SAAS

Presentation Contents

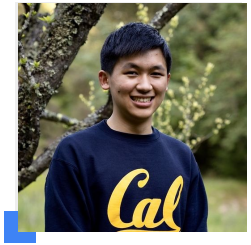
IA Industrial Project

Project Background

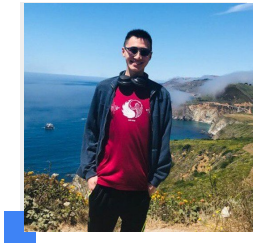
Data Acquiring

Methods
-Word Embeddings
-Transformers
-Transfer Learning

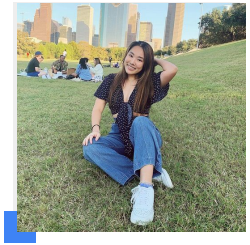
The Team



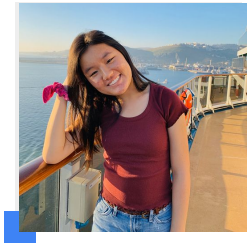
Alex



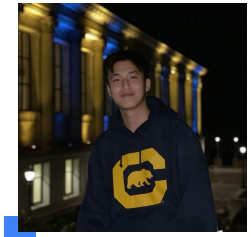
Alvin



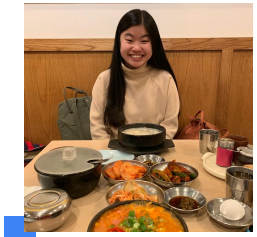
Amy



Ashley

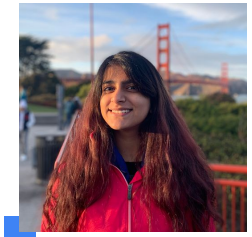


David



Erica

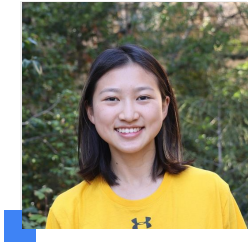
Grace



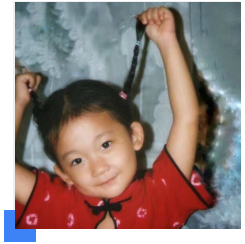
Gunjan

The Team

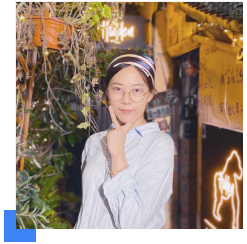
Jay



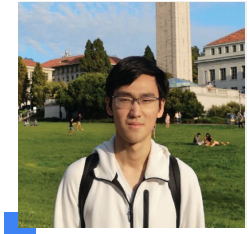
Kaci



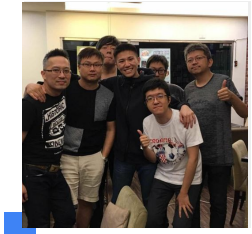
Smin



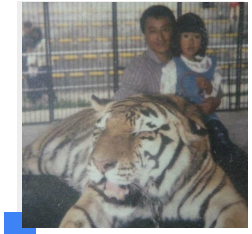
Yiming



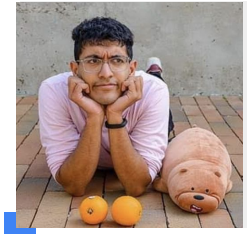
Haoming
(Director)



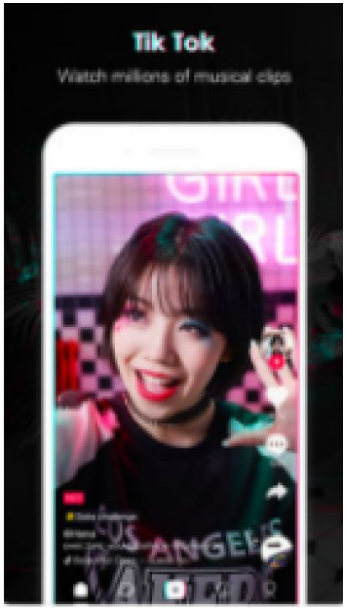
Michael
(Director)



Amanda
(Director)



Ronnie
(Advisor)



Project Background

Problem - Analyze language styles of live streamers (e.g. youtubers)

Sub-problems - **Emotions, Style**, Interactions, Content, etc.

Applications - Recommendation; Finding good live streamers

11/07

12/05

Project Start

Research; Acquire Data

Implement; Present

DC Project?

Present to Client?

