Multiclass Sentiment Analysis and Aspect Segmentation on Movie Reviews

Zelin Li	Wenhao Xu	Shuochen Zhao
zl3611@nyu.edu	wx2056@nyu.edu	sz3315@nyu.edu

Introduction & Motivation

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service or idea.

Movie review is a great corpus for doing sentiment analysis since it contains various emotions from different people based on the movie they have watched. In the paper "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" [1]. The paper introduces the Recursive Neural Tensor Network (RNTN), a model for sentiment analysis that captures semantic compositionality over a sentiment treebank. Using a newly created dataset of sentiment values for movie review sentences, the RNTN outperforms several baselines, achieving state-of-the-art performance at the time. However, outside of this paper, there are still many applicable algorithms and feature engineering methods to be explored. So this project will continue to explore more feasible approaches to multiclass sentiment analysis on movie reviews.

What's more, the paper's[1] approach is limited in only getting the sentiment value of the movie reviews and fails to address specific aspects that influence a viewer's experience. Because people may take different aspects into consideration when they check the movie reviews such as the facilities of the theater, the greatness of the movie plot, the performance of actors and so on. By performing aspect segmentation on movie reviews after multiclass sentiment analysis, we aim to provide a more detailed understanding of the aspects that contribute to a movie's overall reception. The project will identify aspects such as theater facilities, movie plots, and actor performance (will decide this specifically later) and combine with the accurate output of sentiment value together for the sentences in movie reviews in paper [1]. This will not only help viewers make informed decisions about which films to watch but also enable industry professionals to identify strengths and weaknesses in their offerings.

Professor extended the project proposal for this project to Thursday Midnight without penalty.

To conclude, Multiclass Sentiment Analysis and Aspect Segmentation on Movie Reviews is a research project aimed at advancing the understanding of sentiment in movie reviews beyond the simple binary classification of positive or negative sentiment. By leveraging techniques from natural language processing and machine learning, this project will focus on identifying multiple sentiment classes and segmenting them according to various aspects of the movie-watching experience, such as theater facilities and movie plots (will decide this specifically later). This deeper understanding of sentiment in movie reviews might provide valuable insights for both consumers and industry professionals, as well as contribute to the advancement of sentiment analysis research. Furthermore, by extending the scope of sentiment analysis research, this project has the potential to impact various applications beyond movie reviews, such as product reviews and social media analysis.

Description of the Data

The dataset used in the paper "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" [1] is well-formulated, we will use it to perform both multi-class sentiment analysis and aspect segmentation, though different data preprocessing methods are expected.

The dataset has10,605 processed movie review sentences that are splitted into train, test and development data.Within it, 239,223 sentiment phrases and the corresponding sentiment labels, tree parse structure encoders and so on are contained. The data is already downloaded from the website and is ready for preprocessing.

Related Work

The sentiment analysis in the computational linguistics field is fruitful and previous related published papers are also easily accessible. The blueprint of our project is based on the paper "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" [1], in which the authors describe their use of recursive deep models for sentiment analysis of a large corpus of movie reviews. For feature extraction, retrospectively speaking, the previous studies have already concluded conventional and widely-used features. A study by Pang et al. (2002)

Professor extended the project proposal for this project to Thursday Midnight without penalty.

used n-grams as features in a supervised sentiment analysis model for movie reviews [2]. And Go et al. (2009) used POS tags as features in a supervised sentiment analysis model for tweets[3]. Also, sentiment lexicon is used by Hu and Liu (2004) as features in a supervised sentiment analysis model for product reviews [4]. Tang et al. (2014) used word embeddings as features in a supervised sentiment analysis model for product reviews[5]. Moreover, Kiritchenko et al. (2014) used sentiment-specific features as features in a supervised sentiment analysis model for tweets [6]. However, feature extraction is way far from its limitations and remains a huge potential to be explored and enhanced, we would continue research in our project. As for aspect-based sentiment analysis, He, Li and Huang's study inspires us as they focus on a supervised aspect-based sentiment analysis method that uses convolutional neural networks to predict the sentiment polarity of different aspects in product reviews [7]. Detailed implementation and further discussion will be unfolded in the methodology part.

Methodology

The multiclass sentiment analysis stage will contain four parts. Data preprocessing, feature generation & engineering, train models to classify sentiment levels, and finally test and evaluate model performance. Depending on model performance, adjustments will be made on existing features or new features, meanwhile, several machine learning models will be trained and tuned.

First of all, for data preprocessing, we first execute text normalization, which involves converting all text to a standard format, such as converting all text to lowercase, removing punctuation, and expanding contractions, stop word removal, stemming and so on. This step is followed by removing noisy or null data, with the help of pandas and visualization tools. All these techniques manipulate and truncate the raw sentences into a well-designed form that is efficient and friendly to further sentiment analysis. At last, the dataset will be split to train/dev/test or use cross validation to prepare for sentence feature generation.

