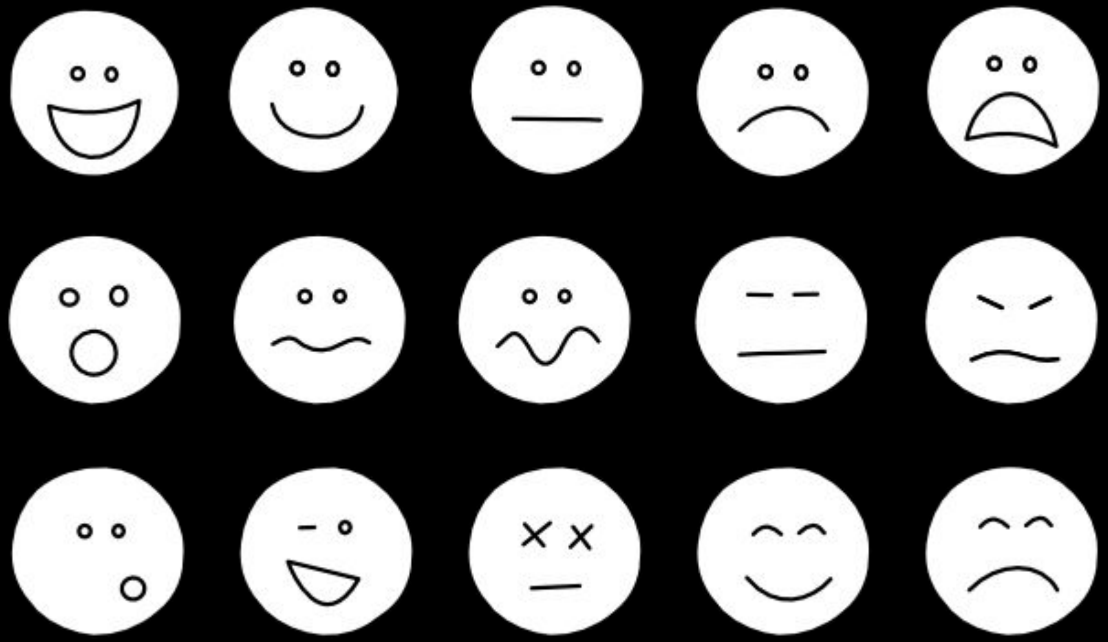
ProQuest Sentiment Analysis MDP

Taylor Murray

Honors Capstone

Dec 16th, 2020



ProQuest

Context

Our Sponsor

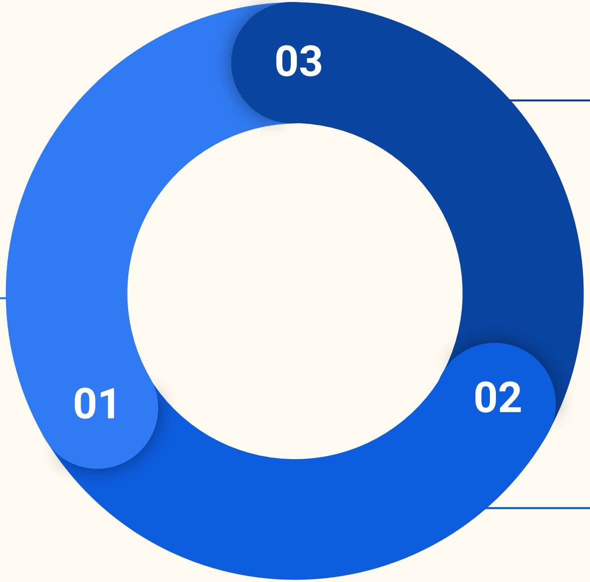
- Education technology company
- Content collection that encompasses 90,000 authoritative sources, 6 billion digital pages and spans 6 centuries

Motivations

- Enhance user experience in the TDM (text-data-mining) environment
- Improve ProQuest’s internal use of data

High Level Goals

Help ProQuest users (commonly academic researchers) create word embeddings and predict sentiment on their own datasets.



Evaluation

Compare which models performed the best on various sentiment analysis datasets.

Word Embeddings System

Generate word embeddings on ProQuest's corpora

Sentiment Analysis System

Detect emotions expressed in the documents

Word Embeddings

Word embeddings (WEs) are representations of words in the vector space.

