SubjECTive-QA

A dataset for the qualitative evaluation of answers in Earnings Call Transcripts (ECTs)

Huzaifa Pardawala , Veer Kejriwal , Siddhant Sukhani , Abhishek Pillai
Rohan Bhasin , Dhruv Adha , Tarun Mandapati
Andrew DiBiasio , Agam Shah , Sudheer Chava
Georgia Institute of Technology

Abstract

SubjECTive-QA represents a pioneering effort to address a significant gap in existing datasets related to sentiment analysis and stock price forecasting. While previous financial datasets have focused on sentiment classification and numerical claim detection using Earnings Call Transcripts, none have delved into the intricate dimensions of answers provided during the Q&A sessions of these calls. In response to this limitation, the paper introduces SubjECTive-QA, the first annotated dataset specifically designed to capture the nuanced and diverse nature of responses during Earnings Call Q&A sessions.

1 Introduction

The primary objective of SubjECTive-QA is to facilitate tone classification, offering a more refined understanding of sentiment beyond simplistic positive or negative categorizations. The dataset not only contributes to sentiment analysis but also serves as a resource for downstream tasks such as numerical claim detection, hawkish-dovish-neutral sentiment classification, and, notably, forecasting potential headwinds or tailwinds for organizations.

2 Methodology

The methodology employed in creating SubjECTive-QA consisted of six steps: identifying Earnings Call Transcripts, scraping data, cleaning data, selecting features, manual annotation, and final dataset creation.

2.1 Dataframe Construction

110 Earnings Call Transcripts (ECTs) were randomly selected from various industry sectors, including Technology, Healthcare, Consumer Staples, Energy, and Industrial. These datasets were collected from Motley Fool (The Motley Fool, 2023) and SeekingAlpha (Seeking Alpha, 2023). The dataset spans from 2007 to 2021, aiming for diversity and comprehensiveness.

2.2 Data Cleaning

The raw ECTs were processed for clarity. Questions and answers were formatted and segregated to maintain their original context and relevance.

2.3 Manual Annotation

The team of annotators comprised nine male students from the Georgia Institute of Technology's College of Computing and School of Mathematics. It consisted of eight Undergraduate students and one Master's student. The team shortlisted 6 key features: Clear, Relevant, Specific, Optimistic, Cautious and Assertive. Following the selection of key features, a double blind annotation process was followed where the 110 Earnings Call Transcripts were randomly assigned to the annotators such that one transcript was assigned to three annotators. The annotation process included rating a particular answer in context to the question being asked across the six features. The annotators were supposed to rate an answer with a score of '2' if the answer showed positive correlation with that particular feature, '1' is there was no clear correlation, or '0' if there was a negative correlation. At the end, the individual annotations were combined based on majority rating. In case there was no clear majority that particular rating was assigned the value '1'.

3 Results

After creating the dataset, we used preliminary data analysis techniques such as creating probability graphs and a correlation matrix for the ratings of the tonality features across all Earnings Call Transcripts. The key findings were as follows:

3.1 Probability Graphs

The probability distribution of tonality features was evaluated at different levels (scores 0, 1, 2) across

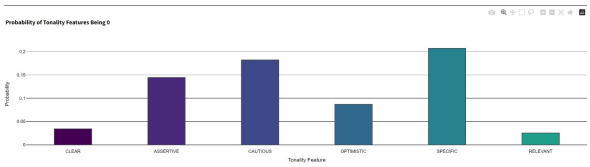

Figure 1: Probability Distribution of Tonality Features equal to 0

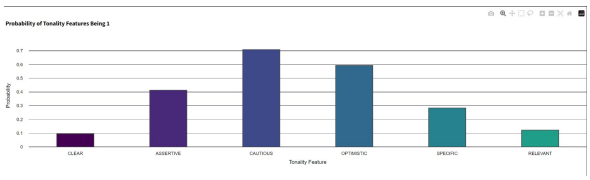

Figure 2: Probability Distribution of Tonality Features equal to 1

various properties. This analysis revealed that most respondents tended to answer questions neutrally. A significant finding was that 87% of answers were clear, and 85% were relevant, reflecting a high degree of cohesiveness and contextual relevance in the responses. Additionally, the data indicated a stronger inclination towards optimistic responses than pessimistic ones, with respondents being 2.9 times more likely to answer questions optimistically.

3.2 Correlation Matrix

Contrary to common expectations, cautiousness and optimism in responses were found to be independent, as indicated by a low correlation R-value of 0.01. This observation suggests that these tonal qualities operate independently in the context of ECTs. Furthermore, the relationship between clarity and specificity in responses was also explored. Despite intuitive expectations of a correlation, the matrix revealed a low R-value of 0.23, indicating no significant relationship between these properties.

The paper's outcomes include a comprehensive dataset meant for improved tone classification and insights into financial forecasting models.

4 Future Work

Future work will expand the dataset, and demonstrate its efficacy via an empirical evaluation involving benchmarking various State-of-the-Art Pre-Trained Language models: FinBERT (Araci, 2019), RoBERTa-base (Liu et al., 2019), FLANG-BERT (Shah et al., 2022) and Large Language Models in the context of tone classification. This evaluation

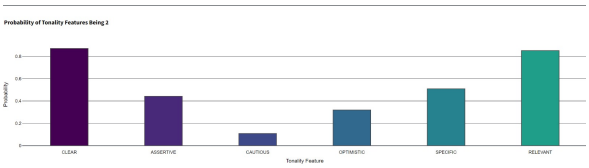

Figure 3: Probability Distribution of Tonality Features equal to 2

Figure 4: Correlation Matrix of Tonality Features

will not only showcase the dataset's utility but also shed light on potential limitations that may impact its application in research and analysis.

Ethics Statement

The research team commits to upholding rigorous ethical standards throughout the project. Data collection will strictly adhere to the terms of service, legal regulations, and ethical guidelines governing publicly accessible sources. To safeguard privacy and confidentiality, any personal or sensitive data encountered will be made anonymous and treated in accordance with applicable data protection laws. Transparency will be maintained in all research procedures, and efforts will be made to identify and mitigate biases. Responsible AI practices will guide the utilization of Pre-trained Language Models. The research team is dedicated to promoting accessibility, fairness, and transparency by communicating any limitations in the research findings to ensure ethical integrity.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach.

Seeking Alpha. 2023. Seeking alpha. Accessed on September 05, 2023.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain.

The Motley Fool. 2023. The motley fool. Accessed on September 05, 2023.