CSE 408: Multimedia Information System (2021 Spring)

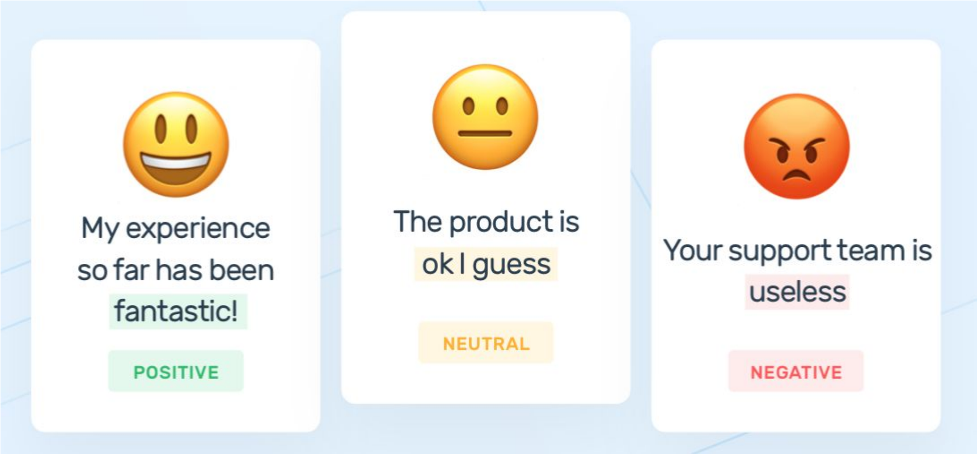
CSE 408 Twitter Sentiment Analysis

Arizona State University

Jielin Wu, Ziming Dong, Jianlun Li

Problem Statement

Sentiment analysis refers to using data, such as text or images, to analyze what you think or feel about something, about almost anything. Twitter sentiment analysis is considered a binary classification problem. According to the two categories of positive and negative tweets, each category of tweets was divided into three or more categories.



My experience so far has been fantastic!

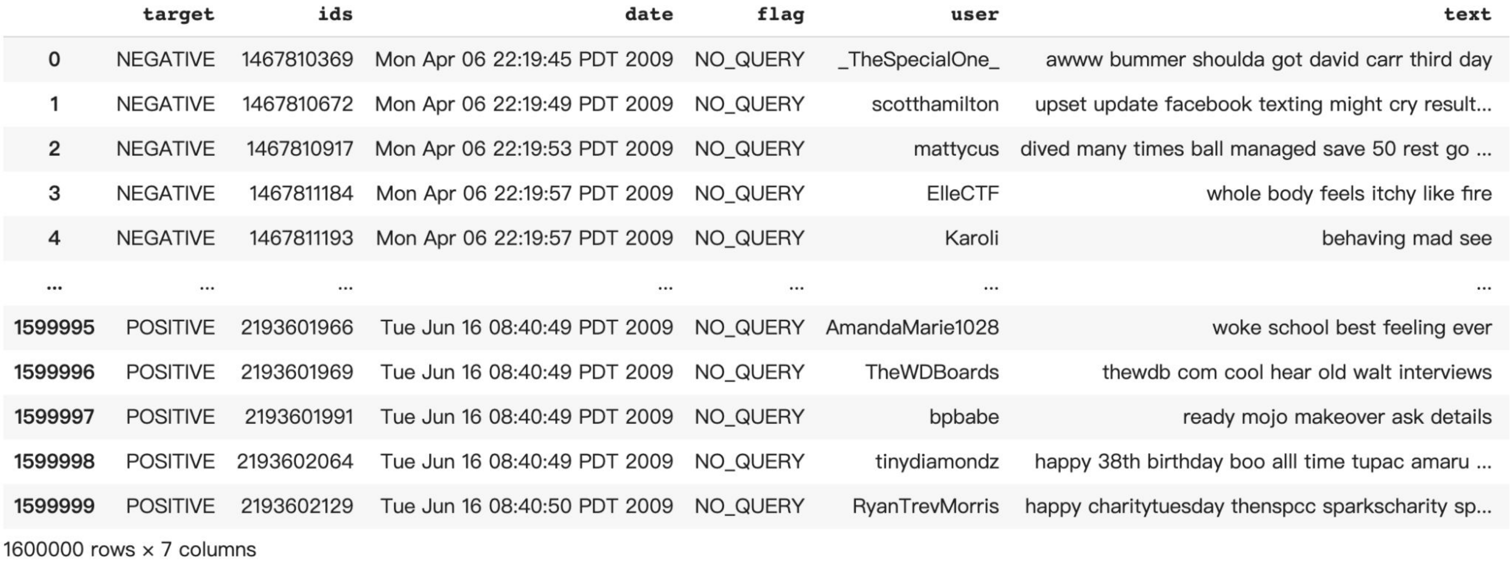
The product is ok I guess

Your support team is useless

Sentiment Analysis also help decision-making process in company, for example, based on the customer reviews of products, company can make more or less production to avoid any economic losses.

