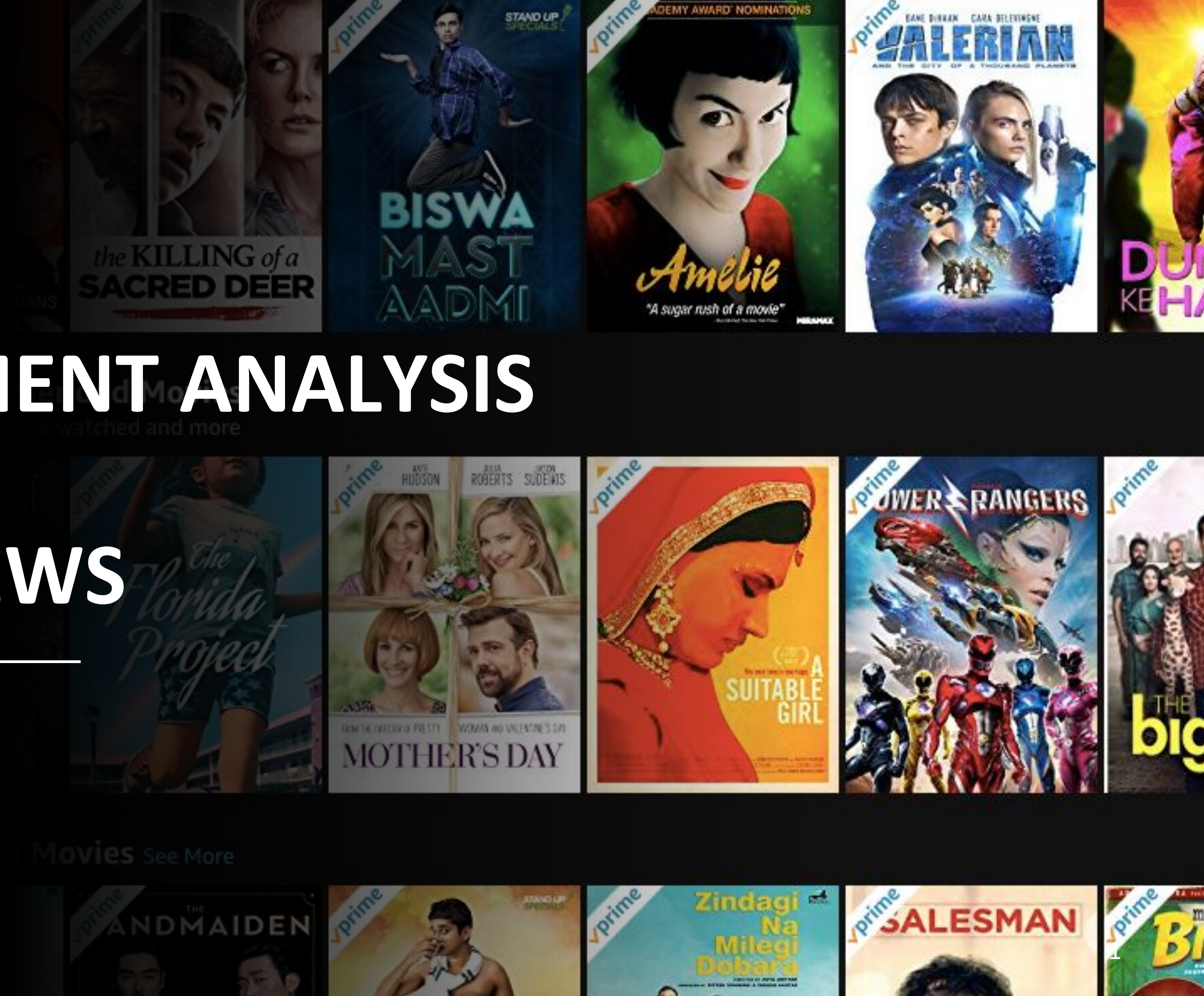
NLP - SENTIMENT ANALYSIS in MOVIE REVIEWS

Thai Linh Bui – Oct.2020



OUTLINE

01 Natural Language Processing (NLP)

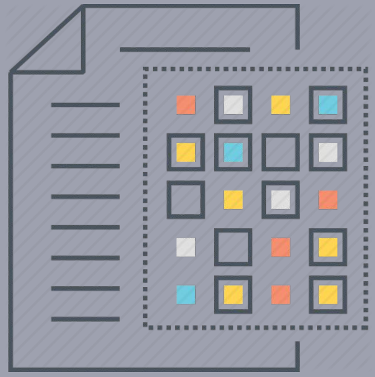
02 Sentimental Analysis

03 Exploratory of the dataset

04 Modeling

80%

Of the world’s data
is UNSTRUCTURED



Source: IBM

1. NATURAL LANGUAGE PROCESSING (NLP)

•Give the machines the ability to read, understand and derive meaning from human languages

•Fields of application:

twitter

Sentiment Analysis

Google YAHOO!

