Deriving Sentiment from English Texts for the Creation of Digital Collages

Tianen (Benjamin) Liu
Advisor: Dr. Natalia Khuri

WAKE FOREST
UNIVERSITY

Computer Science Honors Presentation | April 20, 2020

Introduction & Motivation

- Natural Language Processing
- Phrases / Paragraphs -> Sentiment
- Modern applications

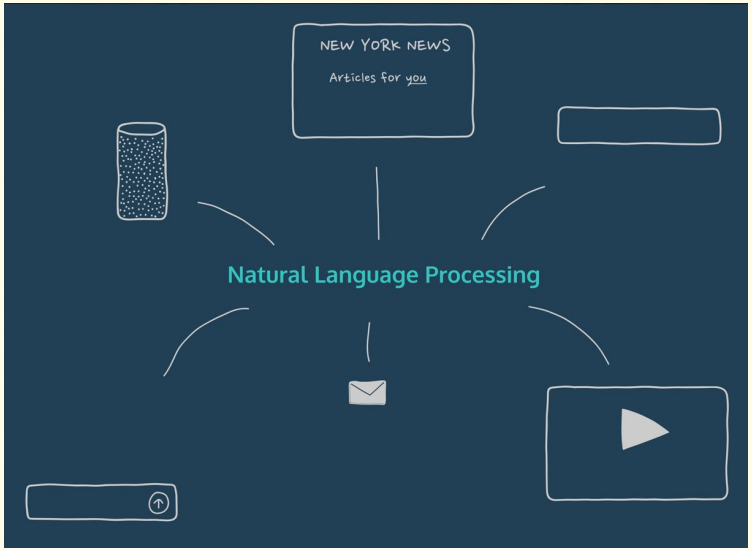


Figure 1

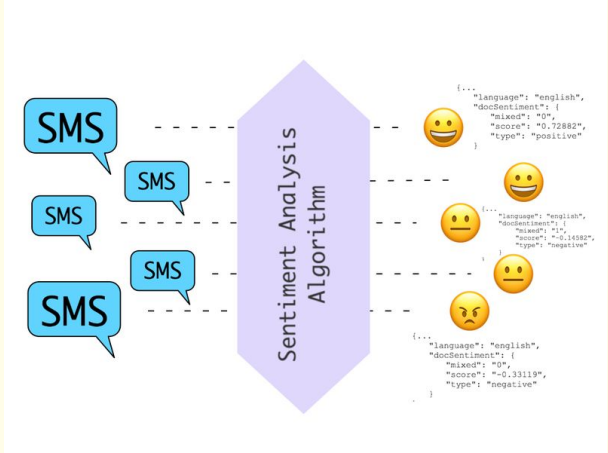


Figure 2

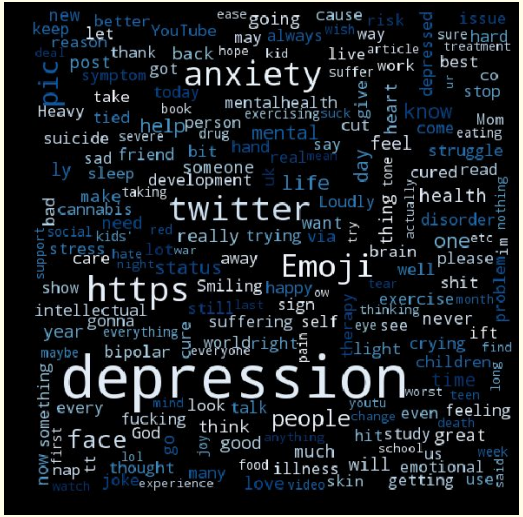


Figure 3

Figure 1: Vijay, https://towardsdatascience.com/a-gentle-introduction-to-natural-language-processing-e716ed3c0863

Figure 2: https://www.twilio.com/docs/glossary/what-is-sentiment-analysis

Figure 3: Martinez, https://medium.com/datadriveninvestor/a-machine-learning-approach-for-detection-of-depression-and-mental-illness-in-twitter-3f3a32a4df60

Research Objective

Explore machine learning methods to predict sentiments in different contexts

Specific Aims

1. Train 3 machine learning algorithms with a data set of movie reviews
2. Compare the performance of trained models
3. Test their accuracy with the data sets from different domains