Dataset Description

We use the famous benchmark twitter analysis datasets: sentiment140 which contains 1.6 millions tweets, the tweets are labeled either positive or negative sentiment. The datasets are balanced which means it contains 800000 positive tweets and negative tweets and no missing values, the team also replace the target “0”, “4” to “NEGATIVE” and “POSITIVE”. The dataset contains six fields shows below:

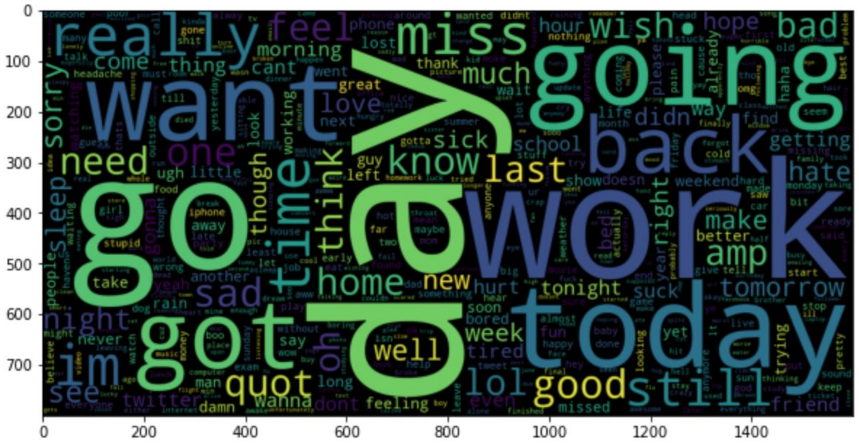


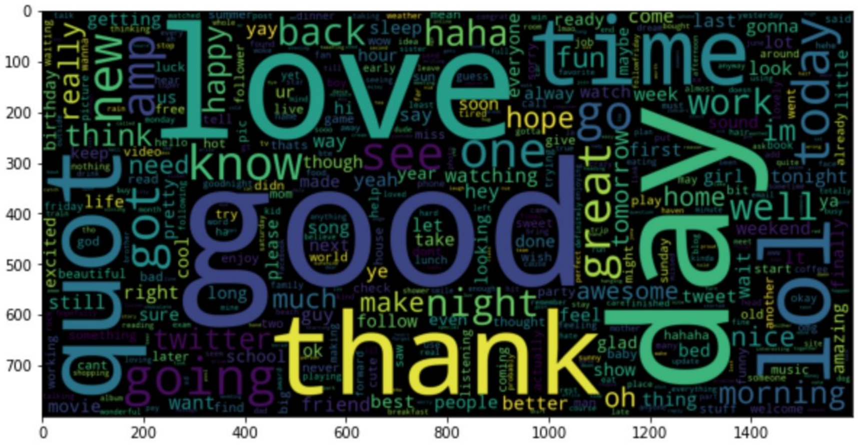
	target	ids	date	flag	user	text
0	NEGATIVE	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	awww bummer shoulda got david carr third day
1	NEGATIVE	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	upset update facebook texting might cry result...
2	NEGATIVE	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	dived many times ball managed save 50 rest go ...
3	NEGATIVE	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	whole body feels itchy like fire
4	NEGATIVE	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	behaving mad see
...	...	...	...	...	...	...
1599995	POSITIVE	2193601966	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	AmandaMarie1028	woke school best feeling ever
1599996	POSITIVE	2193601969	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	TheWDBoards	thewdb com cool hear old walt interviews
1599997	POSITIVE	2193601991	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	bpbabe	ready mojo makeover ask details
1599998	POSITIVE	2193602064	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	tinydiamondz	happy 38th birthday boo alll time tupac amaru ...
1599999	POSITIVE	2193602129	Tue Jun 16 08:40:50 PDT 2009	NO_QUERY	RyanTrevMorris	happy charitytuesday thenspcc sparkscharity sp...

1600000 rows × 7 columns

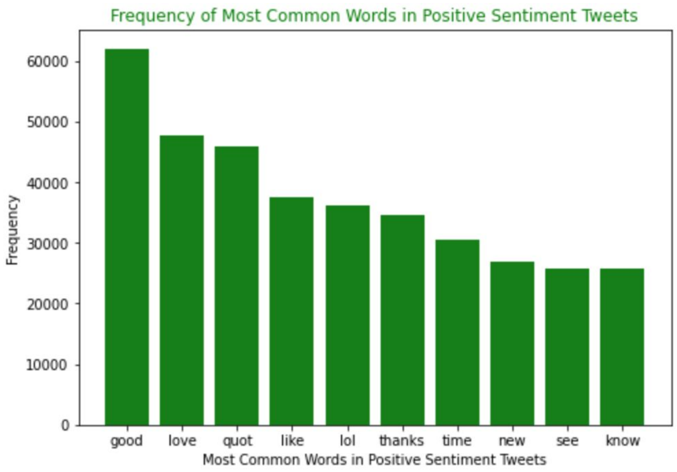
Exploratory Data Analysis(EDA)

After we remove the stop words, special symbols, and transfer all strings to lowercase, we did the Exploratory Data Analysis(EDA) work for the whole datasets, we plot the Word Cloud for positive and negative tweets, we also plot the frequency of most common word in positive and negative tweets, finally, we plot the distribution of words for the tweets texts.