Filter & classify emails

SEPHORA

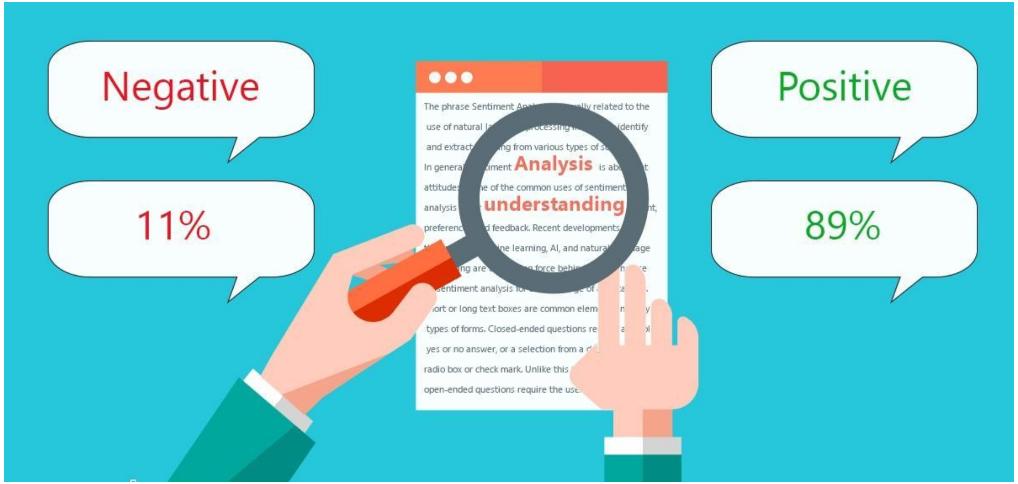
Create chatbot helping customers

amazon

Recognize & predict disease

2. SENTIMENT ANALYSIS

Text analysis technique in machine learning that detects polarity (e.g. a positive or negative opinion) within text.



2. SENTIMENT ANALYSIS & BUSINESS INTEREST

Understand the customers

Crisis management

Monitor market research

Improve products & customer service

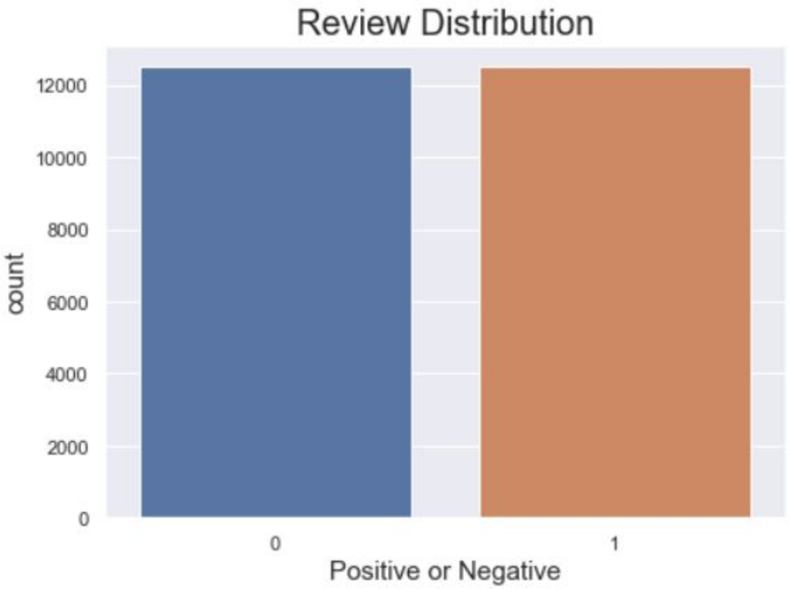
Objective of the project:

