IT356 Project Report

Aakarsh Bansal
Artificial Intelligence
221AI001

Bhuvanesh Singla
Artificial Intelligence
221AI014

Deepak C Nayak
Artificial Intelligence
221AI016

Raajan Rajesh Wankhade
Artificial Intelligence
221AI031

***Abstract*—With advancement in the field of Artificial Intelligence and compute capabilities, applications of AI in domains like Finance Market have gained a lot of attention. One major area of interest is to analyze the current state of a market entity from its news headlines, reports, etc., and determine its sentiment. This analysis can be used for stock movement prediction, investment decision-making, and more. Generally, this is done through supervised learning, requiring a dataset to train a model on correct sentiment classification. While such datasets exist for Western finance markets, they are still lacking for India. This project aims to develop a method to create a similar dataset for the Indian finance market with minimal human intervention, leveraging AI agents for reliable data annotation. This paper presents a streamlined dataset creation pipeline from scratch, employing various NLP techniques for topic modeling, aspect extraction, data annotation, and dataset evaluation. A key feature of this project is the use of AI agents for robust data annotation without human input. All code has been open-sourced and is publicly available: github.com/typos12onlr/FinABSA.**

Keywords - Natural Language Processing, AI agents, Financial Dataset, Aspect Based Sentiment Analysis.

I. Introduction

This project aims to address the lack of structured financial sentiment data for Indian companies by developing an Aspect-Based Sentiment Analysis (ABSA) dataset specifically for the Indian market, modelled after the FiQA 2018 dataset [1]. The field of financial sentiment analysis for Indian companies currently suffers from a scarcity of well-structured datasets which limits the ability to assess market sentiment accurately and negatively impacts stock prediction models. While existing datasets like FiQA 2018 offer insights into Western markets, they lack the necessary specificity for analyzing Indian companies and their unique market dynamics.

To tackle this challenge, we utilize a dataset of Nifty 50 news articles containing attributes such as headlines, descriptions, full articles, and significant keywords. The output of this project is a refined dataset that provides, for each article, a text description and a dictionary containing sentiment analysis results in the form of (target): (aspect), (snippet), and (sentiment). This structure enables a detailed understanding of sentiment related to specific aspects of companies and their activities.

The motivation behind this project lies in the potential impact of such a dataset on research and analysis in the Indian financial sector. By offering a structured, sentiment-labeled dataset, it enables a deeper examination of how news sentiment influences stock prices, supporting researchers, analysts, and investors in making more informed decisions. This, in turn, contributes to the overall advancement of stock prediction models for the Indian market.

The primary contributions of this project include the creation of a novel ABSA dataset tailored for Indian companies, the development of a framework for automating data annotation, and a valuable resource for advancing research in financial sentiment analysis. By leveraging these contributions, this project aims to pave the way for more accurate and context-specific financial news analysis in the Indian market.

This paper is structured as follows: Section 1 discusses the introduction, motivation and the contribution of the paper, followed by Section 2 that explores the previous work in this domain, followed by Section 3 that details the methodology. Section 4 discusses the results and the paper concludes with Section 5 containing the conclusion and possible future works.

II. Literature Survey

In order to understand the domain better and analyse existing related works, we conducted an extensive literature survey.

Aspect-based Sentiment Analysis (ABSA) Techniques

Various studies explored ABSA using different methodologies. Rule-based systems by [2] utilized rule-based sentiment analysis combined with implicit aspect extraction. But the major drawback observed was that it relied on a specific dataset which might introduce bias. Another attempt was to use LLMs for downstream task like ABSA as done in [3], which provide flexibility and adaptability but fine-tuning resource-intensive models or relying heavily on external resources (such as lexicons and ontologies) can introduce biases and challenges.

Another study [4] emphasized fine-grained numerical understanding via specialized components like DigitCNN. This is a unique approach compared to other sentiment analysis models that do not explicitly handle numerical data. Limitations arise in the model's complexity and difficulty in adapting numeral embeddings dynamically to different contexts. Hence, while numerical data incorporation is promising, there is still room to improve context awareness.

Advances in Few-Shot and Zero-Shot Learning

SetFit [5] and Gliner [6] focus on minimizing labeled data requirements using contrastive learning and zero-shot methods, respectively. This contrasts with studies requiring domain-specific fine-tuning like InvestLM [7]. While these methods

reduce data dependency, their effectiveness is sensitive to factors like hyperparameter tuning or the quality of generated data pairs.

AI Generated Datasets

Multi-Hop RAG QA [8] is a dataset for multi-hop retrieval augemented generation question answering dataset, completely generated using large language models. However, the paper does not delve too deep into how the dataset was validated.

Another recent study [9] explored a multi-agent architecture comprising of different language models for different domain specific tasks, in creation of authentic knowledge work datasets. However, similar to [8] their evaluation process consisted of human evaluation and survey.