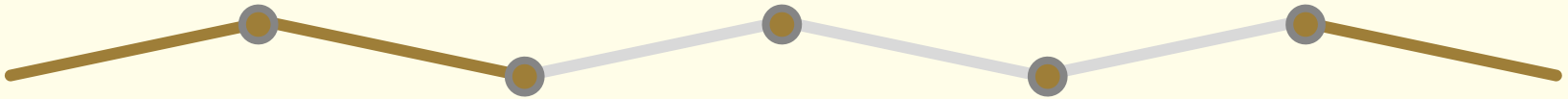
WAKE FOREST UNIVERSITY

Methodology Overview

Natural language processing and text analysis background

Experimentation with 3 different classification methods

Testing the performance of the 3 methods on a different data set *Sentiment140* and novels



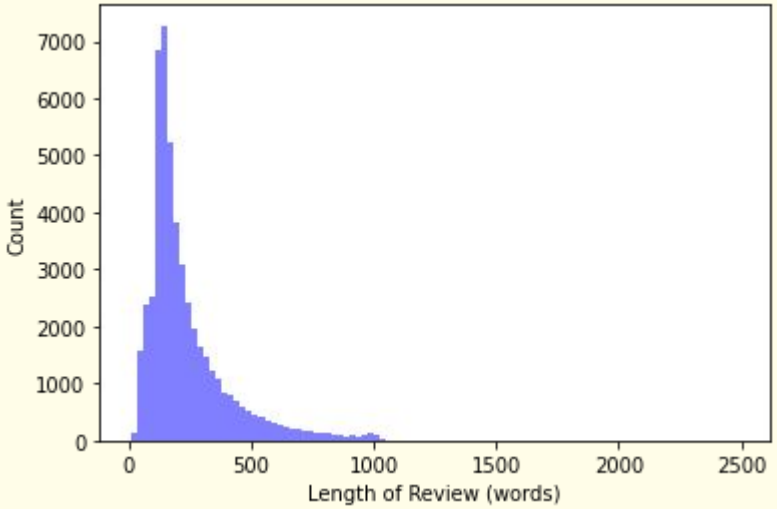
Data collection: transition from psychiatric text to *IMDb Movie Reviews*

Exploration of Long Short Term Memory Network architecture

IMDb: Internet Movie Database

Data Description

- **Source:** Internet Movie Database (IMDb) Movie Reviews
- **Pre-processing**:
  - remove stopwords and punctuation
- **Data Overview:**
  - training size: 25000 reviews (50% positive, 50% negative)
  - median length: 78 words
  - minimum length: 2 words
  - maximum length: 1079 words
  - length distribution



Method

Algorithms used in this work

- Naïve Bayes Classifier
- Turney Algorithm
- Long Short Term Memory (LSTM) Network

Naïve Bayes Classifier

Method: Naïve Bayes Classifier

- Bayes Rule & Naïve Bayes

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$c_{MAP} = argmax_{c \in C} P(c|d) = argmax_{c \in C} \frac{P(d|c)P(c)}{P(d)} = argmax_{c \in C} P(d|c)P(c)$$

likelihood prior

Notation

c: a class
C: all classes
c_{MAP}: max a posteriori (MAP) estimation
d: data (a set of texts)
w: a word in text
$argmax_c$: the class c that makes the following value maximum
V: vocabulary, distinctive words
$|V|$: size of vocabulary

- Likelihood

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)} = \frac{count(w_i, c_j) + 1}{(\sum_{w \in V} count(w, c_j)) + |V|}$$

Example: Naïve Bayes Classifier

- Vocabulary = 11
- Prior: probabilities of classes
  - P(c_positive) = 2/4 = ½
  - P(c_negative) = 2/4 = ½

- Likelihood: Conditional probabilities of individual words

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)} = \frac{count(w_i, c_j) + 1}{(\sum_{w \in V} count(w, c_j)) + |V|}$$

Training	Data 1	i'm happy	Positive
	Data 2	avengers movie so good	Positive
	Data 3	hate movie very boring	Negative
	Data 4	college so stressful	Negative
Test	Data 5	college life so happy	Positive

P(college | positive) = (0+1) / (6+11)
P(life | positive) = (0+1) / (6+11)
P(so | positive) = (1+1) / (6+11)
P(happy | positive) = (1+1) / (6+11)

P(college | negative) = (1+1) / (7+11)
P(life | negative) = (0+1) / (7+11)
P(so | negative) = (1+1) / (7+11)
P(happy | negative) = (0+1) / (7+11)