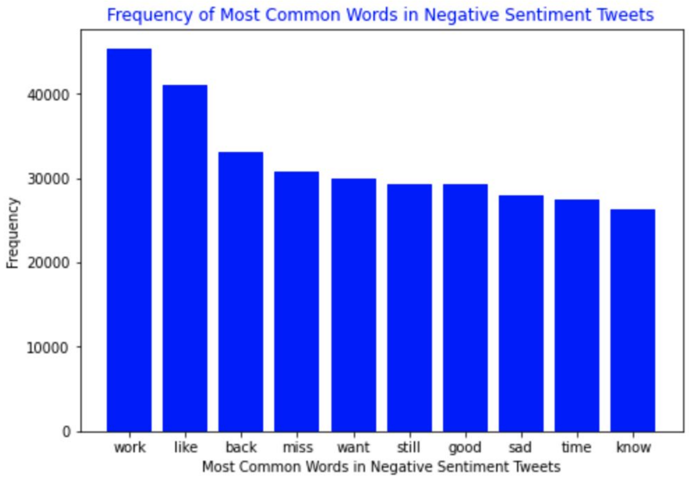




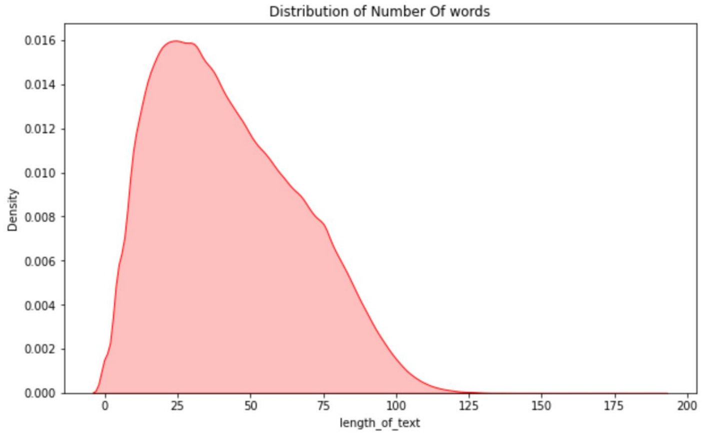
Data Analysis Results



Frequency of Most Common Words in Positive Sentiment Tweets



Frequency of Most Common Words in Negative Sentiment Tweets



Distribution of Number Of words

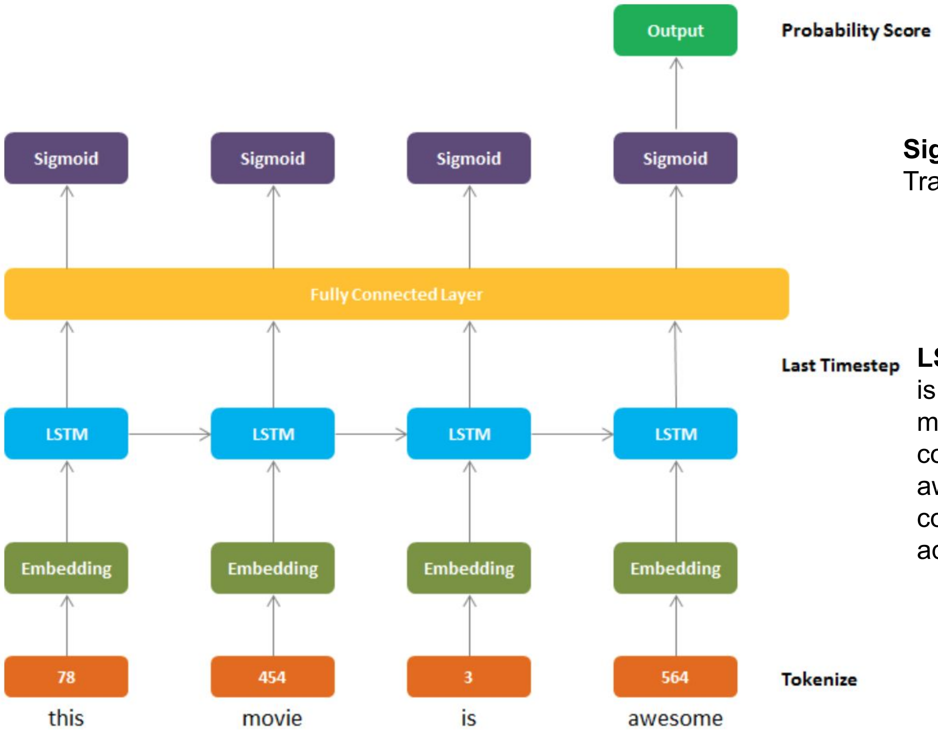
Data Augmentation Technology

Due to the machine limited, we will use 25000 tweets as training dataset and 3000 tweets as testing datasets instead of splitting 1.6 millions. For the training dataset, we use **nlpaug** and **textattack** data augmentation technologies, we use the **SynonymAug** to generate 5000 similar texts and merge them into training dataset, the goal is to generate more sentiment words to avoid overfitting problems.

Origin Text	nlpaug	textattack.augmentation
This week is not going as I had hoped.	Your week is not going as you expected.	This week is not extend going as I had hoped.

Describe Existing Approaches

We develop Deep Learning model by fine tuning hyperparameters from existing state-of-art model architectures.



Sigmoid Activation Function: Transform value between 0.0 to 1.0.

Dense Layer: Fully Connected Layers for classification

LSTM: Long and short time memory is a variant of RNN that has a memory state cell that learns the context of words that are farther away from the text and carry contextual meaning rather than just adjacent words like RNN.

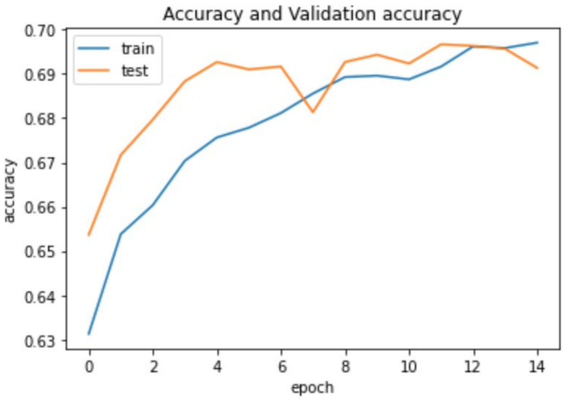