Using NLP to predict the opinion of the movie reviewers

🙂 Positive

☹ Negative

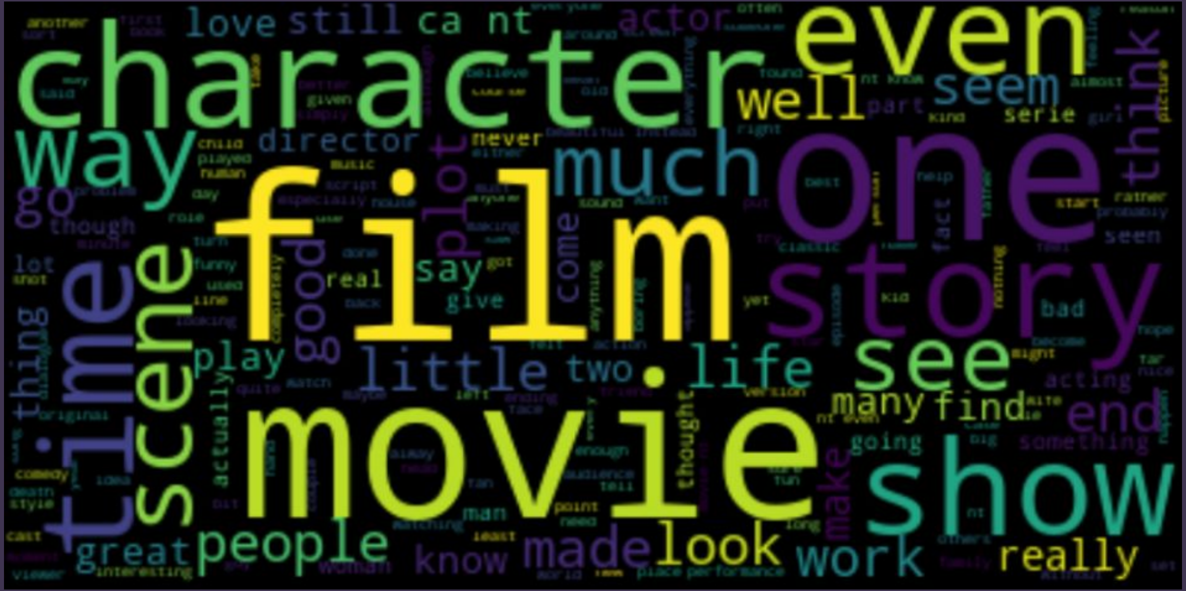
3. EXPLORATORY OF THE DATASET



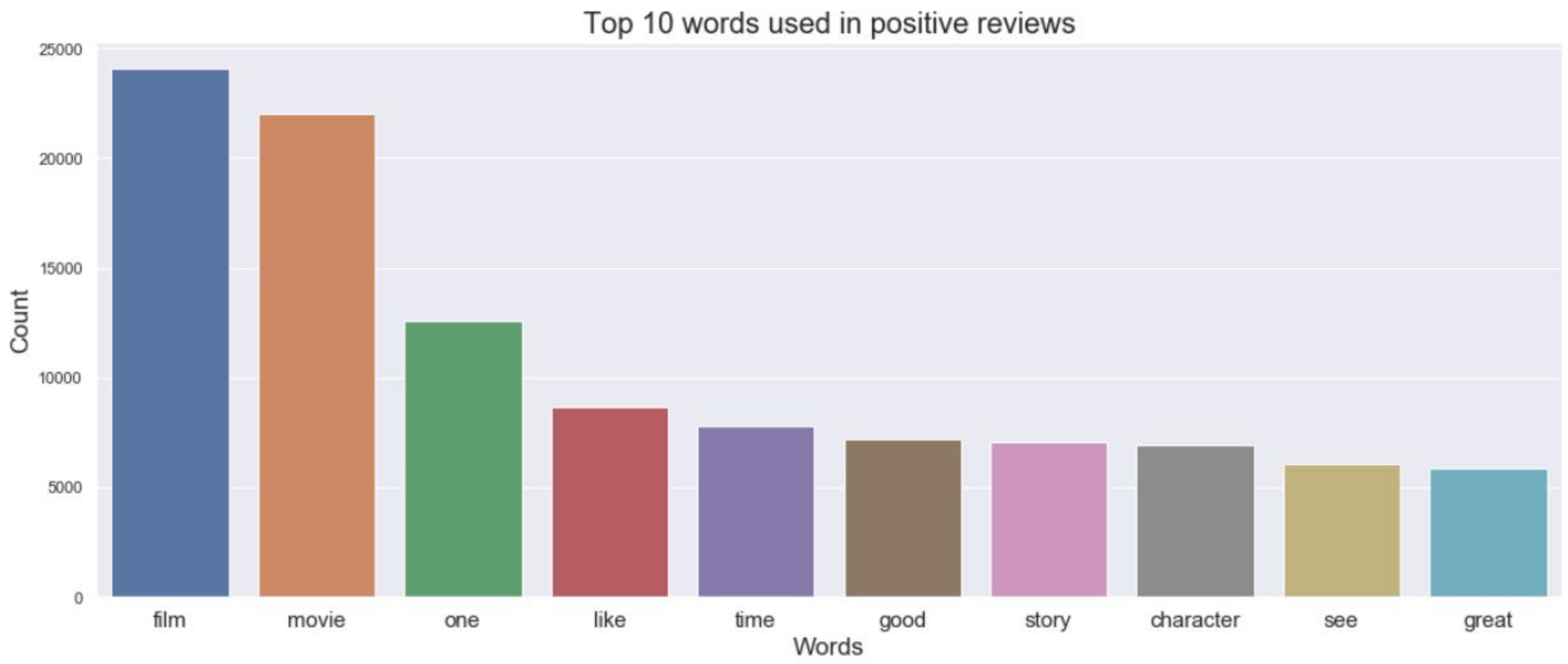
- 25K reviews on IMDB’s website (Source: Kaggle)
- Two columns: Review and Label (Negative or Positive)
- Balance distribution between labels