P(positive | data) = 0.0000239

P(negative | data) = 0.0000191

Turney Algorithm

Method: Turney Algorithm

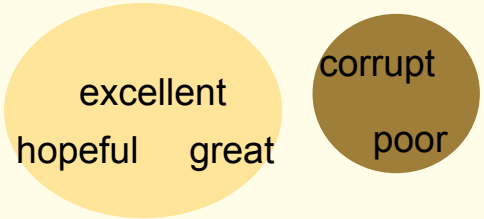
- Part-of-speech tagging
- Bigram Extraction
  "a beautiful day at Wake Forest"

1st word	2nd word	3rd word (not extracted)
JJ *(adjectives)*	NN, NNS *(nouns)*	anything
RB, RBR, RBS *(adverbs)*	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN, NNS	JJ	not NN nor NNS
RBm RBR, RBS	VB, VBD, VBN, VBG *(verbs)*	anything

- Learning the polarity of each phrase: Pointwise Mutual Information (PMI)

$$Polarity(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$

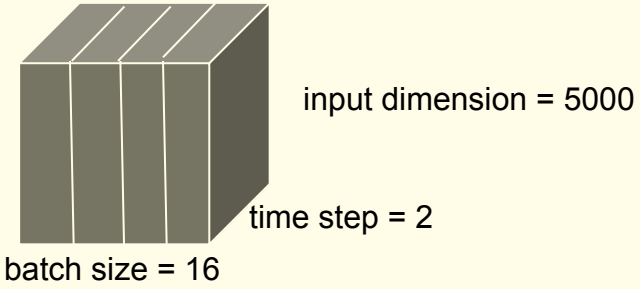
$$PMI(w1, w2) = log_2 \frac{hits(w1 \text{ NEAR } w2)}{hits(w1)hits(w2)}$$