Data Collection - Exploration

Sources we explored:

Kaggle Datasets

Twitch VODs

YouTube Vlogs

Podcast Transcripts

Data Collection - Youtube Vlog Transcripts

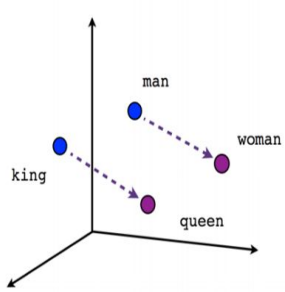
ID	Title	Description	Tags	Channel	Views	Likes	Dislikes	Comments	Transcript
t4lL0SYbO74	Student Diaries | Study sleepover, finishing a...	So I made my boyfriend sleepover and study wit...	['student diaries', 'student diaries vlog', 's...	Briddy Li	1966741	150043	1237	6888	i think i'm gonna start doing work nowum i jus...
LhDW7n4TAAM	CATHERINE TRIED TO CUT OUR SON'S HAIR...AND RU...	CATHERINE TRIED TO CUT OUR SON'S HAIR...AND RU...	["Catherine tried to cut our son's hair", "the...	The ACE Family	1966741	150038	1237	6888	that's good[Music]whatever you need[Music]what...
nTtMY5vsYMY	sam & colby got us lost on a 4 mile ghost hunt...	✧ wild 'til 9 podcast: https://bit.ly/WildTil9...	['laurdiy', 'laurdiy vlog', 'laurdiy vlog new'...	LaurDIY Vlogs	1966741	150035	1237	6887	when i tell you that i'm doing some [__]righ...
st-xjMX46iw	Can't Believe De'arra Ordered This $600 Litter...	*FOLLOW US AROUND THE WORLD* (travel vlogs) ⇣ ...	['litter robots', 'litter', 'litter robot', 'r...	Vlogs By DK4L	1966741	150043	1237	6887	[Music]so today is october 1standi know what y...
LHnJ-hbbMqk	not steady - paloma mami (by manu rios)	😈🔪😈🔪😈 \n\nshop my merch here: https://shopmanu...	['not steady', 'paloma mami', 'manu rios', 'co...	Manu Rios	1966741	150049	1237	6887	

Data Collection - Next Step Ideas

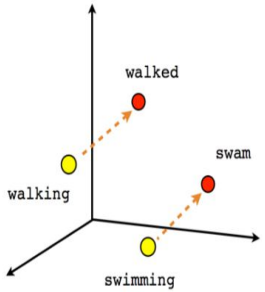
- Transcripts are in chunks of 5 words (like how you see captions in a youtube video), distinguishing the sentences might be useful
- Try to clean the dataset in a way that will allow models to analyze/predict emotion, not content (“angry” and “excited”; not “school”)
    - Remove influential nouns or nouns that appear often if one channel’s transcripts?
    - Giving higher weights/importance to adjectives? (parts-of-speech tagging)
    - Possibly manually labeling the topics/guessed emotions of each video
    - Really depends on the models we want to use

Word Embeddings

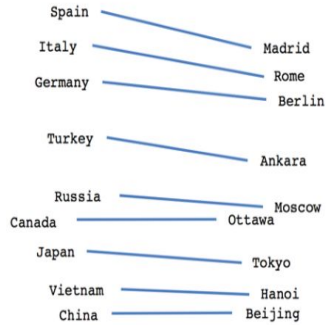
- Word embedding is any of a set of language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped to vectors of real numbers.
- It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.
- Word2Vec is one of the most popular technique to learn word embeddings using shallow neural network.