3. EXPLORATORY OF THE DATASET

The most common words

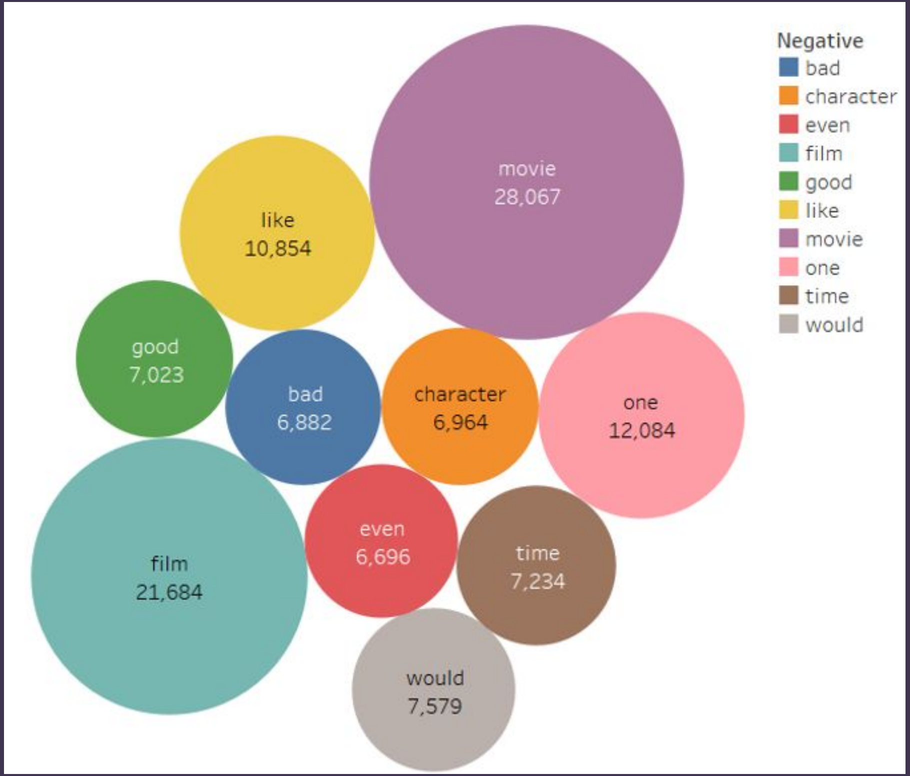


3. EXPLORATORY OF THE DATASET



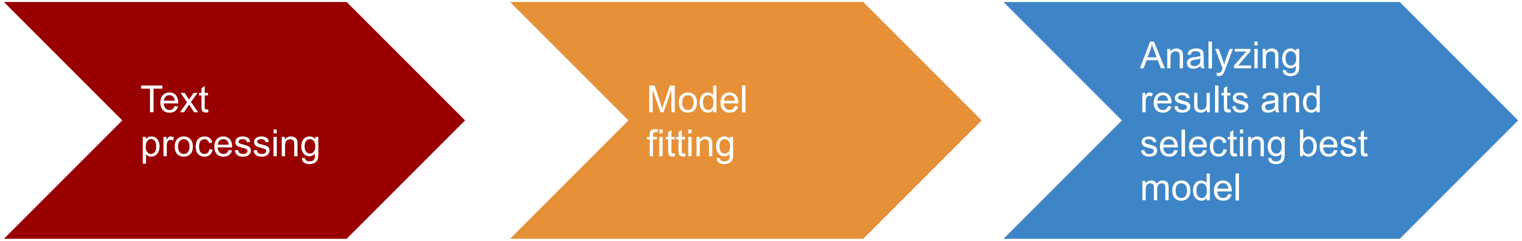
3. EXPLORATORY OF THE DATASET

Top 10 words used in negative reviews



4. MODELING

