Sentiment Analysis on Movie Reviews

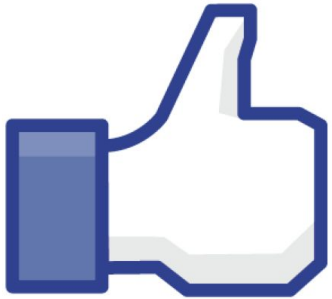
Mitchell Abrams
Shaobo Wang
Zhuoran Wu

Introduction

Introduction

What is sentiment analysis?

Why is it challenging?



“The whole is not necessarily the sum of the parts”. Turney (2002)

Our goal: Achieve high accuracy and discover effective approaches

Literature Review

Background literature and approaches to this task

Liu (2012) *Sentiment Analysis and Opinion Mining*

- Comprehensive overview of the field, current studies, and state of the art methods

Pang et al. (2002) *Thumbs up? Sentiment Classification using Machine Learning Techniques*

- Application of Naive Bayes (NB), Max Entropy (ME), and Support Vector Machine (SVM)

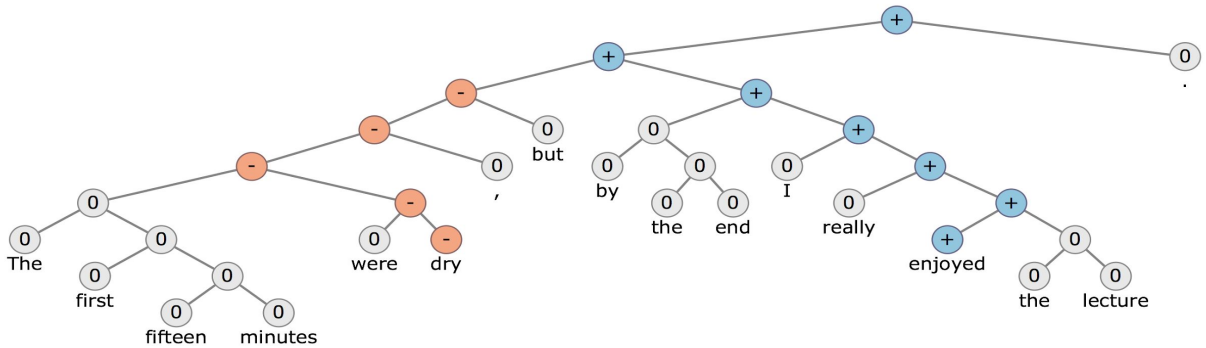
Pang and Lee (2004) *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*

- Ideas from Pang and Lee's method architecture, subjectivity, polarity

Literature Review (Cont)

Socher et al. (EMNLP 2013) *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*

- State of the art methods trained on the Stanford Sentiment Treebank



Methodology

Data Analysis

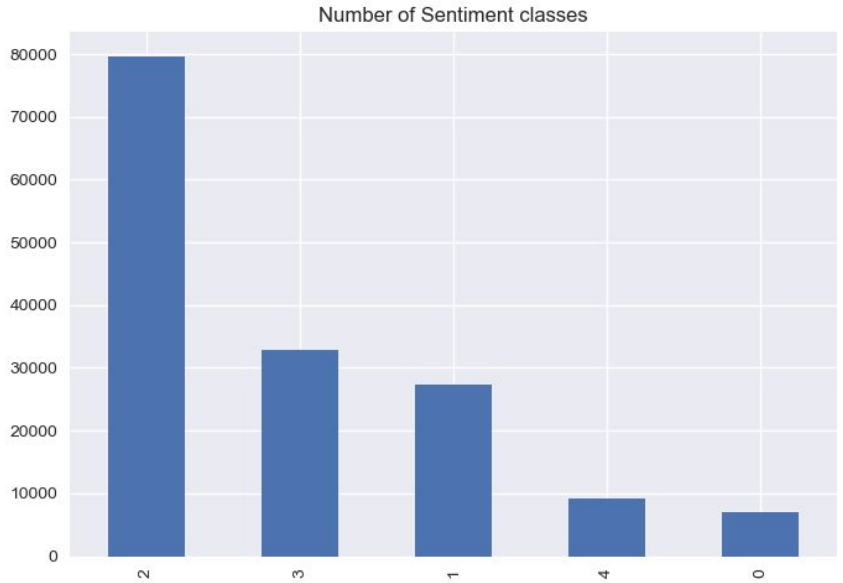
PhraseId	SentenceId	Phrase	Sentiment
1	1	A series of escapades demonstrating the adage that what is good for the goose is also good for the gander , some of which occasionally amuses but none of which amounts to much of a story .	1
2	1	A series of escapades demonstrating the adage that what is good for the goose	2
3	1	A series	2
4	1	A	2
5	1	series	2