Male-Female

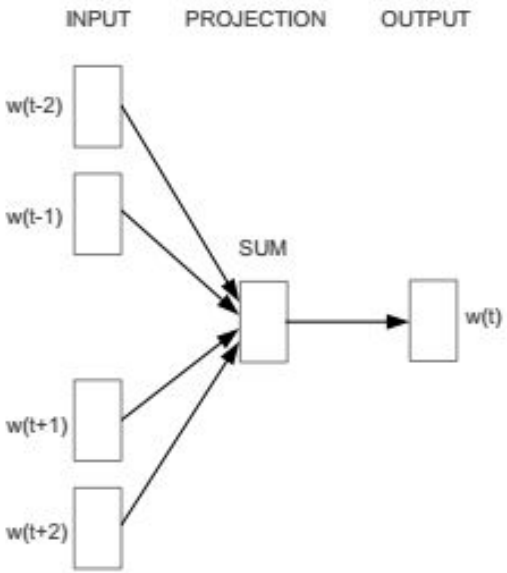


Verb tense

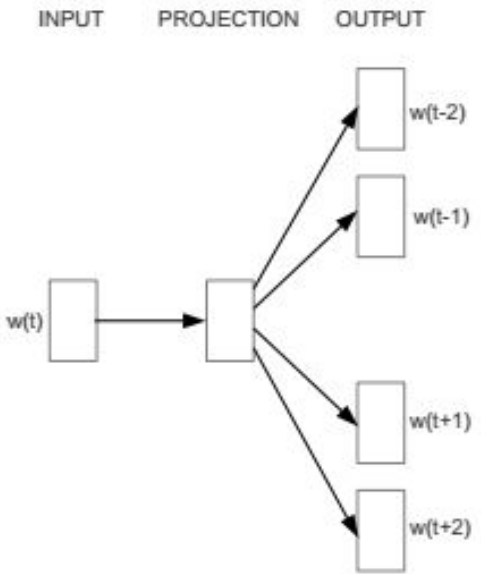


Country-Capital

Models



CBOW



Skip-gram

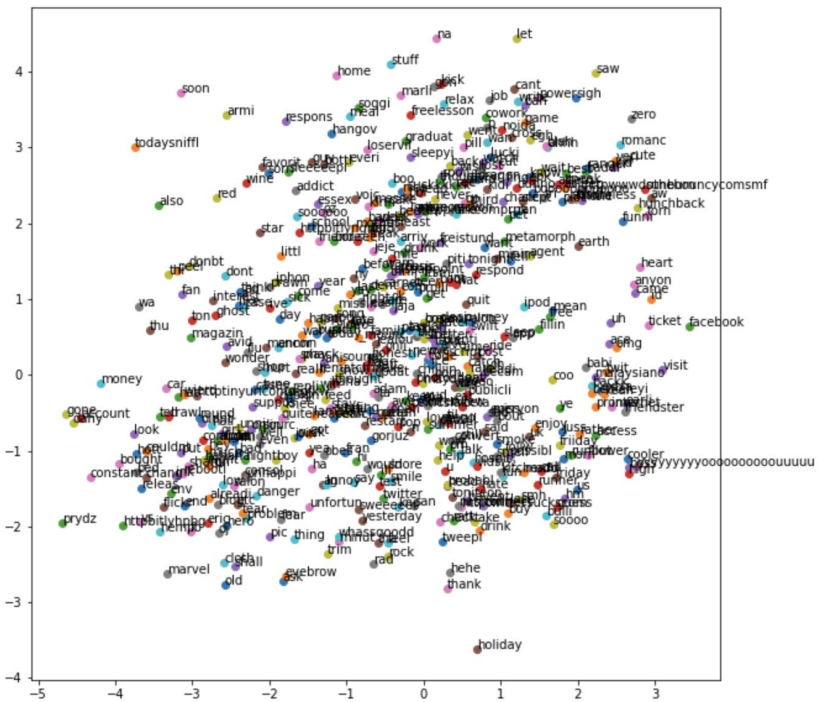
Implementation

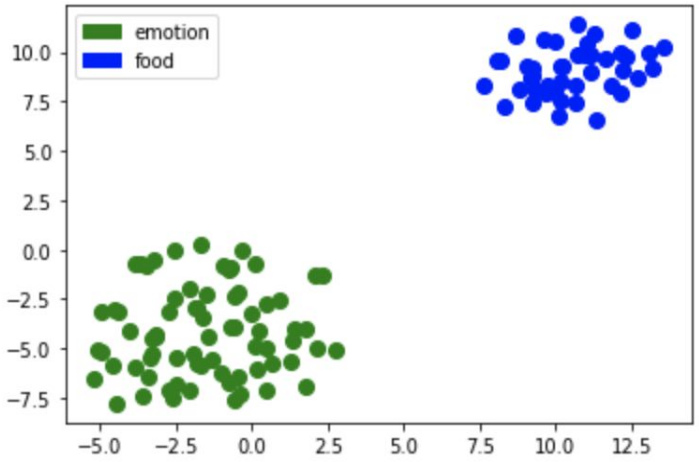
1. Use Deepnote GPU for parallel data processing to share information
2. Download the Twitter emotion dataset from Kaggle and install the word2vec library

1. Standardized the dataset by removing numbers, punctuation, and stopwords
2. Took the word stems of remaining words and tokenized phrases

1. Created a bag of words and used one-hot encoding
2. Set up a dense neural network with 3 layers
3. Trained model with Keras, assigning context scores to words

