DETECTING SENTIMENT

WITH 1.6 MILLION TWEETS

SUMMER '24
DATASCI 207

WILLIAM LEI
XUEYING TIAN
BERNARDO COBOS
REBECCA BARGIACHI

RESEARCH QUESTION

HOW CAN A MACHINE LEARNING MODEL BE OPTIMIZED TO IMPROVE PREDICTION ACCURACY?

SUB-QUESTIONS

What are the key factors that predict the sentiment of a tweet?

What contributes to sentiment classification accuracy?

OUR GOAL

Develop a binary classification model to identify sentiment (positive or negative) based on the language in their tweets.

Potential uses:

- Monitoring brand perception
- Getting a temperature check on product reception
- Quickly identifying negative sentiments to manage public relations crises
- Gauging public reaction to new policies or government actions
- Measure employee satisfaction to proactively discover issues

WHAT HAS BEEN DONE?

Most used text-based emotion recognition techniques:

1. KEYWORD SPOTTING METHOD
2. LEXICAL AFFINITY METHOD
3. LEARNING –BASED METHOD
4. HYBRID METHODS

DATA SOURCE

KAGGLE'S SENTIMENT140

- 1.6 million tweets extracted by twitter API
- Automatically labeled
- Assumption that emojis correlate to sentiment

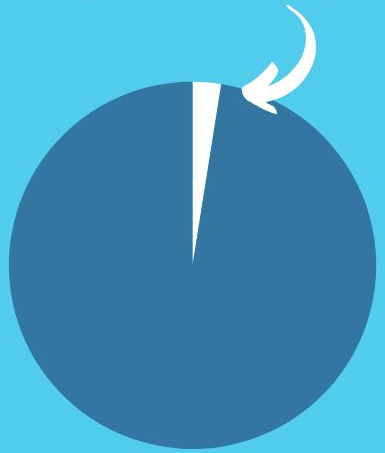
DATA FIELDS:

- **target**: Polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- **ids**: Tweet ID
- **date**: Date of the tweet
- **flag**: Query (if any)
- **user**: User who tweeted
- **text**: Text of the tweet

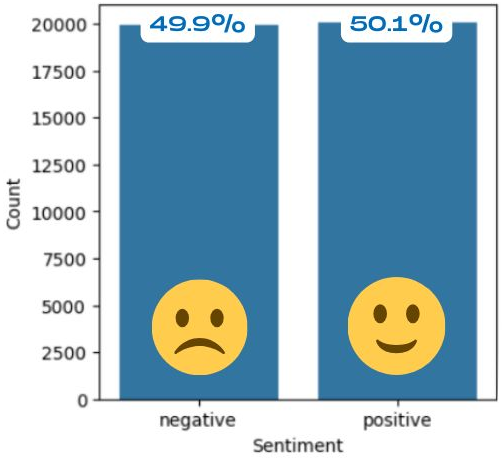
	target	ids	date	flag	user	text	sentiment
514293	0	2190584004	Tue Jun 16 03:08:48 PDT 2009	NO_QUERY	Vicki_Gee	i miss nikki nu nu already shes always there ...	negative
142282	0	1881451988	Fri May 22 04:42:15 PDT 2009	NO_QUERY	PatCashin	So I had a dream last night. I remember a sig...	negative
403727	0	2058252964	Sat Jun 06 14:34:17 PDT 2009	NO_QUERY	deelectable	@girlyghost ohh poor sickly you (((hugs)) ho...	negative
649503	0	2237307600	Fri Jun 19 05:34:22 PDT 2009	NO_QUERY	justinekepa	it is raining again	negative
610789	0	2224301193	Thu Jun 18 09:20:06 PDT 2009	NO_QUERY	cmatt007	@MissKeriBaby wish I was in LA right now	negative