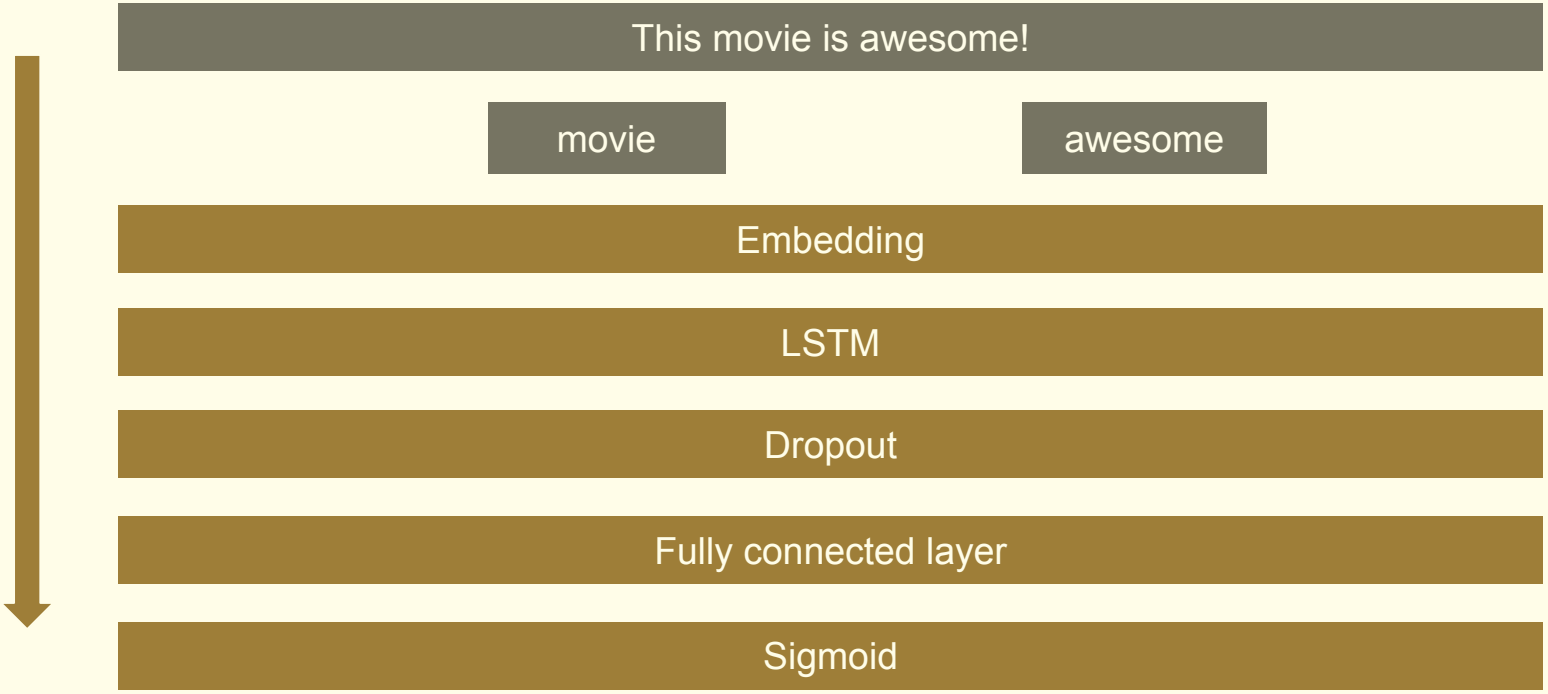
Long Short Term Memory Network

Method: LSTM Network

- **Pre-training**
  - Padding: maximum input word length = 160
- **Parameters**
  - embedding size = 64
  - time step = 2
  - input dimension = vocabulary size = 5000
  - input unit = maximum input word length = 160
  - batch size = 16
  - LSTM input unit = 100
  - dropout = 0.5

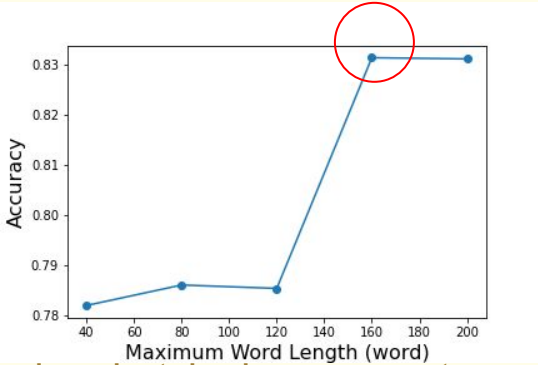


Architecture: LSTM Network

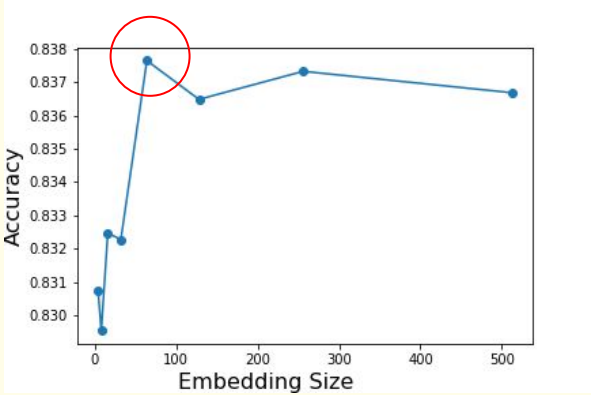


Parameter Tuning: LSTM Network

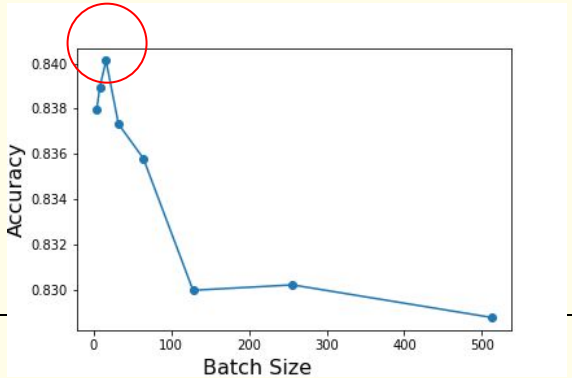
1. Explore maximum word length parameter