Major Gaps

1) **Dataset Limitation**: Most datasets used (like FiQA [1]) derive from Western financial data, limiting their generalizability to markets like India. Financial data in India lack a consistent format, making NLP tasks more challenging.
2) **Limited Exploration of LLMs for ABSA**: Despite the recent success of LLMs in various NLP tasks, their application to ABSA remains underexplored. Research tends to focus on traditional fine-tuning or basic prompting techniques, while more advanced methods like agentic workflows or complex prompting are seldom applied. This gap presents an opportunity to integrate advanced LLM-based techniques for inference, dataset generation and validation. We took inspiration from the annotation guidelines as used in [10].
3) **Extensive Human Expertise Requirement:** Creation of natural language datasets for tasks like sentiment analysis, especially for domain specific tasks generally requires a great amount of human expertise for dataset creation and evaluation. We propose a method that involves minimum human intervention for both generation of a dataset, as well as evaluation.

III. Methodology

To reduce the involvement and need of human experts to annotate domain specific datasets we have come up with the following methodology illustrated in Figure 1

We use NIFTY 50 News data as our data source. It contains news headlines, descriptions, article bodies, and other relevant details. Our focus is on using the headlines and descriptions to extract aspects and sentiments.

The entire methodology can be divided into several phases:

- Preprocessing Phase
- Creating Embeddings
- Clustering
- Aspect Extraction
- Aspect Disambiguation
- Aspect Extraction Evaluation
- Dataset Evaluation

A. Preprocessing Phase

News headlines contain named entities such as names of companies, persons etc. These have the possibility to create problems during the embedding phase by taking attention away from the aspects involved and dragging it towards the entities in question. We have therefore used a pretrained model called "Gliner" [6] for Named Entity Recognition. The named entities found were masked by their corresponding tag
List of entities:

- DATE: Absolute or relative dates or periods (e.g., July 4th, 2020, tomorrow).
- EVENT: Named events (e.g., World War II, Olympics).
- LOC: Locations.
- MONEY: Monetary values (e.g., $10, 500 million euros).
- ORDINAL: Indicates a ranking or order (e.g., first, second, third).
- ORG: Organizations (companies, agencies, institutions).
- PERSON: People, including fictional.
- PRODUCT: Objects, vehicles, foods, etc. (not services).

B. Creating Embeddings

We used the sentence transformer model paraphrase-mpnet-base-v2 [11] from huggingface. This model was used since headlines are paraphrases of the article description. This model takes the masked headlines generates word embeddings of 768 dimensions.

C. Clustering

After obtaining relevant embeddings we clustered the word embeddings using two algorithms: Agglomerative and Gaussian Mean Mixture. This idea was inspired by authors of RAPTOR [12]. We later optimised the number of clusters by plotting the graph for silhouette score vs number of clusters. We then chose the number of clusters above which the rate of increase of silhouette score was negligible. In our case, this number turned out to be 30.

D. Aspect extraction

We randomly sampled 10% of data points from each cluster. These data points were then fed to the gpt-4o-2024-08-06 model from to extract the aspects. We used one-shot prompting with structured outputs. We took inspiration from the FiQA dataset [1] and we prompted the LLM to extract 2 levels of aspects: level 1 and level 2 where level 1 is a broader aspect and level 2 is a more specific aspect under level 1.

E. Aspect Disambiguation

We came up with a hierarchical aspect filtering algorithm that processes a dictionary of clustered aspects to retain unique and relevant Level 2 aspects. It involves three phases: Global Cross-Category Filtering using cosine similarity to remove redundant Level 2 aspects, Within-Category Filtering to ensure distinctiveness within each category, and Final Organization to compile the filtered aspects into a dictionary. FinBert [13] embeddings are utilized for similarity calculations.

