Uncovering Thematic Sentiments in Literary Quotes from Goodreads

A H M Rezaul Karim

CS700 – Fall 2023

Introduction

- Sentiment analysis - essential for understanding emotional undertones in texts.
- Identify themes in literary works, assess NLP models in literary contexts.
- Evaluate how NLP models handle nuanced human expressions.
- Important for enhancing model sophistication and accuracy.
- Contributes to the advancement of sentiment analysis tools and recommendation systems.

Techniques Used

Mean, Median, Quartile – Box Plots and Bar Charts

Sampling from data:

-Small random subset to setup experimental environment

-Random sampling to create Train-Test sets

For thematic sentiment analysis:

1. Used SVM - Lower computational complexity
2. Used BERT – Higher computational complexity

ANOVA test to check if models' performance are different

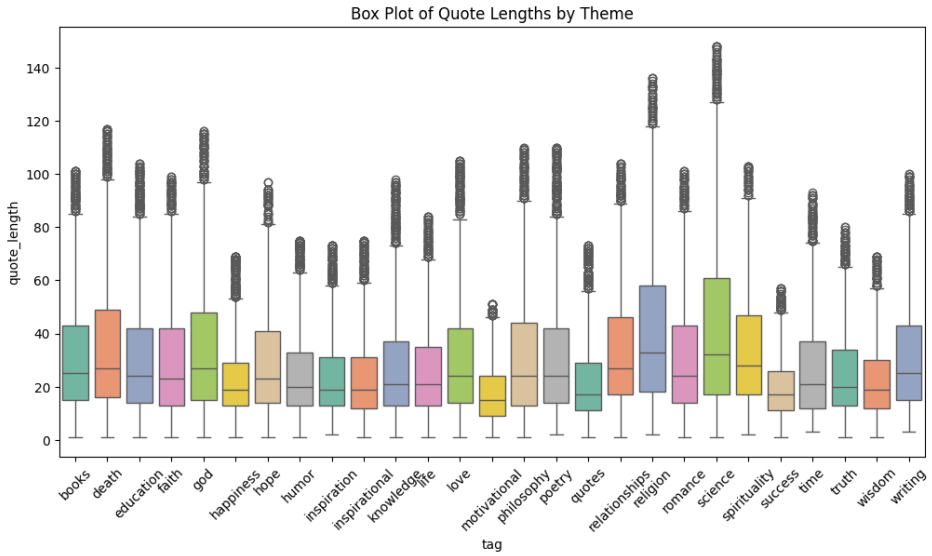
Approximate Visual Test and Paired Observation to find better performing model

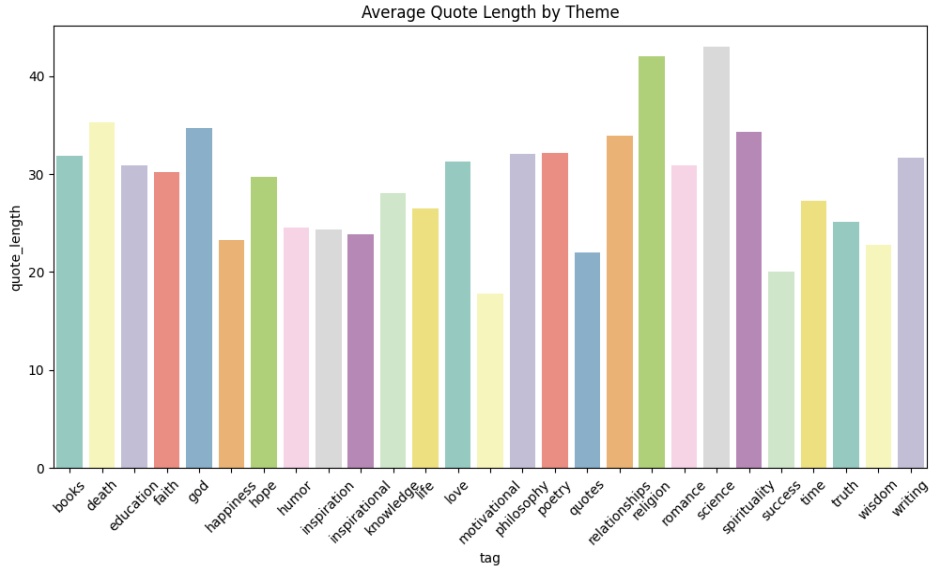
Research Methodology

- Goodreads Quotes dataset from *Kaggle* *(82460 quotes with 27 label)*
- Used box plot to identify and remove outliers, find average quote length to determine appropriate token size.
- Random sampling of 10% data used to create experimental setup.
- 10-fold cross-validation to record accuracies of SVM model performance.
- Fine-tuned BERT model for the task and recorded accuracies for different validation sets.
- Performed ANOVA test to determine if the performance of both models are significantly different.
- Used Approximate Visual Test and Paired Observation to determine which model performs better at specific confidence interval levels.

❖ https://www.kaggle.com/datasets/abdokamr/good-reads-quotes

Experimental Results





Accuracy results of BERT	Accuracy results of SVM
0.4606	0.35696
0.47376	0.36988
0.47149	0.37654
0.46664	0.36793
0.45159	0.37752
0.44353	0.36871
0.44173	0.36891
0.44016	0.38829
0.43484	0.3783
0.4386	0.37811

Test	Result	Decision
ANOVA	F-Statistic: 218.58138, P-Value: 1.64177e-11	Statistically significant difference
Approximate Visual Test	BERT- 95% CI(0.452155851, 0.452432349) SVM- 95% CI(0.373068689, 0.373161511)	BERT model is better
Paired Observation	95% CI(0.07892634, 0.07943166)	BERT model is better