The modeling process



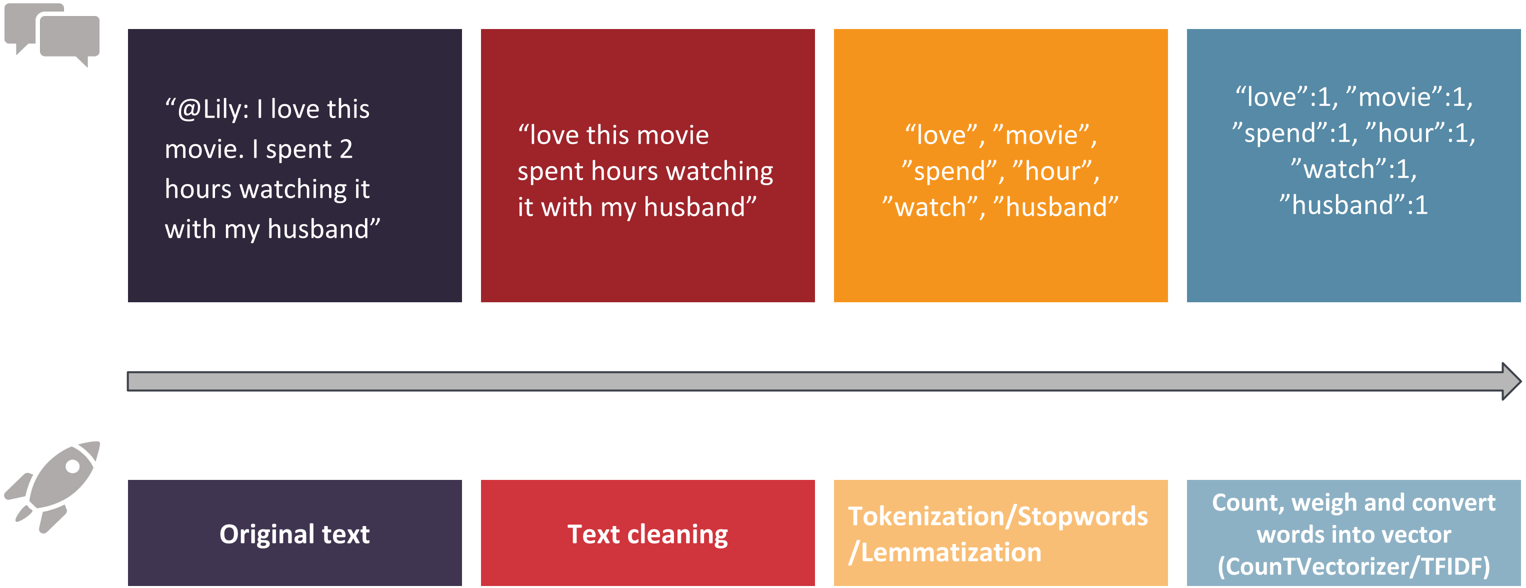
Text processing

Model fitting

Analyzing results and selecting best model

4. MODELING

Text Processing



“@Lily: I love this movie. I spent 2 hours watching it with my husband”