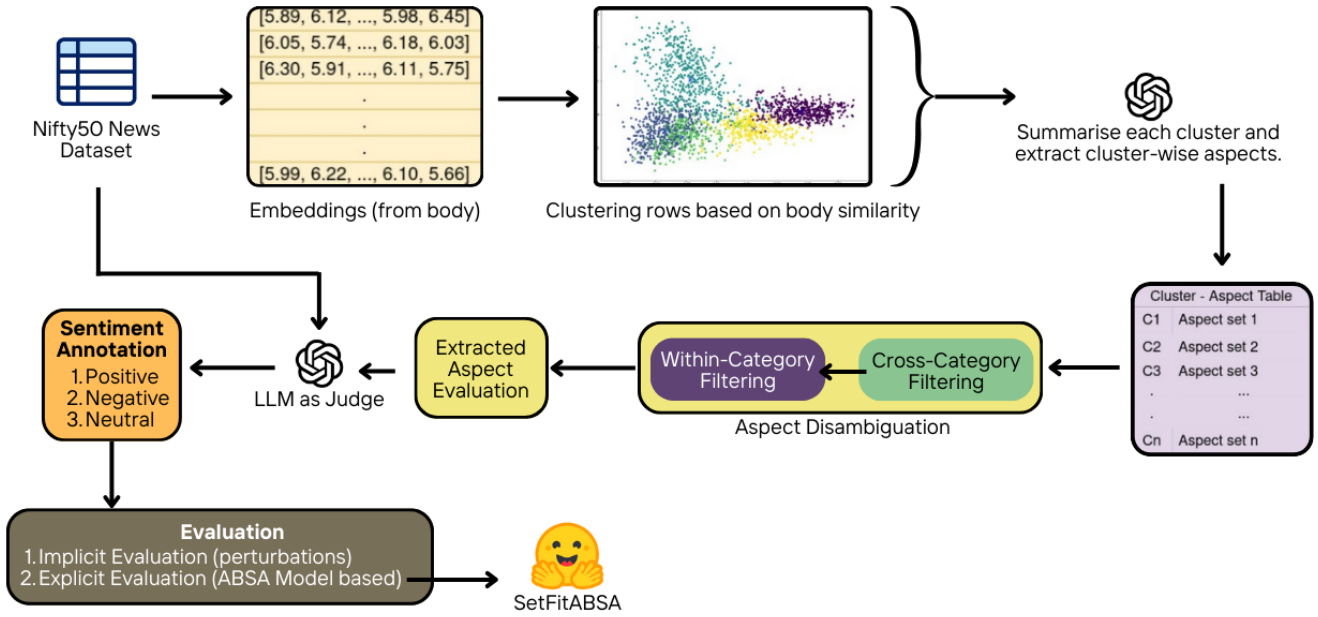

Fig. 1: Methodology

Algorithm 1 Hierarchical Aspect Filtering

Require:

$\mathcal{A}$: Set of hierarchical aspects (Level1/Level2)
T_1, T_2: Similarity thresholds
$E(\cdot)$: FinBERT embedding function

Ensure:

$\mathcal{F}$: Filtered aspect hierarchy

Phase 1: Cross-Category Filtering

$L_2 \leftarrow \text{GETUNIQUEL2}(\mathcal{A})$
$\mathbf{E} \leftarrow \{E(a) \mid a \in L_2\}$
$S_{ij} \leftarrow \text{sim}(\mathbf{E}_i, \mathbf{E}_j)\ \forall i, j$
for $(a_i, a_j) \in L_2 \times L_2$ where $S_{ij} > T_1$ **do**
    $p_i, p_j \leftarrow \text{parent}(a_i), \text{parent}(a_j)$
    Remove aspect with lower $\text{sim}(E(a_k), E(p_k))$ where $k \in \{i, j\}$

Phase 2: Within-Category Filtering

for $l_1 \in \text{UniqueL1}(\mathcal{A})$ **do**
    $L_2^{l_1} \leftarrow \{l_2 \mid (l_1, l_2) \in \mathcal{A}\}$
    $S_{ij}^{l_1} \leftarrow \text{sim}(E(l_2^i), E(l_2^j))\ \forall l_2^i, l_2^j \in L_2^{l_1}$
    for $(a_i, a_j) \in L_2^{l_1} \times L_2^{l_1}$ where $S_{ij}^{l_1} > T_2$ **do**
        Keep aspect with higher $\text{sim}(E(a_k), E(l_1))$ where $k \in \{i, j\}$
    $\mathcal{F}[l_1] \leftarrow$ remaining aspects in $L_2^{l_1}$
return $\mathcal{F}$

Complexity: Time $O(n^2 d)$, Space $O(n^2)$
where n = #aspects, d = embedding dimension

F. Aspect Extraction Evaluation

To evaluate the identified aspects, we sampled a subset of sentences from each cluster, creating a "description set" corresponding to the "aspect set" for that cluster. We then generated embeddings for both sets and calculated cosine similarity scores to measure their alignment.

For each cluster $c \in \{1, \ldots, K\}$ and aspect set A_j, we calculate:

1. Cluster Description Embeddings:

$$D_c = \{E(d_i) : d_i \in \text{Sample}(C_c)\}$$

where C_c is the set of descriptions in cluster c

2. Aspect Embeddings:

$$A_j = \{E(a_k) : a_k \in \text{AspectSet}_j\}$$

3. Mean Similarity Score:

$$S_{c,j} = \frac{1}{|D_c|} \sum_{d \in D_c} \left(\frac{1}{|A_j|} \sum_{a \in A_j} \cos(d, a) \right)$$

We then plot a heatmap that represents the mean similarity score between each pair of aspect set and description set. We use the all-mpnet-base-v2 [11] model from sentence-transformers to get a general dense embedding.

G. Dataset Evaluation

1) Implicit Evaluation: To evaluate the dataset quality, we introduced perturbations to our dataset. The perturbed

sentences change the syntax of the sentence while preserving the meaning. The idea is that aspects and sentiments extracted from the original sentence still apply after these perturbations. Out of the possible choices of synonym replacements, word order replacement, back translation for inducing perturbations, we chose to go ahead with back translation using the Google translate API.

This was done because all the other possible choices had the potential to accidentally disturb the target in question. We chose to translate the sentence to **Hindi, Korean and Chinese** in and then translate back to English in order to obtain 3 perturbed sentences from the original sentence. These languages were chosen because their structural and linguistic dissimilarity to English would result in stronger perturbations. At the end, three perturbed sentences were obtained for every original sentence.

2) Explicit Evaluation (ABSA Model based): Explicit evaluation techniques involve using external models to evaluate created datasets and their performance before and after using the created datasets.

For our task, we utilized a sentence transformer [11] model from Hugging Face, integrated with the SetFit_ABSA framework [14]. This framework facilitates few-shot training for domain-specific ABSA models. The "paraphrase-mpnet-base-v2" sentence transformer features only 110M parameters, significantly fewer than other state-of-the-art large language models.