Dimensionality of the vector is a hyperparameter, typically chosen to be between 50-300.

Each word embedding takes up constant space.

	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

Properties of Word Embeddings

Semantic property: Related to a word's *meaning*

e.g. *doctor, child, woman* → *human*

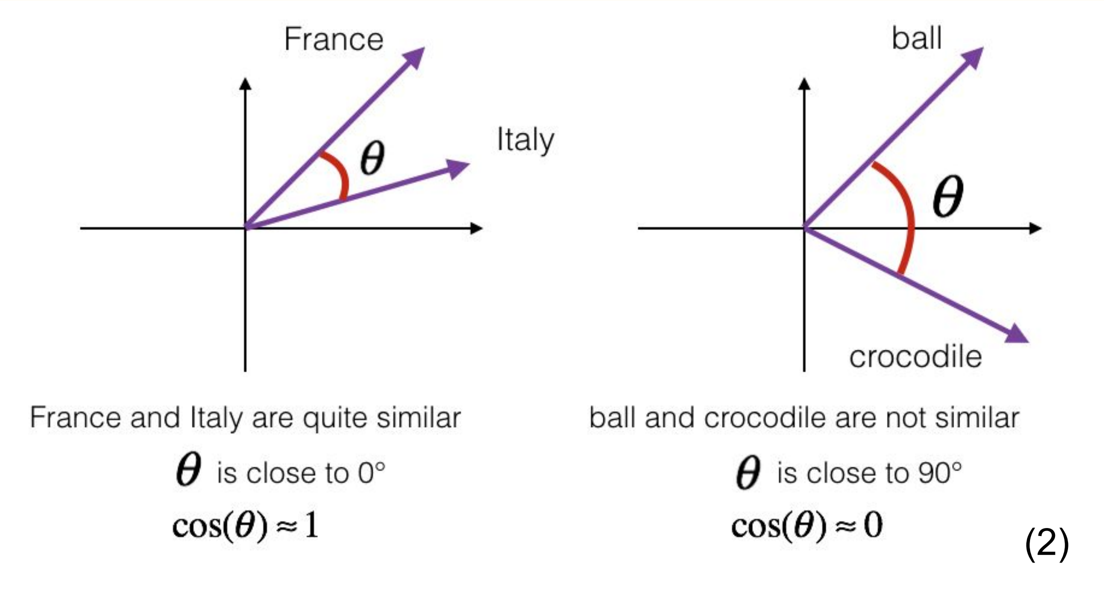
Syntactic property: Related to a word's *structure* and *grammatical features*

e.g. *geese, children, apples* → *plural nouns*; *ran, helped, changed* → *past tense*

Linear Behaviour

Vector("King") - Vector("Man") + Vector("Woman") ≈ Vector("Queen")

