Sentiment Analysis on IMDB movie reviews using NLP

Chandramouli Yalamanchili
DSC680 - T302 Applied Data Science (2215-1)
https://chandu85.github.io/data-science

Which Domain?

I chose to work on NLP, Natural Language Processing analytics for my 3rd project, especially how it can be used to measure human emotions with respect to a particular topic. For this project I chose to work with IMDB movie reviews.

As I am looking to build a machine learning model to classify the text based on NLP, I am going to review the work that is done already using NLP as well as what are different things achieved using sentimental analysis.

Below are several references I am planning to refer as part this project:

1. Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
    - https://surface.syr.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1019&context=cnlp
    - I will review this reference to gain more understanding about NLP.
2. Vishal A. Kharde, S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques" International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.
    - https://arxiv.org/pdf/1601.06971.pdf
    - This reference also talks different techniques for performing the sentiment analysis on twitter data, I will review this to gain understanding on different techniques possible for sentiment analysis.
3. Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143-157.
    - Sentiment analysis: A combined approach
    - This reference talks combining different classification techniques into a combined method to achieve better results. I will review this to gain understanding on different techniques use and how they are combined to achieve better sentiment analysis outcome.
4. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
    - https://www.aclweb.org/anthology/W11-0705.pdf
    - I will review this article to understand how they have performed the sentiment analysis on twitter data and see if I can use any of the techniques in my project.
5. Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 375-384).
    - https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5917&rep=rep1&type=pdf

- This article talks about using different techniques to perform sentiment analysis, I am not familiar with any of the techniques mentioned, so I will review and try to gain understanding on these techniques.

6. Whitelaw, C., Garg, N., & Argamon, S. (2005, October). Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 625-631).
    - Using Appraisal Groups for Sentiment Analysis
    - This article discusses about classification of movie reviews using the appraising adjectives, it definitely looks interesting and closer to my project context. I will review this further to see if there are any components that I can use in my project.
7. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).
    - https://www.aclweb.org/anthology/P11-1015.pdf
    - This article discusses a different approach of combining supervised and un-supervised techniques for sentiment analysis. I will review this article to understand the techniques proposed.
8. Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.
    - https://www.sciencedirect.com/science/article/pii/S1018363916300071
    - This article discusses several challenges with sentiment analysis and evaluation process. I will review this article to understand different challenges related to different NLP approaches and techniques.
9. IMDB Dataset of 50K Movie Reviews - https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
10. Overview about the sentiment analysis
    - Overview about sentiment analysis - https://monkeylearn.com/sentiment-analysis/
    - Use cases for sentiment analysis - https://www.whoson.com/customer-service/top-ten-benefits-of-sentiment-analysis/
    - Referred these links to gain better understanding about sentiment analysis and possible use cases for the same.
11. What is NLP and why is it important? - https://www.analyticsinsight.net/what-is-nlp-and-why-is-it-important/
    - Referred this link to get some basic understanding on why NLP is important and how it can help with some of today's automation requirements.

Which Data?

I am planning to use the below dataset from Kaggle with movie review texts captured from IMDB.

Dataset Link - https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

- This dataset is the collection of 50,000 movie reviews captured from IMDB.
- This dataset has two columns, the text review about a movie, and the sentiment which is class for this project.
- Sentiment has two possible values either positive or negative, indicating a positive review or a negative one.

- By looking at the quick statistics from Kaggle, dataset seems to be very balanced dataset with 50% positive and 50% negative reviews.

Research Questions? Benefits? Why analyze these data?

- NLP, natural language processing is one of the fastest growing areas of AI in the recent history, this technology is coming much closer to our life's day by day. We are seeing more and more call centers and IVR systems started using NLP to interpret the natural responses of the customers instead of having to use the dial pads.
- Sentimental analysis is a use case for NLP, where we will use machine learning algorithms to analyze the natural language in terms of words and classify the text as either positive or negative sentiment.
- Sentiment analysis is playing a key role in automating several areas in enterprises starting from product responses, agent monitoring, sentiment aware chat bots, employee satisfaction, etc.
- Considering the future possibilities, I see in both NLP and sentiment analysis, I want to spend time through this project to gain more understanding on how to analyze the unstructured data in the form of raw text and come up with insights with respect to sentiment by applying NLP machine learning techniques.

Method?

- I am planning to use Python for this project as well.
- I will initially do some data visualization to understand the movie review data.
- I will then build a machine learning model to evaluate the review text and come up with classification of the movie review by applying sentiment analysis.

Potential Issues?

- At this time, I am not anticipating any major roadblocks or issues with respect to data.
- Only challenge for me is that I have not spent a lot of time on text analytics before, so it would be interesting to learn about different models I can use and different types of data analytics I can perform on the text data.

Concluding Remarks

Natural Language Processing (NLP) helps the machines to understand the human language, it is a component of artificial intelligence. NLP helps machines to understand the text data that can come from many sources and is unstructured in nature. Great progress has been made in last few decades in NLP domain mainly because of increased computing capabilities, we see NLP being implemented in lot of applications around us through several smart home devices and chat bots.

Sentiment analysis is an NLP technique that is used to determine the emotion attached to the textual data that generally represents the comments from humans. Sentiment analysis is playing a key role in several enterprises already by enabling the automation in the area of analyzing the customer reviews and feedbacks there by attaining the customer feedback much quicker and being able to make decisions for future products faster while meeting the customer demands.

Through this project I will explore different NLP models and text classification models to analyze the IMDB movie reviews dataset and come up with the sentiment for each review comment to indicate whether it was a positive comment or a negative one.