We conducted our evaluation by training and testing the model on the FiQA Dataset, structured for sentiment analysis and available on Hugging Face (ronenlap/SetFitAbsa_FiQA). Using the FiQA trained weights, we evaluated our dataset.

IV. Results and Analysis

A. Clustering

We tested with Gaussian Mixture Model (GMM) Clustering and Agglomerative Clustering to determine the best clustering, as illustrated in figures 5 and 4, respectively. Both quantitative measurements and the intrinsic properties of our data served as a guide for choosing the best clustering strategy.

We decided to use the Silhouette score as our main indicator of cluster quality. The silhouette score $s(i)$ for a single data point i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where:

- $a(i)$ is the mean distance between point i and all other points in its cluster (intra-cluster distance)
- $b(i)$ is the mean distance between point i and all points in the nearest neighboring cluster (nearest-cluster distance)

The overall Silhouette score is the mean score across all data points. Scores range from -1 to 1. A score close to 1 mean the clusters are well-defined, a score closer to 0 suggest overlap between clusters and negative values indicate poor clustering.

To find the optimal number of clusters, we plotted Silhouette scores for up to 50 clusters, as shown in fig 2 and fig 3. The increase in Silhouette score plateaued after **30 clusters**, and therefore we used 30 as the final number of clusters. This gave us a silhouette score of 0.41 for both Agglomerative and GMM clustering.

In the end, we chose Gaussian Mixture Models over Agglomerative Clustering due to some key advantages in our context. Text embeddings often have subtle overlaps in their semantic space, and GMM's soft assignment approach models this better by allowing descriptions to partially belong to multiple clusters. This aligns with the reality of natural language where a single description might touch upon several related themes.

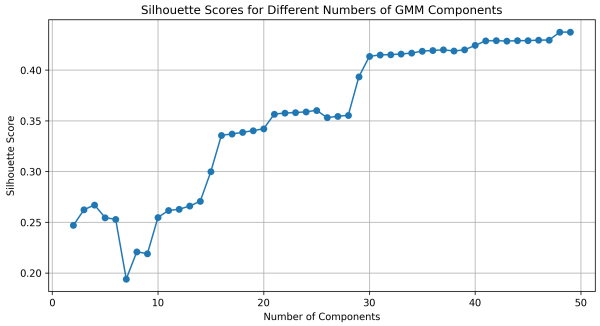

Fig. 2: Silhouette score optimization - GMM

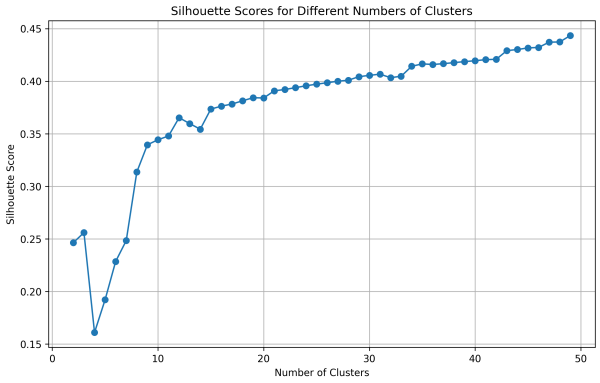

Fig. 3: Silhouette score optimization - Agglomerative

B. Aspects Extracted

We have extracted aspects from 10 of the extracted 30 clusters each representing a unique set of themes relevant to the financial news domain. We were limited to extracting only from 10 clusters due to financial and computational constraints. Below, we detail the key aspects identified in each cluster:

- **Cluster 0:**
  - Market/Stock Recommendation
  - Market/Stock Target Price
- **Cluster 1:**
  - Corporate/Revenue Base
  - Market/Demand