Data Analysis

156060 examples

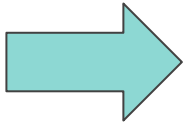
8544 complete sentences

Unbalanced label distribution



Data Analysis (Cont)

- 0 - negative
- 1 - somewhat negative
- 2 - neutral
- 3 - somewhat positive
- 4 - positive



- 0 - negative
- 1 - neutral
- 2 - positive

Feature engineering and ideas

N-grams:

Unigrams: Captures salient positive and negative words: “good”, “bad”, “awful”, “terrific”, etc.

Bigrams: Naive way to capture negation and short phrases: “not bad”, “not good”, “very good”, etc.

Trigrams: “not very good”, “best movie yet”, “must see again”, etc.

4-grams/ngrams: “really not that great”, “this is too much”, “a little too far”, etc.

Character n-grams: “(b)est”, “(wor)st”, “ok”

Feature Engineering

Harvard Inquirer Lexicon

- Lexicon/dictionary that has various labels/categories for words, including ‘Positiv’, ‘Negativ’, ‘IAV’, and ‘strong’

WordNet

- Map words to their synsets so phrases are represented as synset vectors

Feature engineering - Vectorization

Vector types

- Count Vectorizer
- TFIDF Vectorizer

Future implementations...

- Incorporate POS tags
- Or... Word2Vec
- Word2Vec from Corpus
- Pre-trained Word Embedding

Feature engineering and ideas

Applying Stop Word List: “of”, “a”, “to”, “the”, etc.

Alpha smoothing: Adjust smoothing hyperparameter

Other features not applied:

Aspect engineering: mapping opinions to their target entities

Exploit discourse level features from multilayered corpora

Position of words in the overall text

Models

Models- Naïve Bayes

Generative Model

- Generative model that predicts the likelihood that an event will occur- sentiment is negative, neutral, or positive.
- Predictors are independent of each other
- Refining the classifier: n-gram features, stop word list, hyperparameter- alpha smoothing.
- Naive Bayes algorithm from Sci-kit Learn ML library

Model- SVM

A discriminative, non probabilistic model

- Features are not probabilities but points in vector space
- Draws a hyperplane in vector space that maximizes the margin between the data points
- hyperplane is adjusted until optimization is achieved
- Loss function- max margin
- SVM algorithm from Sci-kit Learn ML library

Models- XGBoost

Gradient boosting decision tree algorithm

- Gradient boosting- ensemble method
  - Adds predictors and iteratively adds on models on top of each other
  - Correct errors from previous models and updates model with gradient descent

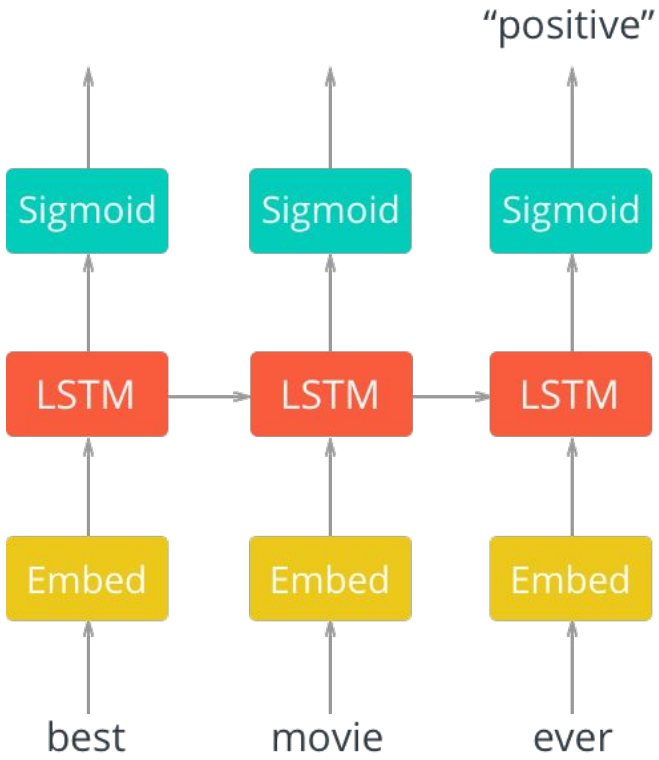
- Doesn't allow for feature engineering or hyper parameter tuning
- But fast tool for learning better models and popular in Kaggle competitions

