Nik Nandi

VIP: Machine Learning in the Financial Markets

Saturday, November 30, 2024

Analysis Summary

The analysis of the dataset reveals meaningful insights into the relationship between sentence characteristics and sentiment labels. First, a correlation between sentence length and sentiment was identified, where certain sentiments, such as hawkish or dovish, were associated with either shorter or longer sentences. This suggests sentence length is a valuable feature for predictive sentiment models. [Figures: a, c]

By transforming sentences into high-dimensional embeddings using a pre-trained language model and reducing dimensionality via PCA, clear clustering patterns emerged based on sentiment labels. This demonstrates the power of embeddings in capturing semantic and sentiment-related information. Additionally, mutual information analysis confirmed that sentence length is informative, helping reduce uncertainty about sentiment labels. [Figures: b & 1-3]

These findings highlight the combined value of simple features like sentence length and advanced embedding techniques in enhancing sentiment analysis models. Furthermore, visualizations of reduced embedding dimensions revealed patterns useful for feature engineering and thematic clustering. [Figures: d-h]

PS: A more refined analysis of the data can be done looking for more related features, but admittedly I do not have as much time to work on this as I wished I could have.

Works Used

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

https://www.youtube.com/watch?v=0MaJzw6z4AY&ab_channel=OrangeDataMining

https://www.youtube.com/watch?v=a5pxJgKFeP4&ab_channel=OrangeDataMining

https://en.wikipedia.org/wiki/Multidimensional_scaling

https://www.youtube.com/watch?v=_CzYVI8axao

https://ai.google.dev/gemini-api/tutorials/anomaly_detection

https://www.learnpytorch.io/