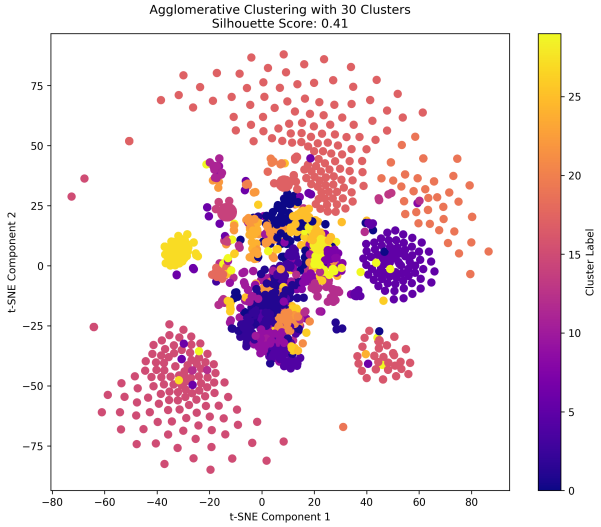

Fig. 4: Agglomerative Clustering

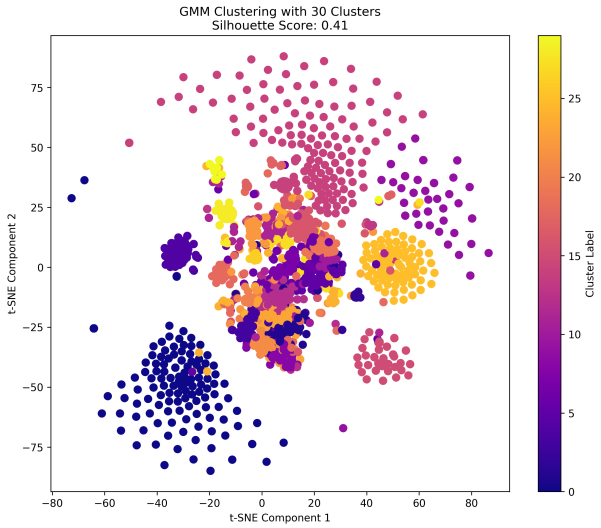

Fig. 5: GMM based clustering

- **Cluster 2:**
  - Market/Stock Rating
- **Cluster 3:**
  - Financial/Revenue Growth
  - Financial/Revenue Decline
  - Market/Analyst Estimates
- **Cluster 4:**
  - Market/Stock Recommendation
  - Market/Stock Target Price
- **Cluster 5:**
  - Financial/Revenue
  - Financial/Profitability
  - Corporate/Partnership
  - Corporate/Product Offering
- **Cluster 6:**
  - Financial/Pricing
  - Market/Market Share
  - Market/Demand Forecast
  - Corporate/Product Recall
- **Cluster 7:**
  - Corporate/Operational Health
- **Cluster 8:**
  - Financial/Revenue Growth
  - Market/Operational Efficiency
- **Cluster 9:**
  - Market/Stock Rating

The clustering process had initially identified themes within the financial data, with each cluster encompassing a range of aspects that are critical for analyzing company performance, market dynamics, and regulatory impact. We later used our aspect disambiguation process to filter out the aspects and then re-assigned them to the clusters.

C. Aspect Extraction Evaluation Result

To evaluate the obtained aspects, we sampled 10% of sentences from each cluster to obtain a "description set" corresponding to every "aspect set" for each cluster. We then obtained embeddings for the "description sets" and "aspect sets" to obtain cosine-similarity scores as mentioned in the methodology for aspect extraction evaluation. The below heatmap represents the aspect-cluster similarity as computed and averaged over their dense embedding. We observe that the trace of the heatmap matrix is made of cells that often have a higher value. This indicates that the aspect sets are more or less similar to their corresponding clusters. However, we also observe that some clusters have a high mean similarity score with aspect sets other than their own. We have addressed this in more detail in the future work section.

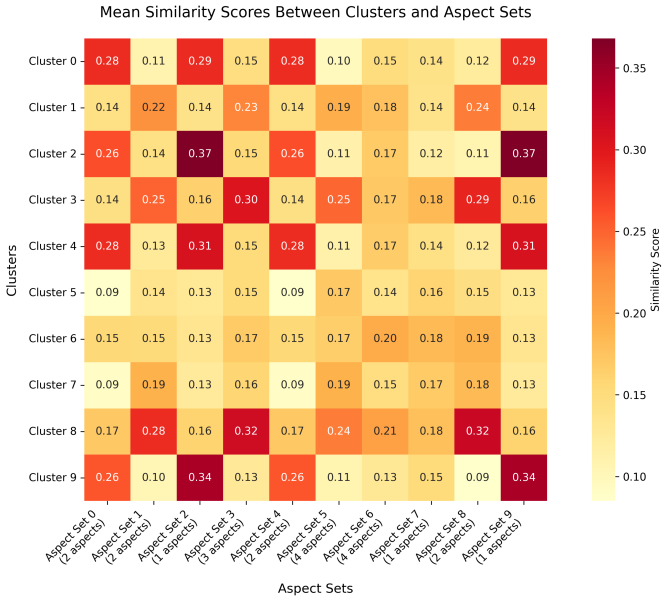

Fig. 6: Cluster-Aspects similarity matrix

D. Dataset Evaluation

1) Implicit Evaluation: In order to check the quality of dataset, we perturbed the original dataset as explained in the *Dataset Evaluation* section. To quantify our results, we compared the predictions of the LLM on perturbed sentences with the labels it had assigned to the original sentences and obtained results as shown in Table **??**. Accuracy of **83%** was obtained between the LLM's sentiment assignments across perturbed and original sentences.

Sentiment	Precision	Recall	F1-score	Support
Negative	0.59	0.80	0.68	20
Neutral	0.38	1.00	0.55	20
Positive	1.00	0.82	0.90	220
Macro Avg.	0.66	0.87	0.71	260
Weighted Avg.	0.92	0.83	0.86	260

TABLE I: Classification Report on Perturbed Sentences

2) Explicit Evaluation: As mentioned in our methodology, we evaluated our dataset with a model trained on FiQA dataset, and obtained results as shown in table II. This shows that the quality of our dataset is on par with FiQA dataset.

Dataset	F1 Score	Accuracy
FiQA SA	0.772	76.2%
Our Dataset	0.764	73.6%

TABLE II: SetFit scores on our dataset

V. Conclusion

This paper demonstrates the use of different NLP techniques to automate a task that requires a high degree of human involvement and supervision - dataset creation. By utilizing various facets of Natural Language Processing, we successfully compile a Financial ABSA dataset that is on par with an human-expert annotated dataset (FiQA). We demonstrate how the breakthroughs in language modelling, especially with the advent of LLMs, we can simplify and automate complex tasks requiring high domain expertise. We extracted the aspects using contextual embeddings and clustering, identified target entities and assigned them aspects, and annotated the sentiment polarity for the target entities with respect to the assigned aspects using SoTA LLMs.