In order to make data usable for machine learning models, we will engineer on sentence features. From our initial point of thinking, we would include the statistical and linguistic features. Statistical features may contain sentimental word's frequencies, TF-IDF, and word embeddings. Linguistic features like N(1,2,3) grams, Bag-of-Words, etc. will be included.

Professor extended the project proposal for this project to Thursday Midnight without penalty.

Beyond that, through our project, more experimental features like phrase structure trees will be tried and introduced, some feature generation models like Word2Vec will also be applied to improve data vector (word embedding) quality.

With regard to the label preprocessing, since the Stanford Dataset contains no direct labels to the sentences, rather it annotates phrases in review sentences with id and sentiment score. One of our potential ways to handle it is tracking back which sentence(s) the phrase belongs to, and assigning the phrase's score to the sentence. The sentence's ultimate label would be the average of phrases' scores inside. Another solution would be coherently processing the features and model on the level of phrase. Both ways are viable and need to be compared in the further project application. It's feasible for us to compartmentalize scores into more detailed sentiment categories for multi-class sentiment analysis, because labels are numeric scores.

In terms of the baseline system, we plan to use simple linear classifiers, like Support Vector Machine, which is capable of making multi-class classifications in approaches such as pairs comparison or one-vs-all. Hopefully, it will generate results of our pursuing labels like positive, extreme positive, neutral etc. This model serves the fundamental purpose of sentiment classification, therefore a relatively poor accuracy is expected. So in order to upgrade our system, where higher accuracy is expected to occur, we would like to implement models that go beyond SVM's capability.

After implementing the baseline system, we would then experiment with and tune more complicated classification models, like Random Forest, RNN, Naive Bayes and so on. We plan to use GridSearchCV to tune model parameters to reach better performance. Together with the feature improvement, we would expect that our well-trained model can make sentiment classification with greater accuracy compared to our baseline system.

In terms of metrics, we would use F1 score to measure the performance (accuracy) of our model and also serve as a reflection to the quality of our word embedding features. Finally, we would expect an increase in score on sentiment prediction.

To continue toward aspect segmentation, it links our predicted sentiment to detailed aspects of movie review. Relevant data sets / corpus will be generated for segmentation usage. We will select 3-4 movie or theater related major aspects, like "plot", "actor/actress", "theater facility", and some uncovered small aspects will be categorized toward "others". Corresponding corpus about these aspects (including "others", general film-related random aspects) will be

Professor extended the project proposal for this project to Thursday Midnight without penalty.

generated via ChatGPT or from reliable online resources, which are of high quality and relation. A ChatGPT API will be requested and used for the convenience of use. The specific aspects will be decided based on the ChatGPT feedback and the training data. See appendix for possible query and output for ChatGPT. Similar operations will be performed to word, phrases and sentences. This will help the TF-IDF process in the following paragraph.

Algorithms for aspect segmentation is to make TF-IDF for movie review sentences first. During this process, to ensure movie review sentence's TF-IDF vectors are usable, we will first filter out those sentences that only contain OOV words (not aspect related words), and segment them as "others" directly. Then for the remaining data, we will calculate cosine similarities with our target aspect corpus to perform segmentation. Additionally, The application of clustering may be experimented, in which we convert cluster sentiment and aspect into vectors. The cluster centroid is considered as sentiments and points can be aspects in which we meticulously observe the deviation distance and direction of points away from the centroid. Finally, we will use cluster-related metrics to evaluate our aspect labeling's effectiveness, using inter/intra-cluster distance. Further detailed implementation still requires consulting with the professor and our mentor.

Collaboration Plan

Generally, Zelin will be in charge of paper writing, data collection and analysis, proposal writing, brainstorming ideas and use ChatGPT to find the suitable aspects for segmentation. Wenhao and Shuochen will focus more on implementing the system baselines, explore sentiment analysis required feature engineering, coding part of implementation of the machine learning and NLP algorithms, output testing and evaluation.

Professor extended the project proposal for this project to Thursday Midnight without penalty.

Appendix

Q: For movie reviews, generate 10 separate words about the facilities in the movie theatre.

A: Cleanliness, Comfort, Seating, Accessibility, Sound, Projection, Ambiance, Concessions, Restrooms, Parking.

Q: For movie reviews, generate 10 separate Phrases about the facilities in the movie theatre.

A: Impeccably clean auditoriums

Spacious, comfortable seating

Crystal-clear surround sound

State-of-the-art digital projection

Spotless restroom facilities

Inviting, well-stocked concession stand

Efficient, friendly staff

Well-maintained temperature control

Convenient, ample parking

Cozy, relaxing lounge area.

References

[1] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

[2] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 79-86.

[3] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

[4] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

Professor extended the project proposal for this project to Thursday Midnight without penalty.

[5] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1555-1565.
[6] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. Proceedings of the seventh international workshop on semantic evaluation, 321-327.
[7] He, X., Li, Y., & Huang, M. (2018). A unified model for aspect-based sentiment analysis with gated convolutional networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2516-2525.

Professor extended the project proposal for this project to Thursday Midnight without penalty.