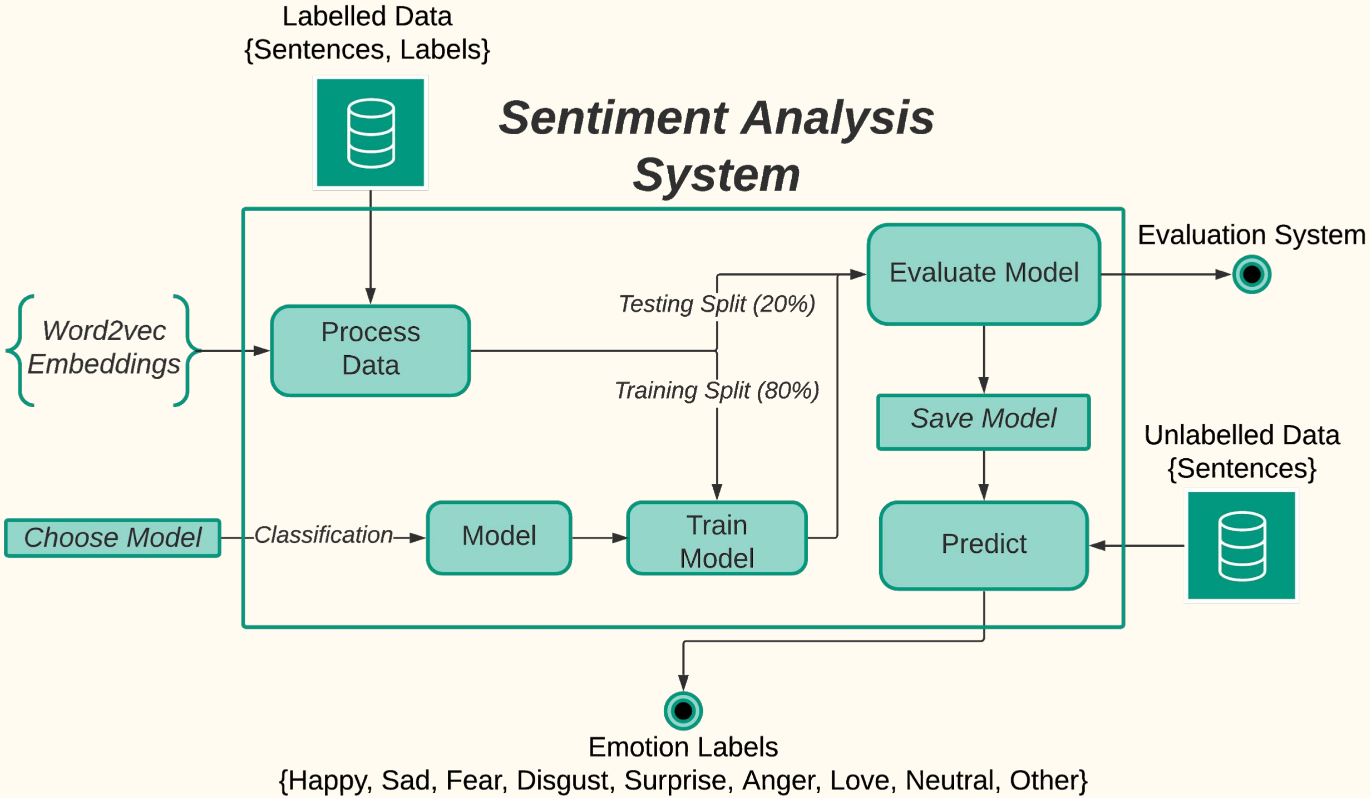
Cosine Similarity



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Calculate similarity between different words

Sentiment Analysis System



Emotion prediction

Goal: Classify text as having one of nine sentiments

- Happy
- Sad
- Fear
- Disgust
- Surprise
- Anger
- Love
- Neutral
- Other

Ekman's 6 Basic Emotions[3]

Example Text:

Sentiment:

Happy

ProQuest corpora

1. New York Times

2. Book Blurbs

3. LION (Literature Online) Poems

Key Question:

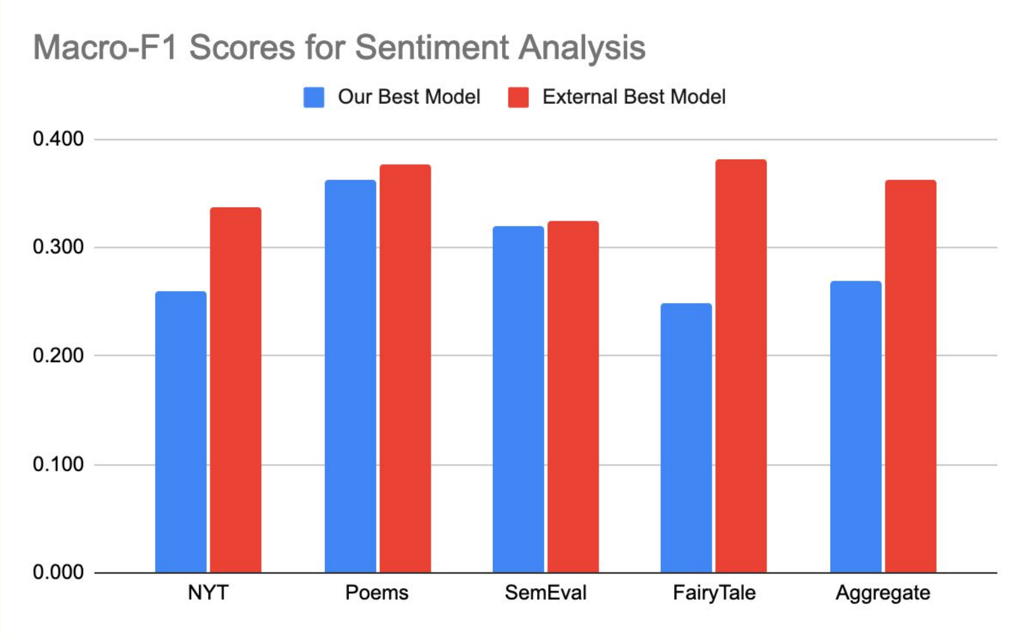
Can using **in-domain** corpora for word embeddings improve sentiment analysis performance?