However, we realize there is scope for improvement in all the aspects we have explored and documented here. We will discuss about the possible future directions in the next section.

VI. Future Work

- **Topic Modelling as an Additional Step**: The use of pretrained models such as BERTopic [15] introduces an effective mechanism for embedding the text and re-clustering, followed by the extraction of dominant and recurring keywords. One of the key advantages of integrating topic modelling is that it allows for an indirect evaluation of clusters. Specifically, by using coherence scores—metrics that assess how semantically similar the top words within each topic are—we can gauge the quality of each cluster [16]. A high average coherence score for a cluster indicates strong semantic consistency, suggesting that the clustering has successfully grouped related data points.
- **Language Agents for better annotation**: Data annotation conventionally is not a one person job. Often times the annotations done by an annotator are validated by other annotators, and the agreement is taken in account. Most annotation teams consist of an expert annotator and a few junior annotators. This human-like structure could be simulated using a multi-agent data annotation pipeline with different language models wearing different hats [17].
- **Evaluation Strategies**: While generating a dataset using generative AI is extremely useful, not many evaluation strategies exist to evaluate the effectiveness of these datasets. One could explore the potential ways to measure the goodness of an AI generated dataset. Most AI generated datsets are still evaluated by humans even today [9].

VII. Conclusion

This paper demonstrates the use of SoTA LLMs and for data annotation when used with robust agentic workflows. We also explore a streamlined pipeline for dataset generation from extraction to evaluation without any human intervention. We further plan to improve the pipeline using robust evaluation strategies mentioned in futher development.

References

[1] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=wCu6T5xFjeJ

[2] O. Alqaryouti, N. Siyam, A. Abdel Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data," *Applied Computing and Informatics*, vol. 20, no. 1/2, pp. 142–161, 2024.

[3] P. F. Simmering and P. Huoviala, "Large language models for aspect-based sentiment analysis," *arXiv preprint arXiv:2310.18025*, 2023.

[4] C. Qin, C. Yu, Y. Meng, and J. Chang, "A numeral and affective knowledge enhanced network for aspect-based financial sentiment analysis," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2023, pp. 926–933.

[5] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," 2022. [Online]. Available: https://arxiv.org/abs/2209.11055

[6] I. Stepanov and M. Shtopko, "Gliner multi-task: Generalist lightweight model for various information extraction tasks," 2024. [Online]. Available: https://arxiv.org/abs/2406.12925

[7] Y. Yang, Y. Tang, and K. Y. Tam, "Investlm: A large language model for investment using financial domain instruction tuning," *arXiv preprint arXiv:2309.13064*, 2023.

[8] Y. Tang and Y. Yang, "Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries," 2024. [Online]. Available: https://arxiv.org/abs/2401.15391

[9] D. Heim, C. Jilek, A. Ulges, and A. Dengel, "Using large language models to generate authentic multi-agent knowledge work datasets," 2024. [Online]. Available: https://dl.gi.de/handle/20.500.12116/45090

[10] G. Kontonatsios, J. Clive, G. Harrison, T. Metcalfe, P. Sliwiak, H. Tahir, and A. Ghose, "Fabsa: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing*, vol. 562, p. 126867, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223009906

[11] Sentence Transformers, "paraphrase-mpnet-base-v2 (revision e6981e5)," 2024. [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[12] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "Raptor: Recursive abstractive processing for tree-organized retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2401.18059

[13] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," 2019. [Online]. Available: https://arxiv.org/abs/1908.10063

[14] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," 2022. [Online]. Available: https://arxiv.org/abs/2209.11055

[15] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022. [Online]. Available: https://arxiv.org/abs/2203.05794

[16] H. Fei, T.-S. Chua, C. Li, D. Ji, M. Zhang, and Y. Ren, "On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–32, 2022.

[17] M. Lin, Z. Chen, Y. Liu, X. Zhao, Z. Wu, J. Wang, X. Zhang, S. Wang, and H. Chen, "Decoding time series with llms: A multi-agent framework for cross-domain annotation," 2024. [Online]. Available: https://arxiv.org/abs/2410.17462

APPENDIX A: ASPECT EXTRACTION PROMPTS

System Prompt for Aspect Extraction

The following system prompt was used to initialize the model for aspect extraction:

```
You are an experienced financial analyst with several years of experience. You are hired to
    annotate a financial ABSA dataset.
Your job is to look at the given sentence, and extract two levels of aspects from it.

Instructions for annotation:
- Analyze the sentence and identify the snippets. It is possible that one sentence has more
    than 1 snippet, but not always necessary.
- A snippet should be a coherent phrase containing specific business or financial information,
    either relating to an entity in the sentence or an event.
- For each snippet you identify, extract level 1 and level 2 aspects
- Level 1 aspects should be broad financial/business categories that could apply to multiple
    companies/events (e.g., Corporate, Financial, Market)
- Level 2 aspects should be specific subcategories of the Level 1 aspect (e.g., Revenue Growth,
     Profitability, Partnership)
- Both level 1 and level 2 aspects should be concise (1-2 words maximum) and relate to business
    /financial concepts
- Provide clear, concise reasoning that explains both the financial relevance and
    categorization logic
```