Results



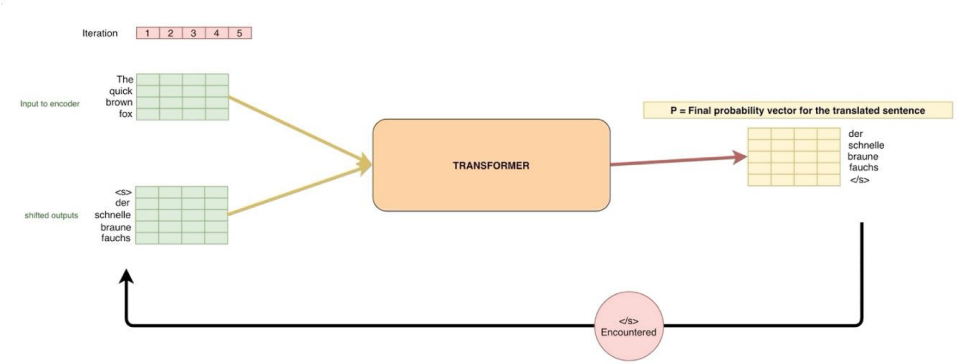



```
model.most_similar(positive=['woman', 'king'], negative=['man'], topn=5)
[(u'queen', 0.711819589138031),
 (u'monarch', 0.618967592716217),
 (u'princess', 0.5902432799339294),
 (u'crown_prince', 0.5499461889266968),
 (u'prince', 0.5377323031425476)]
```

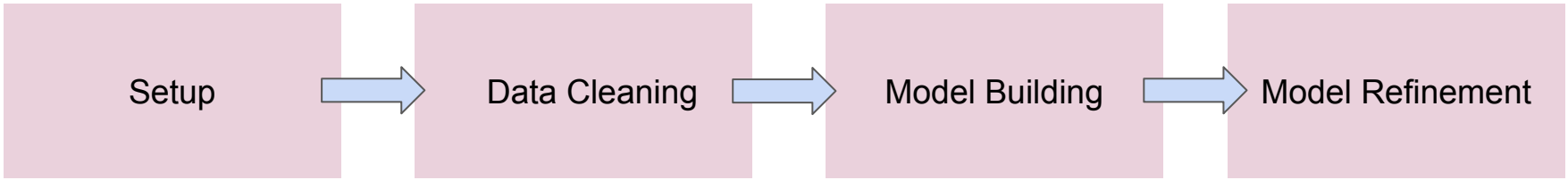
```
[('queen', 0.711819589138031),
 ('monarch', 0.618967592716217),
 ('princess', 0.5902432799339294),
 ('crown_prince', 0.5499461889266968),
 ('prince', 0.5377323031425476)]
```

BERT

- **Transformers** - sequence transduction/ neural machine translation
    - Input sequence → output sequence (i.e. speech recognition, text-to-speech)
- **Bidirectional Encoder Representations from Transformers (BERT)**, is a transformer-based pre-trained NLP model developed by Google in 2018
- **Bidirectional** - input sequence and reverse of input sequence
- Processes each word **in relation to all other words** in the sentence then fine tunes smaller NLP tasks



Implementation



1. Use Colab GPU for parallel data processing to save time
2. Install Hugging Face Library for a Pytorch interface

1. Removed unneeded columns and dropped NA values
2. Split data into train, validation and test sets

1. Tokenization-special tokens
2. Padding & Truncating to make all sentences constant length
3. Conversion to Pytorch datatype

1. Compare training accuracy and test accuracy
2. Go back to the Model Building stage to refine the model

Youtube Genre Predictor

SAAS

Building Youtube Video Caption Dataset	Sentiment Analysis	Predicting the Genre	Expansions
Youtube Transcript library: Gets the caption transcripts of the Youtube Data API video IDs Correlating its video IDs with data from using the Youtube Data API which gives Channel, Video Description, Title and Tags. This along with the transcripts will make a video dataset which will be used to train and test our model	Researched models and libraries such as Word2Vec, FastText, and Skip Gram.	Predict the genre using the sentiment analysis and compare that to the tags of the videos in the test dataset to find the accuracy of our model	Test if sentiment analysis of reviews helps make the genre predictions more accurate

Transferring the Knowledge (Embeddings)

- Predict topic/genre of transcript
  - Comedy
  - Beauty
  - Gaming
  - Lifestyle
  - Commentary
- Sentiment and Emotion Scores/Classification with Discrete Emotion Models, Dimensional Emotion Models or NLP Style Analysis
- Go beyond Youtube with Twitch VODs, Ted Talks, NPR recordings and transcript
- To better analyze the speaker's style, can remove genre related keywords from transcription

SAAS

Recap

- Dataset
- Word2Vec
- Transformers
- Knowledge Transfer

Questions?