What did we discover?

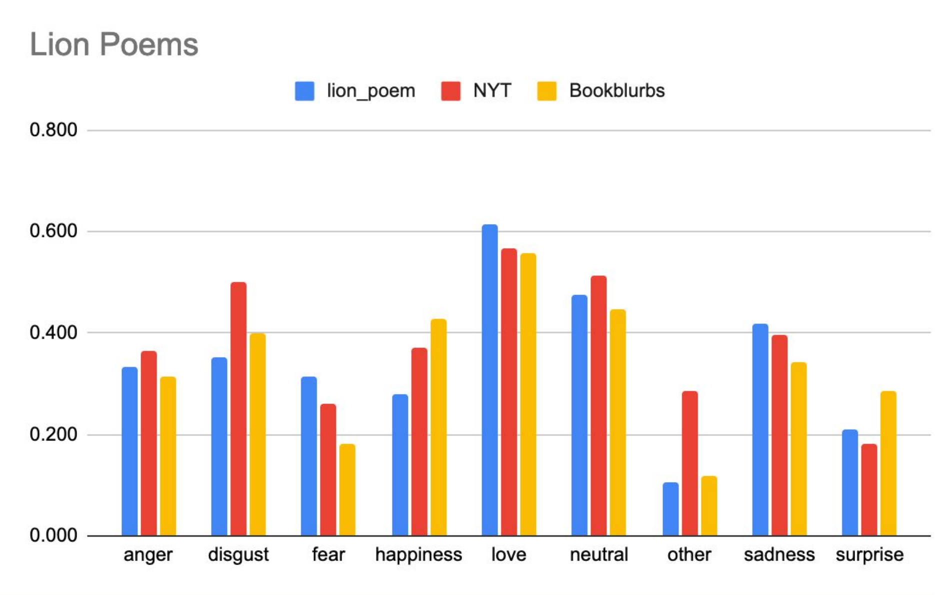
We tested 18 models, 9 of which used custom word embeddings.

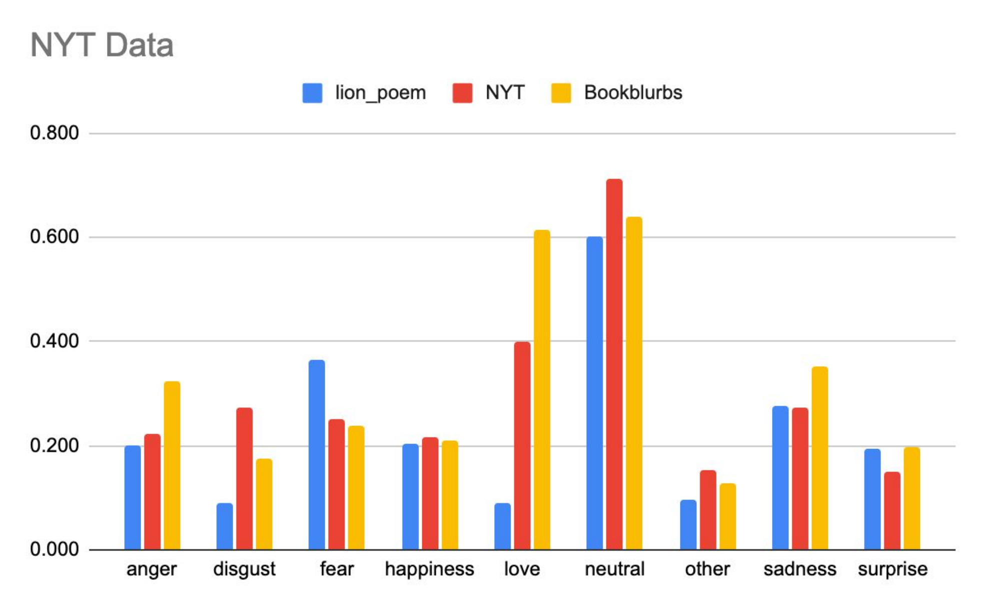
We used different sampling techniques, filtering methods, ML models, and sentence embeddings.

