Sentiment Analysis

Yingjie Zhu, Vincent Shen, Feifei Han, Clara Chan, Harry Liu

Rotman

Sentiment Analysis Overview

Definition of Sentiment Analysis

Sentiment Analysis is the use of Natural Language Processing techniques to systematically identify and study useful information such as emotion and preferences.

Importance of Sentiment Analysis

Scalability

Centralized criteria

Provide timely analysis

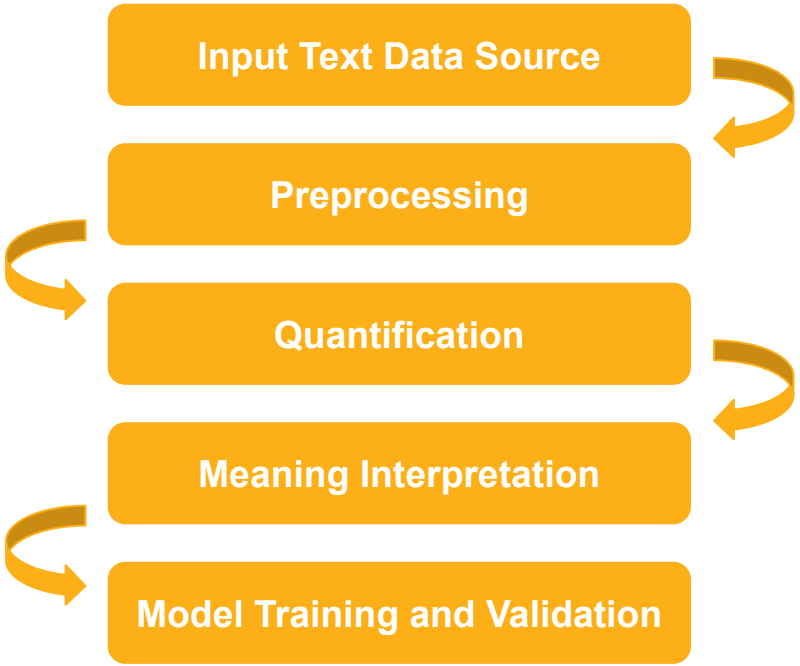
Motivation

Build a pre-trained model that can be applied to a wide variety of textual datasets

- Predicting voter preferences on candidates based on political surveys
- Product improvement using public opinion

Natural Language Processing

Steps and Processes



Introduction and Summary of Methodologies

Three Methodologies

TextBlob

- Python library
- Based on Naive Bayes
- Determines the polarity of phrases ranging from negative to positive

Neural Network

- Based on a paper done by Liu Bing and Mingqin Hu
- Identified the sentiment of opinion sentences
- Used WordNet: a lexical database of semantic relations between words for over 200 languages

TF-IDF

- Converting text data into vector
- Counts the frequency of words to determine sentiment of word (frequency within and across document)
- Two types of algorithm: Naive Bayes and SVM

Data Preprocessing

1. Tokenize the comments into vectors
2. Transform all words into lowercase
3. Remove stop words using Natural Language Toolkit
4. Drop punctuations
5. Punctuations which cannot be dropped are replaced with space.
6. Separate all words using space

	Review	Liked
0	Friendly staff, good food and homely environme...	1
1	Well...The Food was Good__Intrerior design is...	1
2	The man who is foodie like me for him arabian ...	1
3	ordered pizza and they were unable to serve th...	0
4	This place is too much comfortable & food is d...	0

ReviewTokenize
[friendly, staff, good, food, homely, environm...
[well, food, good, intrerior, design, nice, en...
[man, foodie, like, arabian, master, nice, pla...
[ordered, pizza, unable, serve, ordered, set, ...
[place, much, comfortable, food, delicious, ev...

Implementation and Experimentation

TextBlob

TextBlob uses a Movies Reviews dataset in which reviews have already been labelled as positive or negative. The data is trained on a **Naive Bayes Classifier**.

Passed the tokens to a **TextBlob sentiment classifier** which classifies the reviews as positive, negative or neutral by assigning it a polarity between -1.0 to 1.0. If the polarity > 0, the review is classified as positive, and if the polarity < 0, the review is classified as negative.