EXPLORATORY DATA ANALYSIS

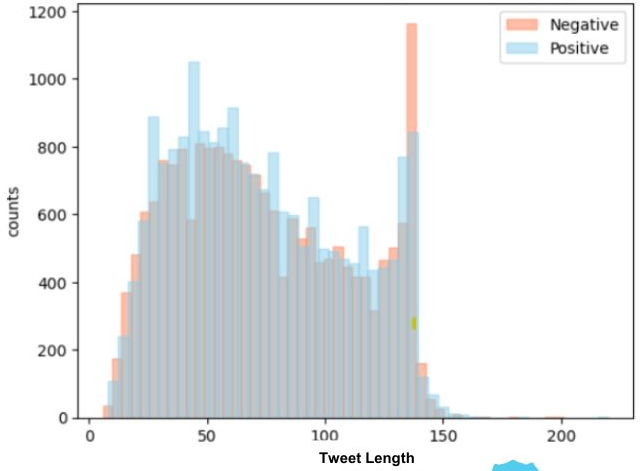
OUR SAMPLE: 40,000 TWEETS, 2.5% OF POPULATION



SENTIMENT CLASS DISTRIBUTION



POSITIVE & NEGATIVE TWEET DISTRIBUTION



DATA PRE-PROCESSING

REMOVED URLS, @ MENTIONS, & ALL SPECIAL CHARACTERS

DATA CLEANING

VOCAB SIZE: 2000
MAX SEQ. LENGTH: 100

TOKENIZATION & PADDING

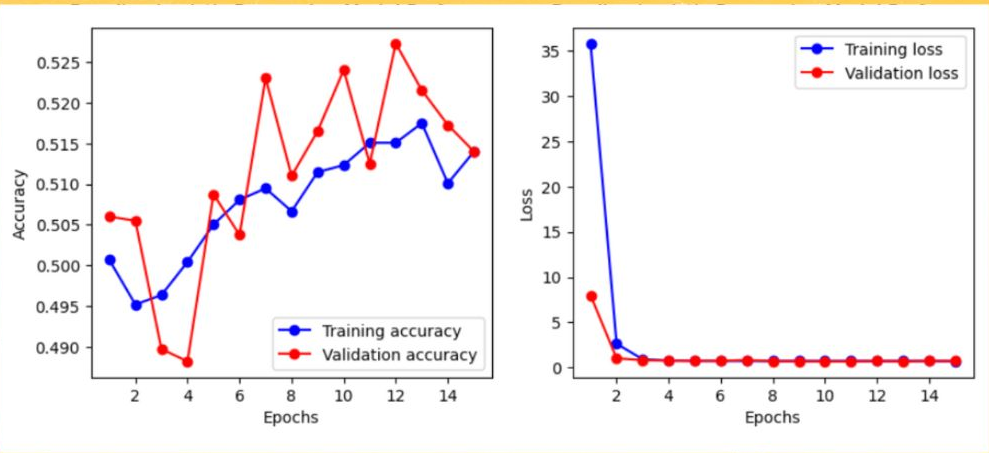
60
20
20

TRAIN, TEST, & VALIDATION SPLIT

BASELINE MODEL

TRAINING ACCURACY

TRAINING LOSS



ARCHITECTURE

- 1 Fully Connected Dense Layer
- Activation: Sigmoid
- Loss: Binary Crossentropy

Hyperparameter tuning results:

- LR = 0.0005
- Num Epochs = 15
- Optimizer = ‘Adam’