In the end, it seemed conclusive that more generally trained models performed the best.



Using in-domain corpora not correlated with better performance on sentiment tasks





End Deliverables

- Made Jupyter Notebooks for:
  - Word Embeddings Generation
  - Word Embeddings Usage
  - Predicting affective state (“happiness”,“sadness”, etc.)
  - Predicting valence state (“very negative”, “positive”, etc.)
  - Visualizations for Sentiment Analysis Results
- Set of word embeddings
- Set of sentiment analysis models
- Emotion-labelled data

Lesson: You should record **all** of your parameters and brainstorm new ones you might try later

The parameters we noted:

```
parameters
----------
Note: param2 is deprecated because we haven't named all the
      feed in an array of paths to the models

param1: a list of datasets ['Stanford', 'semEval_emotions']
param2: choose one WE model ['cbow', 'sg-lion', 'sg-1m', 'sg
param3: a list of SEs ['AVG', 'SIF', 'concat', 'SBERT']
param4: choose ML models -> ['LR', 'SVM', 'Tree', 'LSTM']

return value: Dict['param1']['{param2}_{param3}']['param4']
Exceptions: when param4 == LSTM, then Dict[param1][param2][p
"""
```

Parameter specification

The parameters we failed to note:

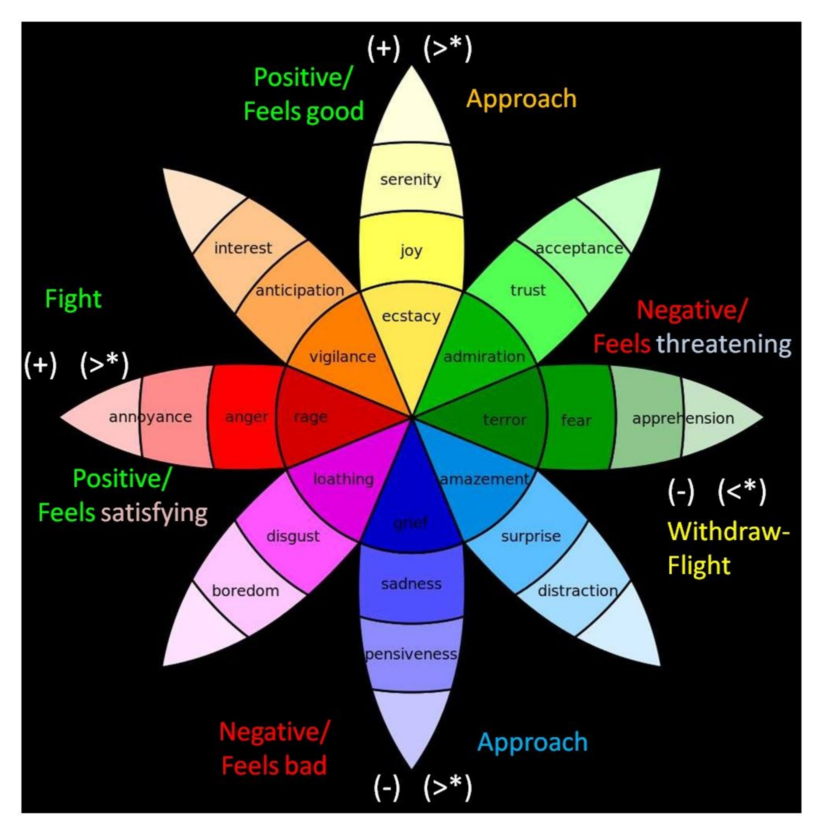
- What data was trained on
- If there was over/undersampling done
- Whether or not the data trained on was stratified
- What the ML model was maximizing (F1-score or accuracy)

How I got involved?

- Interested in machine learning
- Wanted to try my hand at research
- Enjoyed the psychological aspect

Next steps for the project

- Emotion as a vector
- Using word embeddings for optical character recognition
- Neutral filter



https://inchoatepsyche.files.wordpress.com/2015/06/modified-plutarch-wheel3.jpg

Thanks!

Sponsor Partners: John Dillon and Dan Hepp

Advisor: Sugih Jamin

Team: Pranay Shah, Vishnu Nair, Arun Annamalai, Eamy Mo, Rakshit Gogia, Sebastian Jin