Models- Random Forest

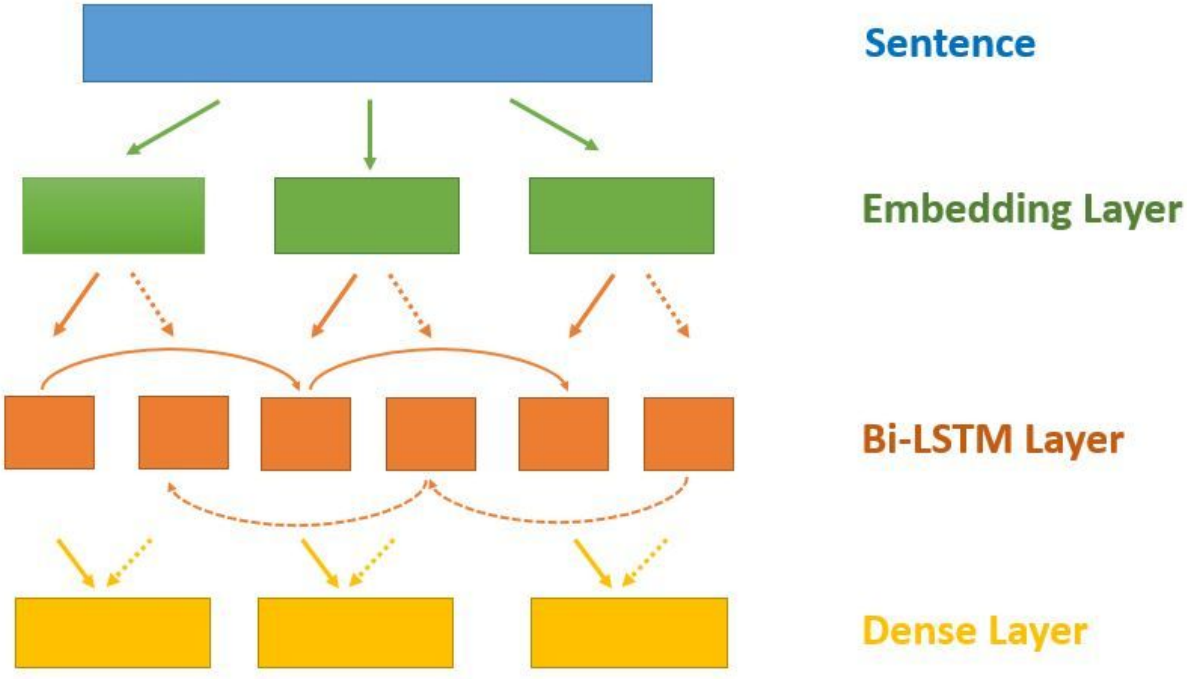
Decision Tree Classifier

- Ensemble learning method- divide and conquer approach
- Constructs many decision trees and predicts the class that has the most votes
- Corrects for overfitting
- But little control over model

Models - RNN-LSTM



Models - Bi-LSTM



Experiments

Experiments - Notification

- **Extracted Sentences**
  - Only contains all whole sentences and mapped sentiment without data clean
- **Whole Dataset**
  - Contains whole dataset and mapped sentiment without data clean.

- **Test Size:** 0.2 (with Shuffle & downsampling)

- Evaluation: **Accuracy**, **Precision**, **F1**, **Recall** and **mlogloss** for Deep Learning.