Embedding Layer: Generates Embedding Vector for each input pad sequence.

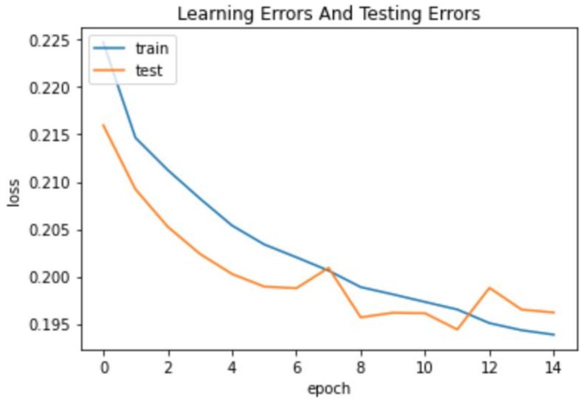
Model Evaluations

	Accuracy	Val_accuracy
Baseline code model trained with 1.6millions tweets	0.7774	0.7890

	Accuracy	Val_accuracy
Fine Tuning model trained with 25k tweets **without** data augmentation work	0.6699	0.6950

	Accuracy	Val_accuracy
Fine Tuning model trained with 25k tweets **with** data augmentation work	0.7000	0.6967


Accuracy and Validation accuracy


Learning Errors And Testing Errors

Updated Project Plan: Task, Deadlines, Division of Work

Task #	Description	Team Member	Deadline
1	Research on Twitter Sentiment Analysis Kaggle competition	Ziming, Jianlun, Jielin	3/20(Done)
2	Project Proposal	Ziming, Jianlun, Jielin	4/6(Done)
3	Exploratory Dataset Analysis	Ziming Dong	4/10(Done)
4	Research on Existing Approaches	Ziming, Jianlun, Jielin	4/13(Done)
6	Model Evaluation	All Members	04/14(Done)
7	Final Report	Ziming, Jianlun, Jielin	4/20

https://github.com/AllenX-Li/CSE408-Tweet-Analysis

References

- Nlpaug: https://nlpaug.readthedocs.io/en/latest/

- TextAttack: https://textattack.readthedocs.io/en/latest/

- Kaggle Twitter Sentiment Analysis: https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis/output

- Twitter Feature Extraction: https://www.kaggle.com/tanulsingh077/twitter-sentiment-extaction-analysis-eda-and-model/data

- Model Training LSTM: https://www.kaggle.com/arunrk7/nlp-beginner-text-classification-using-lstm

Demo Time!
Thank You For Watching!

Q&A