PERFORMANCE

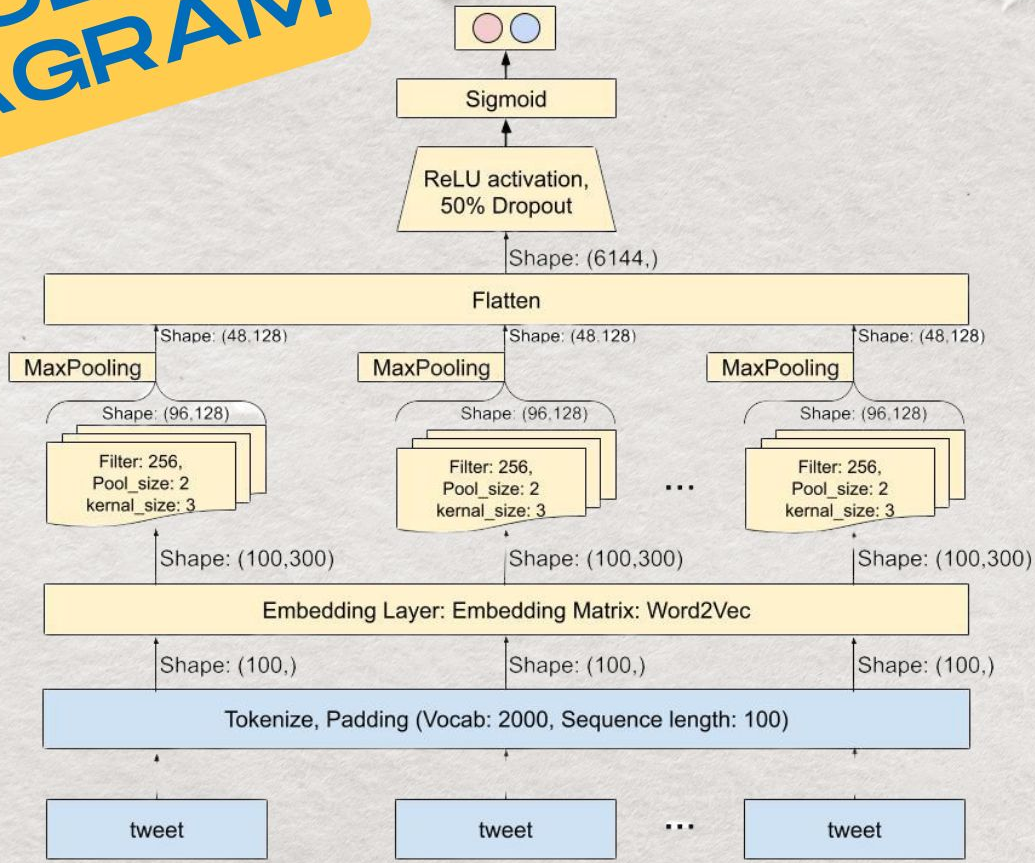
LOSS: 0.7319

ACCURACY: 52.4%

IMPROVED MODEL

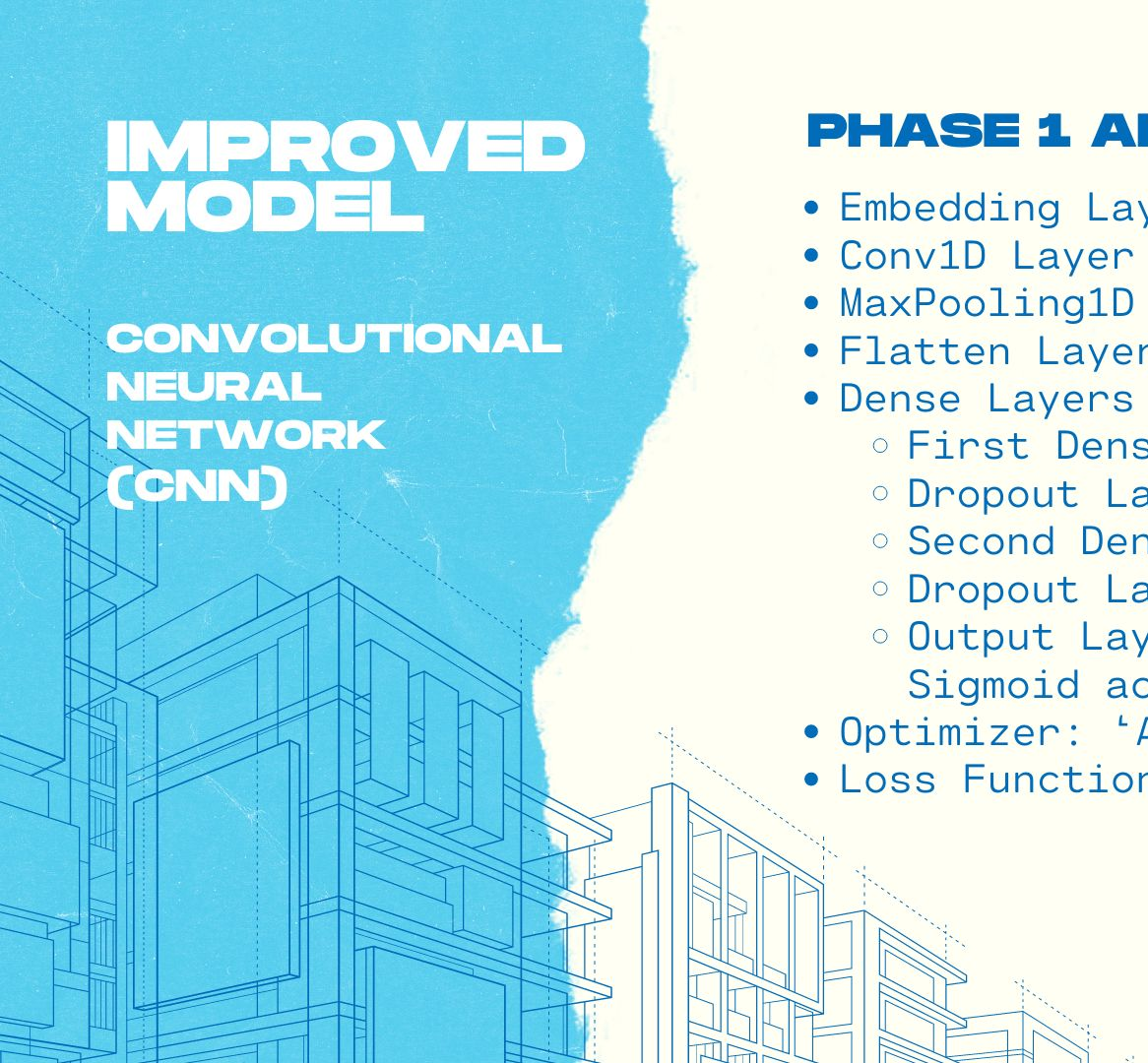
CONVOLUTIONAL NEURAL NETWORK (CNN)

MODEL DIAGRAM



IMPROVED MODEL

CONVOLUTIONAL NEURAL NETWORK (CNN)

