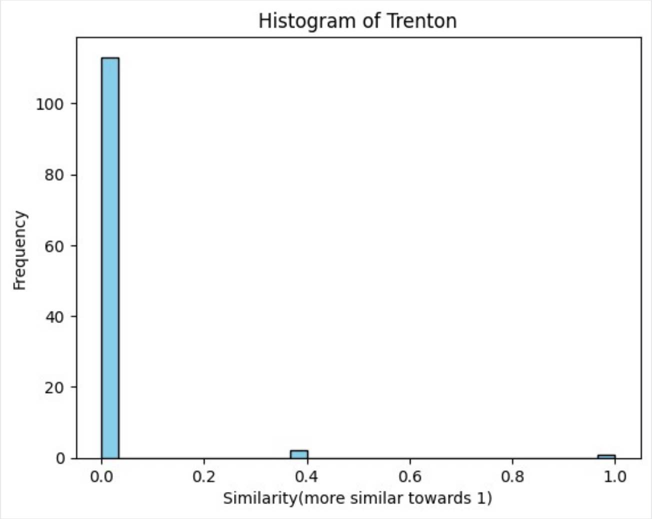
Fall 2024 – AI Practicum

- Vectorized tweet(text) Clustering


Stephen Checkoway
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe… #procrastiworking
Vectorized Model
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5

- Progress
  - create vector of text tokens which are 500 dimensions vector
  - we can clearly see not that much similarity there for same city so we can not make cluster

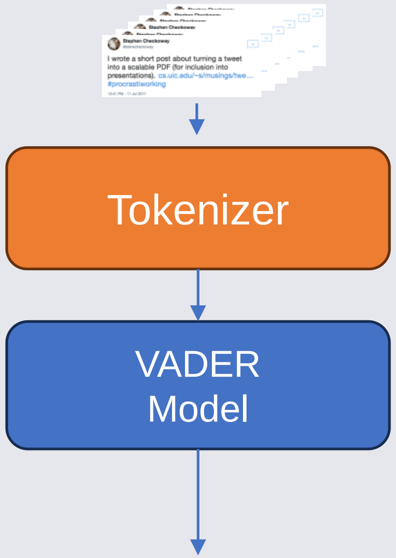




Additional Feature – Sentiment Analysis

Fall 2024 – AI Practicum

- Foundation Model: VADER-Sentiment-Analysis

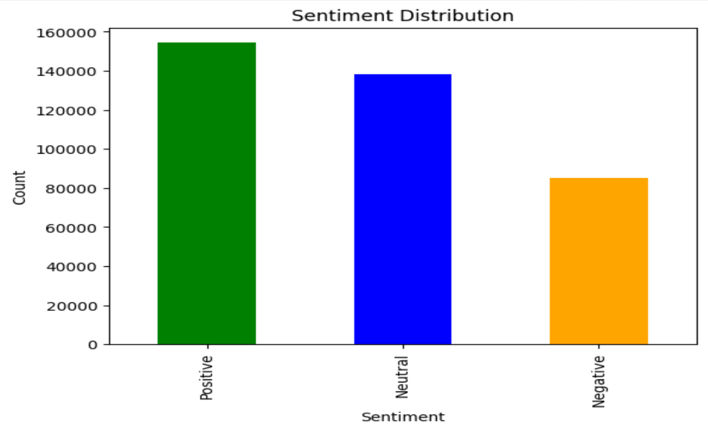

Stephen Checkoway
@stevecheckoway
I wrote a short post about turning a tweet into a scalable PDF (for inclusion into presentations). cs.uic.edu/~s/musings/twe…
#procrastiworking
12:41 PM - 11 Jul 2017

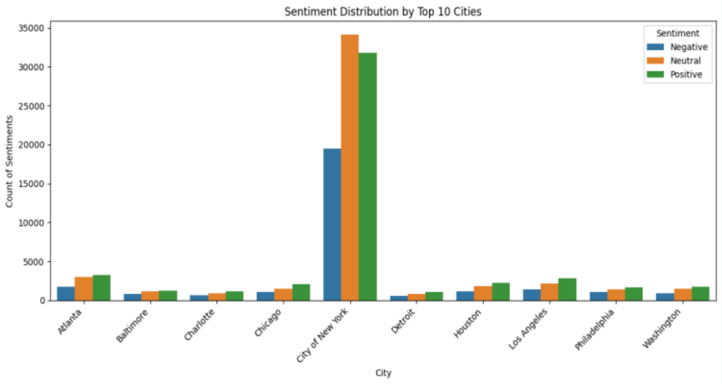
Tokenizer

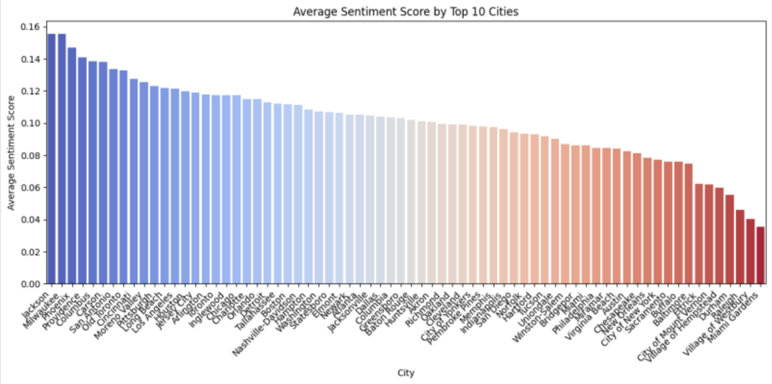
VADER Model

(Batch, 3)

'neg': Negative sentiment score
'neu': Neutral sentiment score
'pos': Positive sentiment score







- Conclusion : In small text, it is difficult to determine joy, surprise, shock, angry or any other specific emotion. So Vader worked well in this context. Bert would be better if there are larger texts with more contexts.