2. Explore embedding size parameter



3. Explore batch size parameter



Aim 2: 5 Fold Cross Validation Results

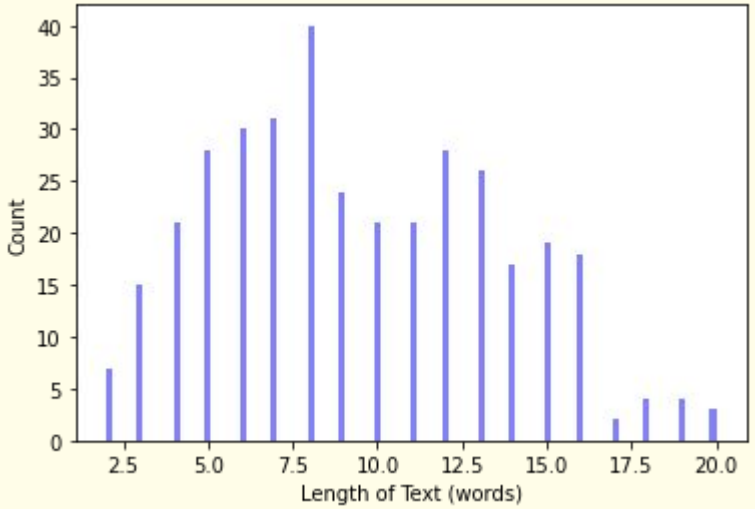
Comparison of the 3 methods

	Naïve Bayes	Turney Algorithm	LSTM Network
5-fold CV accuracy	83.73%	50.87%	86.17%
Standard deviation	0.0029	0.011	0.016

Aim 3: Application to Different Domain

Sample testing data of size 359 from *sentiment140*, a Twitter sentiment dataset

- Text length distribution



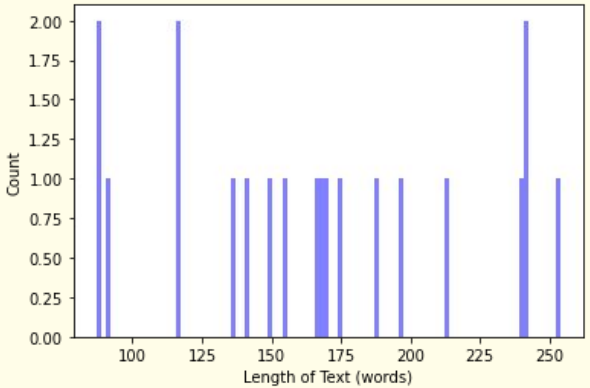
- Data format: csv
  - 1st column values of 0 (neg) and 4 (pos)
  - last column of texts
- Accuracies
  - Naïve Bayes accuracy: 53.48%
  - Turney Algorithm accuracy: 37.68%
  - LSTM accuracy: 52.09%

4	26	Mon May 11	iphone app	CocoSavanna	downloading apps for my iphone! So much fun :-) There literally is an app for just about anything.
4	33	Mon May 11	visa	DreambigRa	good news, just had a call from the Visa office, saying everything is fine.....what a relief! I am sick of scams out there! Stealing!
4	34	Mon May 11	fredwilson	andrewwats	http://twurl.nl/epkr4b - awesome come back from @biz (via @fredwilson)
4	35	Mon May 11	fredwilson	fredwilson	In montreal for a long weekend of R&R. Much needed.
4	46	Thu May 14 (	"booz allen"	JoeSchueller	Booz Allen Hamilton has a bad ass homegrown social collaboration platform. Way cool! #ttiv
4	47	Thu May 14 (	"booz allen"	scottabel	[#MLUC09] Customer Innovation Award Winner: Booz Allen Hamilton -- http://ping.fm/c2hPP

Aim 3: Application to Different Domain

Sample testing data of size 10 from self-created novel excerpts

- Text length distribution



- Accuracies
  - Naïve Bayes accuracy: 40.00%
  - Turney Algorithm accuracy: 40.00%
  - LSTM accuracy: 55.00%

- Data format: txt
  Excerpted from *Harry Potter* and *Ulysses*

Neg 8.txt

there was another empty chair too, and Harry couldn't think who else was missing.

"Where's the new Defense Against the Dark Arts teacher?" said Hermione, who was also looking up at the teachers.

They had never yet had a Defense Against the Dark Arts teacher who had lasted more than three terms. Harry's favorite by far had been Professor Lupin, who had resigned last year. He looked up and down the staff table. There was definitely no new face there.

"Maybe they couldn't get anyone!" said Hermione, looking anxious.

Harry scanned the table more carefully. Tiny little Professor Flitwick, the Charms teacher, was sitting on a large pile of cushions beside Professor Sprout, the Herbology teacher, whose hat was askew over her flyaway gray hair. She was talking to Professor Sinistra of the Astronomy department. On Professor Sinistra 's other side was the sallow-faced, hook-nosed, greasy-haired Potions master, Snape –

Conclusion & Future Work

- In 5 fold cross validation, LSTM has the best performance, followed by Naïve Bayes and Turney Algorithm.
- Domain transfer is a challenge. LSTM and Naïve Bayes trained with movie reviews and Turney Algorithm fail to predict sentiments in other contexts due to overfitting and difference in phrasing words in different contexts.
- Application of *text2collage* is a challenge.