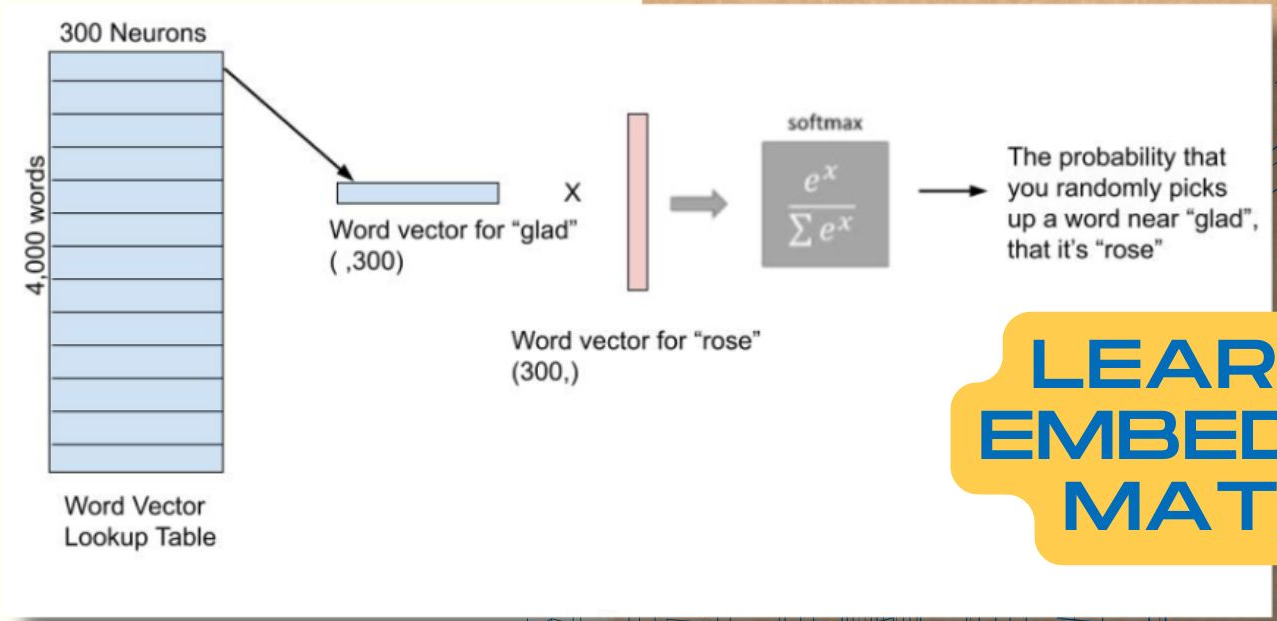


PHASE 1 ARCHITECTURE

- Embedding Layer
- Conv1D Layer
- MaxPooling1D Layer
- Flatten Layer
- Dense Layers
  - First Dense Layer: 128 neurons
  - Dropout Layer: 50% dropout rate
  - Second Dense Layer: 64 neurons
  - Dropout Layer: 50% dropout rate
  - Output Layer: 2 neurons with Sigmoid activation
- Optimizer: ‘Adam’
- Loss Function: Binary Crossentropy

WORD2VEC EMBEDDING LAYER

- Pre-trained: ‘word2vec-google-news-300’ (Trained by Google News dataset)
- Each word is represented by a 300-dimensional vector



