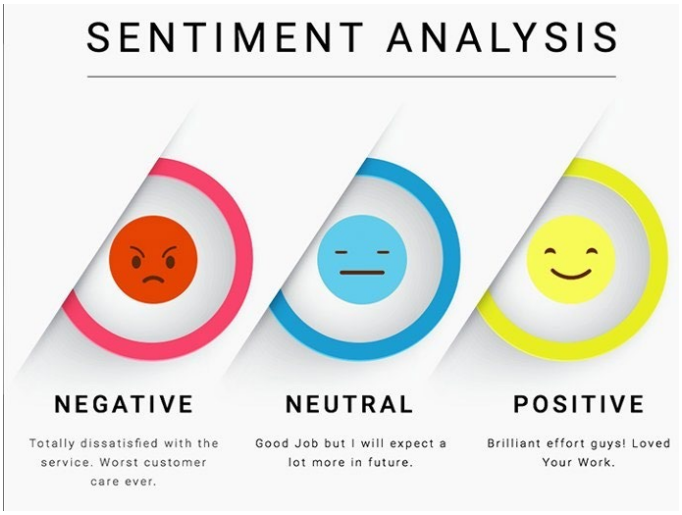
IE6483 Artificial Intelligence and Data Mining

Mini Project (Option 1)

Due: Friday (11:59pm), 15 Nov, 2024



Sentiments of Product Reviews

[Source: https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html]

Sentiment analysis (a.k.a opinion mining) is the use of natural language processing (NLP) to determine the polarity (i.e., positive, negative and neutral) of a given piece of text. You are required to build a learning-based classifier to classify the sentiments of product reviews. The whole process involves the following steps:

1. **Reading in data:** In this project, you will be using product review data [1]. You will need to load the data from the provided *json* files. There are two csv files, namely training set *("train.json")* and testing set *("test.json")*, each contains examples in the dataset splits. The attributes of the data are as follows:
    a. *reviews* – the user review of the product (in raw text)
    b. *sentiments* – the ground truth sentiment of the review, where *0* represents *negative* and *1* represents *positive*.
2. **Data processing:** You will need to convert the raw data into appropriate feature format such as via bag-of-words models (e.g., TF-IDF), pretrained word embeddings model [2,3] or pretrained language models [4]. The procedures in this step may vary according to the feature format selected. You may refer to the *"tfidf_features_example.ipynb"* for an example on how to convert the review from raw text to TF-IDF features.
3. **Model Selection:** You are to design and develop a classification model to classify the sentiments of product reviews. You are encouraged to use neural networks for this task. For example, you can treat the review as a sequence and embed them using recurrent neural network (RNN) or pretrained language model such as BERT [4], then pass the embeddings to a classifier layer.

4. **Training:** You will need to try different model parameters to obtain good classification results, e.g., feature dimension, weights, initialization, and learning rate. In the training stage, your algorithm should take only the inputs from the training set and predict their corresponding labels.
5. **Prediction:** By setting up the correct learning algorithm, you can classify the sentiment of the reviews in the test set. You need to report the classification results of the testing data in the file ("submission.csv").

Submit your report to answer the following questions:

a) Literature Survey on the following aspects:
   (1) Study and review tutorials, books, or surveys related to the field to gain a comprehensive understanding of the entire domain. Please clarify problem definitions, state the challenges and common solution types under different settings, e.g., supervised vs. unsupervised, closed-set vs. open-set, with vs. without domain shift.
   (2) Use paper references, keyword searches, and conference paper lists to find papers that have received significant attention in recent years. Quickly read and summarize dozens of papers to understand the latest developments in the field.
   (3) Review the recent progress in this field: Based on your understanding, identify one or two top-performing groups in the field, select some papers of interest for in-depth reading, and write a summary of these selected works in detail.
   (4) Based on the survey and review, explain which approach or method is suitable as the baseline for tackling the targeted problem here. Explain and motivate what improvements need to be made to implement the proposed solution for the project.

   (20% marks)

b) State your choice of feature format. Describe the data preprocessing procedures to convert the raw data into appropriate feature format for input to the model. (5% marks)

c) Select at least one appropriate model (e.g., RNN + linear layer, etc.) to build your classifier. Clearly describe the model you use, including model architecture figure, the input and output dimensions, structure of the model, loss function(s), training strategy, etc. Include your code and instructions on how to run the code if you are solving the problem by programming. If non-deterministic method is used, ensure reproducibility of your results by averaging the result over multiple runs, cross-validation, fixing random seed, etc. (15% marks)

d) Discuss how you consider and determine the parameters (e.g., learning rate, etc.) / settings of your model as well as your reasons of doing so. (5% marks)

e) Apply the classifier(s) built to the test set ("test.csv"). Submit the "submission.csv" with the results you obtained. (15% marks)

f) Analyse some correctly and incorrectly classified (if any) samples in the test set. Select 1-2 cases to discuss the strength and weakness of the model. (10% marks)

g) Discuss how different choice of feature format may affect the project in terms of resource consumption and accuracy. (10% marks)

h) Assuming that in a separate project, you need to perform sentiment classification on hotel reviews. However, the hotel reviews consist only of reviews in raw text and does **NOT** come with **rating scores**. Discuss how your classification algorithm would perform in this new project and how can it be modified to perform well in this new problem. (10% marks)

i) For the same sentiment classification problem in (g), if the rating scores are given, but **inaccurate** (i.e., noisy annotation). Discuss how your classification algorithm would perform in this new project. Name three approaches that you can improve your algorithm to perform well in this new problem. (10% marks)

Notes:

- If you couldn't obtain any meaningful results or answers to the questions above, you may describe what you have done and attach the relevant working, codes, or screenshots, if available.

- Work in group of **THREE** students and submit one report which must clearly indicate (1) the group members and (2) the **respective contribution** of each group member to answering the questions.

- You should clearly cite all the references and sources of information used in your report.

- You are expected to uphold NTU Honour Code.

- Submit your report and the file "submission.csv" with your results to NTULearn by the **deadline**: Friday (11:59pm), 15 Nov, 2024.

References

1. John Blitzer, Mark Dredze, Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. Association of Computational Linguistics (ACL), 2007.
2. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
3. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.
4. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. 2019.
5. https://nlp.stanford.edu/projects/glove/