Example Outputs

The system was trained with the following examples:

```
Example 1:
Input: "Reliance's Q4 revenue grew 15% while operating margins declined due to higher raw
    material costs"

Output:
{
    "sentence": "Reliance's Q4 revenue grew 15% while operating margins declined due to higher
    raw material costs",
    "info": [
        {
            "snippet": "Q4 revenue grew 15%",
            "level_1_aspect": "Corporate",
            "level_2_aspect": "Revenue Growth",
            "reasoning": "this snippet talks about revenue of the corporate entity (Reliance)
    growing"
        },
        {
            "snippet": "operating margins declined due to higher raw material costs",
            "level_1_aspect": "Corporate",
            "level_2_aspect": "Profitability",
            "reasoning": "This snippet describes how increased input costs are negatively
    impacting the corporate entity's operating margins, directly affecting company-wide
    profitability"
        }
    ]
}

Example 2:
Input: "Joining hands with partners including HUL, SBI and OYO, Apollo Hospitals plans to scale
     this capacity to up to 5,000 rooms where patients can stay in isolation."

Output:
{
    "sentence": "Joining hands with partners including HUL, SBI and OYO, Apollo Hospitals plans
     to scale this capacity to up to 5,000 rooms where patients can stay in isolation.",
    "info": [
        {
            "snippet": "Joining hands with partners including HUL, SBI and OYO",
            "level_1_aspect": "Corporate",
```

```
            "level_2_aspect": "Partnership",
            "reasoning": "This snippet describes corporate-level strategic partnerships for
    business expansion"
        }
    ]
}
```

User Prompt Template

For each analysis, the following template was used:

```
Given the sentence: {sentence}

Previously extracted aspects (in the format level1/level2):
{aspects}

1. Extract the snippets that contain distinct financial/business information from the sentence
    (1 or more)
2. For each snippet:
    - FIRST check if any existing Level 1 and Level 2 aspect combinations from above lists
    accurately describe the financial concept
    - Only create new aspects if none of the existing ones represent the financial information
    - New aspects must follow business/financial terminology conventions
    - Use exactly the same aspect names when matching existing aspects
    - Ensure Level 2 aspects logically relate to their Level 1 parent category
    - Avoid creating new aspects that are synonymous with existing ones
```

APPENDIX B: ANNOTATION PROMPTS

System Prompt

The following system prompt was used to initialize the GPT model for financial statement analysis:

```
You are an experienced financial analyst with expertise in Aspect-Based Sentiment Analysis (
    ABSA) for financial text. Your task is to analyze financial statements and identify target
    entities, aspects, and sentiments.

ANNOTATION RULES:
1. Entity Identification:
   - Identify ALL target entities in the sentence
   - Valid entities: companies, products and FINANCIAL entities
   - Person names or generic named entities are NOT valid target entities
   - A sentence may contain multiple target entities

2. Snippet Extraction:
   - Extract relevant phrase(s) containing business/financial information
   - Snippets must be direct quotes from the sentence
   - Snippets should be short and concise, usually surrounding the target entity
   - Each snippet should focus on one specific piece of information
   - Ensure there is no redundancy between snippets

3. Aspect Assignment:
   - Format: "Level1/Level2"
   - Each target entity MUST be assigned ONE aspect
   - Choose ONLY from the provided aspect list
   - No custom aspects allowed
   - The same aspect cannot be assigned to a target entity more than once

4. Sentiment Analysis:
   - Analyze sentiment for each aspect-entity pair
   - Consider context within the specific snippet
   - Base sentiment on factual financial implications

5. Reasoning:
   - Provide clear, concise justification for each aspect-sentiment pair
   - Focus on financial/business impact
   - Reference specific parts of the snippet
```

Example Format

The system was provided with the following example format:

```
Input:
Sentence: "AstraZeneca's MedImmune Inks Licensing Deal With Omnis Pharmaceuticals"
Aspects: ["Corporate/Sales", "Product/Service", "Financial/Loss"]

Output:
{
    "sentence": "AstraZeneca's MedImmune Inks Licensing Deal With Omnis Pharmaceuticals",
    "annotations": [
        {
            "target": "MedImmune",
            "snippet": "MedImmune Inks Licensing Deal",
            "aspect": "Corporate/Sales",
            "sentiment": "positive",
            "reasoning": "Securing a licensing deal indicates business expansion and new
    revenue potential"
        },
        {
            "target": "Omnis Pharmaceuticals",
            "snippet": "Licensing Deal With Omnis Pharmaceuticals",
            "aspect": "Corporate/Sales",
            "sentiment": "positive",
            "reasoning": "Partnership through licensing deal suggests business growth
    opportunity"
        }
    ]
}
```