LEARNED EMBEDDING MATRIX

HYPERPARAMETER TUNING RESULTS

PERFORMANCE

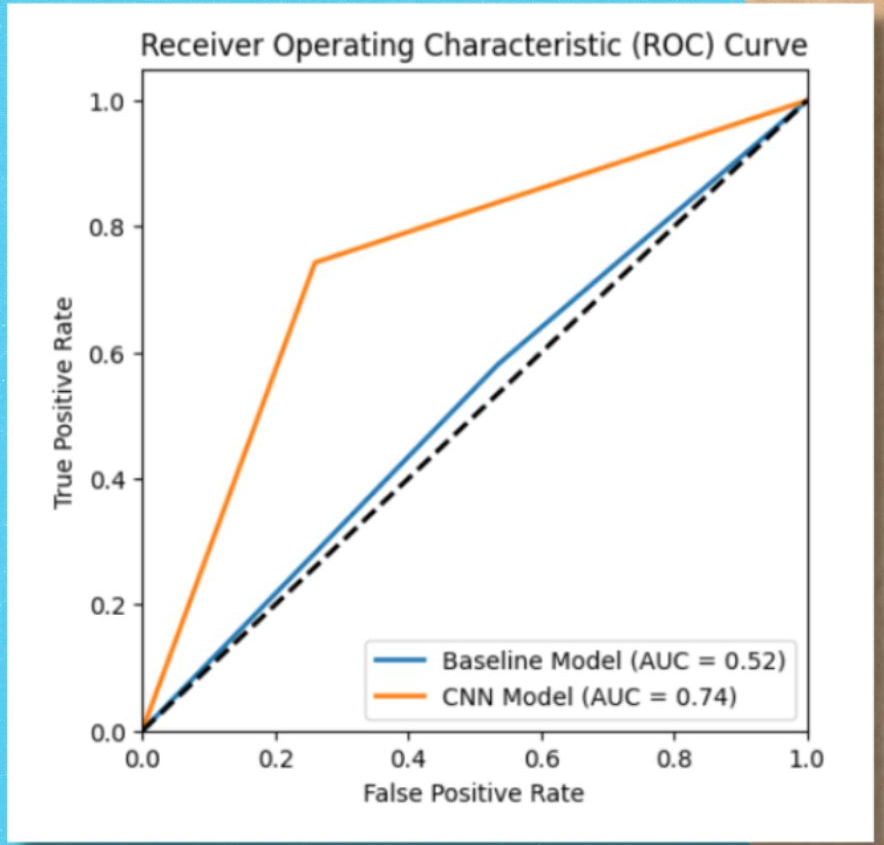
TEST LOSS: 0.5301

TEST ACCURACY: 74.3%

HYPERPARAMETERS	EXPERIMENTS	FINDINGS	FINAL
Filter Size	[64, 128, 256]		256
Kernel Size	[3, 5, 7]		3
Pooling Method	[Max, Avg]		Max
Pooling Size	[2, 5]	Smaller pooling size is better for shorter sequence (<100).	2
Optimizer	[Adam, SGD]		Adam
Epoch	[5, 10, 15]	Longer training epoch bears greater risk in overfitting.	5

PERFORMANCE COMPARISON

	Baseline Model	CNN Model
Accuracy	0.52375	0.74150
Precision	0.52345	0.74151
Recall	0.52375	0.74150
F1 Score	0.52211	0.74150



THE "NO NEUTRALS" BARRIER

- **60%** of dataset is correctly labeled positive or negative
- **20%** have neutral sentiment but are labeled as positive
- **20%** have neutral sentiment but are labeled as negative

The maximum attainable accuracy with this dataset is ~80%.

ADDITIONAL TINKERING

1. Run an analysis only using emoticons
2. Data augmentation to address missing neutral values (with the goal of running a multiclass analysis and recoding "true neutrals")

EMOTICONS ONLY

DATA PREP

- Regex pattern to identify emoticons like ":)", ":-P", and "XD" and store them in a new column “emoticons”
- Combine Text and Emoticons

EVALUATION

- Despite adding emoticons, the model accuracy did not improve
- Suggests emoticons may not be strong sentiment signals in this dataset

QUALITATIVE ANALYSIS

"POSITIVE" RATING 0.999 AND ABOVE

“get followers a day using wwwtweeteraddercom once you add everyone you are on the train or pay vip”

“heyy whats up”

“get followers a day using wwwtweeteraddercom once you add everyone you are on the train or pay vip”

“heyy girl follow me”

"NEUTRAL" RATING 0.499 – 0.501

“its fine ill get over it somehow”

“gosh i hate my horrible lets fix it are you two having fun xd”

“time for a cup of tea and im going to turn the day around”

"NEGATIVE" RATING BELOW 0.0001

“left min early cancelled next train”

“cancelled on me”

“after tweetdeck updated the day it died”

“home and about the hit the day bummed that paris hasnt gotten my messages”

TAKEAWAYS

- This model suffers from the badly-labeled data that was put into it
- To build a good model from this data, another step would be needed
- We imagine using an unsupervised method to find tweets with truly “neutral” sentiments. This would allow us to filter these tweets out, leading to better-labeled data

TEAM CONTRIBUTIONS

- **Bernardo Cobos:** Model Performance Evaluation, Qualitative Analysis of Dataset, Next Steps
- **Rebecca Bargiachi:** Topics and database searching, Existing Literature Review, Model Performance Evaluation, Presentation Slides
- **William Lei:** Data Preprocessing, EDA, Improved Models, Visualizations, Emoticon-Only Analysis
- **Xueying Tian:** Baseline Model, Improved Models, Visualizations, Model Performance Evaluation, Presentation Slides