Siqi (David) Liu
20428295

CS 686, Winter 2020

Project Proposal

- The application domain is **sentiment analysis/text classification**
- The data I will be using is the Yelp dataset, stored in JSON format. The following information is available
  - 200K businesses, with location data and categories
  - 6.7 million reviews, with full text data and ratings (1 - 5)
- The input variable is the **full review text**. For simplicity purpose, the target variable, **rating**, will be transformed into a binary class:
  - 1 - 2: negative (class 0)
  - 4 - 5: positive (class 1)
  - 3: neutral, discard
- I will be implementing and comparing the effectiveness of various machine learning/AI algorithms. Since this is a binary classification problem, I am considering the following:
  - Naive Bayes [1] [2]
  - Logistic regression [3] [4]
  - Tree-based ensemble models (e.g., LightGBM, XGBoost) [5] [6]
  - Long-short term memory (LSTM) [7] [8]
  - Bidirectional Encoder Representations from Transformers (BERT) [9] [10]
- The analytical aspect of this project would be on why certain algorithms out-perform/under-perform compared to the others
- Extension, if time permits - I will try to identify any common theme among positive reviews that businesses can leverage to achieve higher ratings

References

[1] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI 1998*, 1998.

[2] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[3] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 753–758, 2015.

[4] A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," *Entropy*, vol. 17, p. 252, 2009.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, pp. 3146–3154, 2017.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] X. Wang, Y. Liu, C.-J. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1343–1353, 2015.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using bert," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, vol. 1, pp. 1–5, IEEE, 2019.