“love this movie spent hours watching it with my husband”

“love”, ”movie”, ”spend”, ”hour”, ”watch”, ”husband”

“love”:1, ”movie”:1, ”spend”:1, ”hour”:1, ”watch”:1, ”husband”:1

Original text

Text cleaning

Tokenization/Stopwords /Lemmatization

Count, weigh and convert words into vector (CounTVectorizer/TFIDF)

4. MODELING

Selected Machine Learning Algorithms

Machine Learning

Naives Bayes (unigram)

Logistic Regression (unigram)

Deep Learning

Convolutional Neural Network

Text input goes through different layers

These layers:

- try to find a pattern or useful information of the data
- reduce the dimensional complexity and still keeps the significant information

The process then returns the outputs

4. MODELING

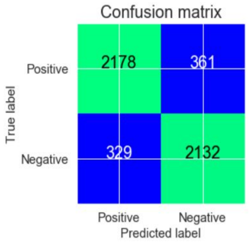
Results

Machine Learning

Naives Bayes (unigram)

Accuracy: 86%

1min



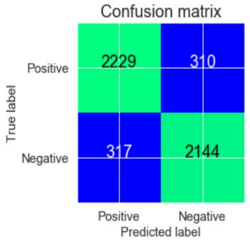
Confusion matrix

True label	Positive	Negative
Positive	2178	361
Negative	329	2132

Logistic Regression (unigram)

Accuracy: 88%

1min



Confusion matrix

True label	Positive	Negative
Positive	2229	310
Negative	317	2144

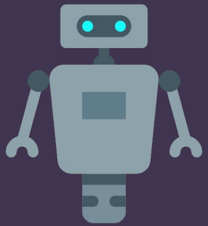
Deep Learning

Convolutional Neural Network

Accuracy: 84%

45 min

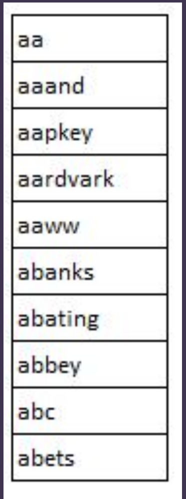
4. MODELING



The most important words according to the machine

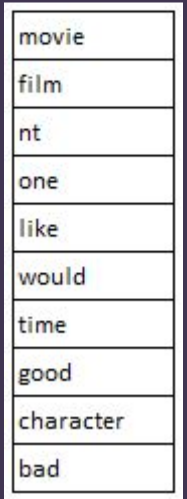
Naives Bayes

🙂



aa
aaand
aapkey
aardvark
aaww
abanks
abating
abbey
abc
abets

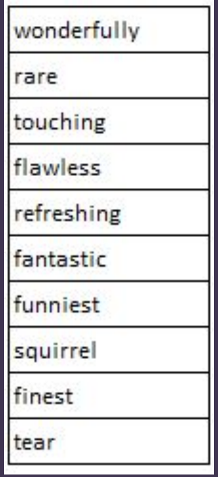
☹



movie
film
nt
one
like
would
time
good
character
bad

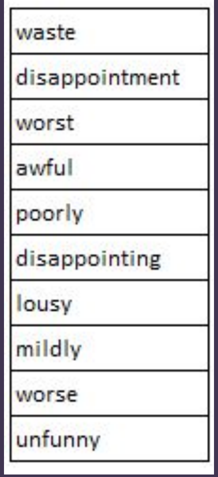
Logistic Regression

🙂



wonderfully
rare
touching
flawless
refreshing
fantastic
funniest
squirrel
finest
tear

☹



waste
disappointment
worst
awful
poorly
disappointing
lousy
mildly
worse
unfunny

4. MODELING

And the champion is:

Logistic Regression (unigram)

Accuracy: 88%

THANK YOU!

APPENDIX

My github: https://github.com/EricBui0201?tab=repositories

Sources:

https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

https://machinelearningmastery.com/develop-word-embedding-model-predicting-movie-review-sentiment/

https://towardsdatascience.com/sentiment-analysis-with-python-part-1-5ce197074184

https://github.com/jps1001/Sentiment-Analysis-On-Movie-Reviews/blob/master/Sentiment%20Analysis%20On%20Rotten%20Tomatoes%20Reviews.ipynb

Edit graph negative words and put in the same place as the Run third model slide 15