User Prompt Template

For each analysis, the following template was used to structure the input:

```
Please analyze the following financial statement:

Sentence: {sentence}

Available aspects (select only from this list):
{aspects}

Provide a complete analysis following the annotation guidelines. Include all target entities,
    their relevant snippets, aspects, sentiments, and reasoning.
```

APPENDIX C: PERTURBATION EVALUATION PROMPTS

System Prompt for Sentiment Analysis

The following prompt was used for sentiment evaluation of perturbed sentences:

```
You are an experienced financial analyst who can understand complex financial statements. You
    are tasked with analyzing the sentiment of a sentence with respect to a target entity in
    the context of the given aspect. Please use the given information and predict the sentiment
    . Also provide a reasoning for your prediction.

Details:
- A target entity is a company or a product mentioned or a financial entity mentioned in the
    sentence.
- An aspect is a category or a topic that the target entity is associated with. It is in the
    format "level1/level2". For example, "Market/Stock Recommendation".

Example:
Input-
sentence: Prabhudas Lilader is a boom on Mahindra and Mahindra in its research report on
    November 10, 2020 recommended a purchase rating on stock with a target price of Rs 728.
Target: Mahindra and Mahindra
Aspect: Market/Stock Recommendation
```

```
Output-
{
    "target": "Mahindra and Mahindra",
    "aspect": "Market/Stock Recommendation",
    "sentiment": "positive",
    "reasoning": "The sentence contains a positive sentiment as the research report recommended
     a purchase rating on the stock."
}
```

User Prompt Template

For each analysis, the following template was used:

```
Given the sentence: "{sentence}"
With the Target entity: {target}
And the Aspect: {aspect}

Predict the sentiment and provide a reasoning for your prediction.
```

APPENDIX D: EXAMPLE ASPECT ASSIGNMENTS

The following examples demonstrate how aspects were assigned to different financial statements:

```
Example 1:
Input: "Prabhudas Lilladher is bullish on Bharti Airtel has recommended buy rating on the stock
     with a target price of Rs 1032."

Output:
{
  "info": [
    {
      "snippet": "Prabhudas Lilladher is bullish on Bharti Airtel has recommended buy rating on
     the stock",
      "level_1_aspect": "Market",
      "level_2_aspect": "Stock Recommendation",
      "reasoning": "This snippet describes a market-level recommendation by Prabhudas Lilladher
    , indicating a positive outlook on Bharti Airtel's stock, which aligns with the existing
    aspect of 'Stock Recommendation'."
    },
    {
      "snippet": "target price of Rs 1032",
      "level_1_aspect": "Market",
      "level_2_aspect": "Stock Target Price",
      "reasoning": "This snippet provides a specific target price for Bharti Airtel's stock,
    which is a market-level financial forecast, matching the existing aspect of 'Stock Target
    Price'."
    }
  ]
}

Example 2:
Input: "Prabhudas Lilladher is bullish on Asian Paints has recommended buy rating on the stock
    with a target price of Rs 3675."

Output:
{
  "info": [
    {
      "snippet": "bullish on Asian Paints has recommended buy rating on the stock",
      "level_1_aspect": "Market",
      "level_2_aspect": "Stock Recommendation",
      "reasoning": "This snippet indicates a positive outlook and a buy recommendation for
    Asian Paints, which aligns with the existing aspect of stock recommendation."
    },
    {
      "snippet": "target price of Rs 3675",
```

```
      "level_1_aspect": "Market",
      "level_2_aspect": "Stock Target Price",
      "reasoning": "This snippet provides a specific target price for the stock, which matches
    the existing aspect of stock target price."
    }
  ]
}
```

Note: These examples demonstrate consistent aspect assignment across similar financial statements, showing how the system maintains uniformity in categorizing market recommendations and target prices.

APPENDIX E: EXAMPLE ANNOTATIONS

The following examples demonstrate the annotation outputs from gpt-4o-2024-08-06:

```
Example 1:
Input: "Prabhudas Lilladher is bullish on Tata Consultancy Services has recommended buy rating
    on the stock with a target price of Rs 3200 in its research report dated October 08, 2020."

Output:
{
  "annotation": [
    {
      "target": "Tata Consultancy Services",
      "snippet": "Prabhudas Lilladher is bullish on Tata Consultancy Services has recommended
    buy rating on the stock",
      "aspect": "Market/Stock Recommendation",
      "sentiment": "positive",
      "reasoning": "The recommendation to buy indicates a positive outlook on the stock's
    future performance."
    },
    {
      "target": "Tata Consultancy Services",
      "snippet": "target price of Rs 3200",
      "aspect": "Market/Stock Target Price",
      "sentiment": "positive",
      "reasoning": "A target price of Rs 3200 suggests an expected increase in stock value,
    indicating a positive sentiment."
    }
  ]
}

Example 2:
Input: "The Mumbai-based auto major had dispatched 61,114 units to dealers in October 2022."

Output:
{
  "annotation": [
    {
      "target": "Mumbai-based auto major",
      "snippet": "dispatched 61,114 units to dealers in October 2022",
      "aspect": "Corporate/Operational Health",
      "sentiment": "positive",
      "reasoning": "The dispatch of a large number of units indicates strong operational
    performance and demand fulfillment capability."
    }
  ]
}
```

Note: The API configuration details have been omitted for security reasons.