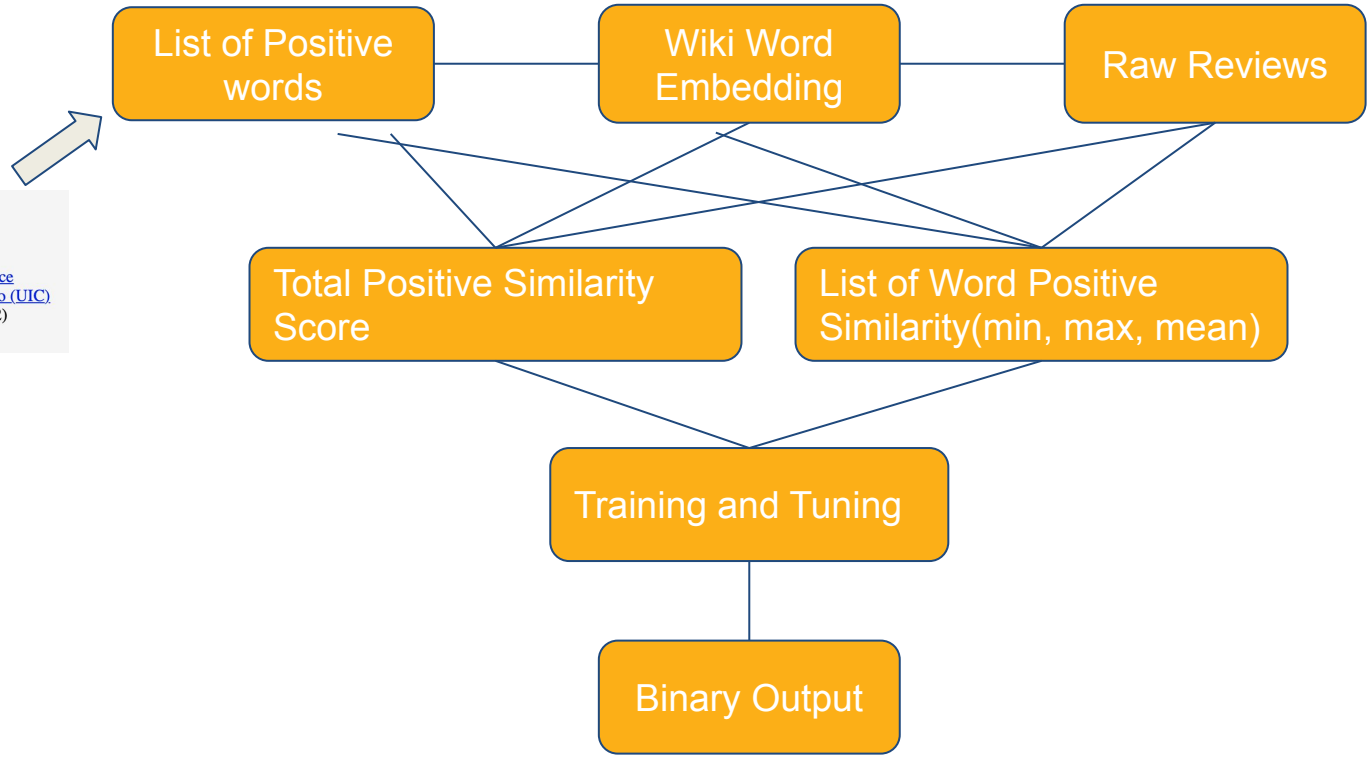
Data Type	Accuracy
Validation Data	66.75%
Test (New Data)	68.82%

Implementation and Experimentation

Neural Network

Bing Liu

Distinguished Professor
Department of Computer Science
University of Illinois at Chicago (UIC)
851 S. Morgan Street (M/C 152)
Chicago, IL 60607-7053



Implementation and Experimentation

Neural Network

Data Type	Accuracy
Validation Data	47%
Test (New Data)	54%

Reasons for Failure

1) Too much preprocessing, **lost meaning** of the original text

2) The research is based on social media data from 2004, which is **too old**

Implementation and Experimentation

TF-IDF with Naive Bayes and Support Vector Machine

Term Frequency - Inverse Document Frequency (TF-IDF)
Term Frequency looks how often each word appears in the document
Inverse Document Frequency is the log to the all document divided by the document words appears

Since TF-IDF is a vectorization method, we need to use classification model to categorize the vectorized data points.

Naive Bayes

Type	Accuracy Score
Validation data	81.42
New data	75.4

Support Vector Machine

Kernel	SVM Accuracy Score (Validation)	SVM Accuracy Score (New Data)
Linear	81.75	72.06
Polynomial	79.44	64.50
rbf	82.27	67.20
sigmoid	79.98	69.30

Performance Comparison

Accuracy

Data Type	Accuracy Scores			
	TextBlob (Naive Bayes)	TF-IDF (Naive Bayes)	TF-IDF (SVM)	Neural Network
Validation Data	66.75%	81.42%	**82.27%**	54.24%
Test (New) Data	68.82%	**73.40%**	72.06%	51.20%

Conclusion

The sentiment analysis is one of the most commonly used NLP to help us understand public opinions towards certain matter. Based on the experiments, there is no “the best” algorithm for text mining.

Recommendation

- According to the performance results, we recommend TF-IDF with both Naive Bayes and SVM

Next Step

- Cross-validation to split dataset
- Customized kernel for SVM
- Train the model based on more datasets across all industries
- Neural Network may perform better if resources are allowed

Thank you!

Any Questions?