Experiments - Majority Baseline

Model	Test Acc	F1	Recall	5-Fold CV
Baseline Majority	**42.09%**	**None**	**None**	**None**

Experiments - TextBlob Baseline

Model	Test Acc	F1	Recall	5-Fold CV
XGBoost + Textblob	**54.10%**	**0.54**	**0.49**	**52.3%**

Experiments - Naive Bayes

Model	Test Acc	F1	Recall
Naive Bayes + Unigram + Whole Dataset	**68.33%**	**0.69**	**0.69**
Naive Bayes + Bigram + Whole Dataset	67.25%	0.67	0.67
Naive Bayes + Trigram + Whole Dataset	66.63%	0.66	0.67
Naive Bayes + [1-3]gram + Whole Dataset	66.75%	0.67	0.67
Naive Bayes + [1-7]gram + Whole Dataset	68.08%	0.68	0.68

Experiments - Naive Bayes (Cont)

Model	Test Acc	F1	Recall
Naive Bayes + Unigram + Remove Stop Words + Whole Dataset	**63.18%**	**0.60**	**0.63**
Naive Bayes + Unigram + Remove Stop Words + Extracted Sentence	62.84%	0.57	0.63

Experiments - Support Vector Machine

Model	Test Acc	F1	Recall	5-Fold CV
SVM + Unigram + Whole Dataset	73.4%	0.73	0.73	None
SVM + Bigram + Whole Dataset	71.74%	0.71	0.72	None
SVM + Trigram + Whole Dataset	69.62%	0.69	0.70	None
SVM + [1-3]gram + Whole Dataset	**73.86%**	**0.74**	**0.74**	**72.1%**

Experiments - Support Vector Machine

Model	Test Acc	F1	Recall	5-Fold CV
SVM + Unigram + Extracted Sentence	56.21%	0.56	0.55	None
SVM + [1-3]gram + Extracted Sentence	59.67%	0.58	0.60	None

Experiments - Random Forest

Model	Test Acc	F1	Recall	OOB Score
Random Forest + TFIDF + ChiSquare Feature Selection[1] + Whole Dataset	65.36%	0.65	0.65	0.66
Random Forest + GI Lexicon + WordNet + Whole Dataset	**72.49%**	**0.72**	**0.73**	**0.72**

χ-square Feature Selection https://nlp.stanford.edu/IR-book/html/htmledition/feature-selectionchi2-feature-selection-1.html

Experiments - RNN-LSTM

Model	Test Acc	Train Acc	Val Acc	5-Fold CV
LSTM + Tokenizer + 3 layers + Whole Dataset	**70.67%**	**74%**	**70%**	**None**
LSTM + Tokenizer + 3 layers + Extracted Sentence	61.58%	69.37%	61.58%	None
Bi-LSTM + Tokenizer + 3 layers + Whole Dataset	65.29%	74.3%	65.29%	None

Experiments - Overall Result

Model	Test Acc	F1	Recall
Baseline Majority	42.09%	None	None
XGBoost	54.10%	0.54	0.49
Naive Bayes	68.33%	0.69	0.69
SVM	**73.86%**	**0.74**	0.74
Random Forest	72.49%	0.72	0.73
LSTM	70.67%	0.73	**0.75**

One More Thing...

Experiments - 5 Classes Test

Model	Kaggle Test Acc	10-Fold CV
Majority Baseline	0.411	None
SVM	0.622	None
LSTM	**0.643**	60.86%
Ensemble Random Forest[1]	0.658	61.32%
Current Best[2]	0.765	None

1 Kaggle Current 62nd Place Score: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/discussion/9856
2 Current Best Score: 0.76526: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/leaderboard

Conclusion

Conclusion

Sentiment analysis is still a challenging research topic.

“8526 The Santa Clause 2 is a barely adequate babysitter for older kids , but I 've got to give it thumbs down . 2”

SVM performs better in our experiments but RNN is the most widely used in contemporary studies.

Need better preprocessing methods for the data set (stritifying?).

Need to study more on binarized treebank, kernel for SVM.

Reference

[1] Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In ACL, pages 115–124.

[2] Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng and Chris Potts. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).

[3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.

[4] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004.

Q&A

Thank you!