Towards Robust Urdu Aspect-based Sentiment Analysis through Weakly-Supervised Annotation Framework

Zoya Maqsood[1] **Seemab Latif**[1] **Rabia Latif**[2]
[1]School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST), Islamabad, Pakistan
{zmaqsood.phdcs17seecs, seemab.latif}@seecs.edu.pk
[2]College of Computer and Information Sciences (CCIS),
Prince Sultan University, Riyadh, Saudi Arabia
rlatif@psu.edu.sa

Abstract

Aspect-Based Sentiment Analysis (ABSA) remains largely unexplored in low-resource languages like Urdu due to the absence of large-scale, publicly available, and domain-diverse annotated corpora. Additional challenges like the scarcity of lexical resources, unstructured Urdu websites, and linguistic complexities, further hinder corpus development. These limitations create a critical bottleneck that prevents robust Urdu ABSA systems from being deployed in practical scenarios. We address this gap by proposing a weakly supervised framework that automates corpus annotation (∼10K Budget tweets) leveraging seed-based pattern matching with dynamic window analysis. Through a comparative analysis of Large Language Models (LLMs), and human annotations on expertly curated datasets, we further demonstrate the inherent complexity of Urdu ABSA. Suboptimal results from a conventional LSTM model that achieved a mean performance of 0.52 precision, 0.49 recall, and 0.50 F1 score across various ABSA tasks validate this challenge. In short, this work establishes a scalable and cost-effective annotation framework that advances ABSA research for Urdu and similar low-resource languages.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained Opinion Mining (OM) domain that evaluates sentiment toward specific attributes of entities, offering valuable insights for customer feedback analysis, product benchmarking, and market trend monitoring (Zhang et al., 2022). ABSA comprises four key elements: aspect category *(c)*, aspect term *(a)*, opinion term *(o)*, and sentiment polarity *(p)*. Figure 1 illustrates these elements through an annotation example of a customer's review. Examining single or multiple combinations of these elements to understand opinions in diverse scenarios gives rise to various ABSA tasks (Maqsood, 2023). For instance, the extraction of aspect categories constitutes the Aspect Category Detection (ACD) task, whereas sentiment analysis over these categories leads to the Aspect Category Sentiment (ACS) task. Similarly, evaluating sentiment toward explicit aspect terms constitutes the Aspect Sentiment Classification (ASC) task.

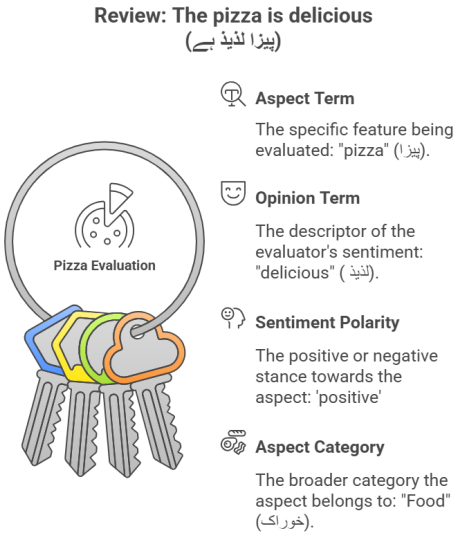

Figure 1: Aspect Based Sentiment Analysis

A fundamental prerequisite for any OM system is an accessible benchmark corpus of annotated reviews (Zhou et al., 2019; Hu et al., 2021). This requirement becomes particularly acute for low-resource languages like Urdu, where despite substantial social media presence, available ABSA datasets remain inadequate characterized by non-public availability, absence of benchmark standards, sparse annotations, and limited-domain coverage (Rani and Anwar, 2020; Ahmad and Wan, 2021). While manual annotation of ABSA elements becomes prohibitively expensive for large-scale datasets containing multi-aspect sentences in various domains. Additional challenges include the

scarcity of lexical resources, prevalent use of non-standard encoding in Urdu web content, unique linguistic features, and informal language on social media (Khattak et al., 2021). These constraints collectively impede corpus development, creating significant barriers in building robust Urdu ABSA models and complicating the adaptation of existing methodologies (Zhou et al., 2019; Liu et al., 2020; Zhang et al., 2022).

Besides, leveraging weak supervision has demonstrated potential in the realm of social media mining (Maqsood, 2023; Tekumalla and Banda, 2023). Although weak labels may not achieve manual-level precision but they enable rapid dataset expansion and robust model training especially when combined with a subset of high-quality manual labels (Zhang et al., 2022). Despite the success of Large Language Models (LLMs) like GPT-4.0 and DeepSeek in capturing linguistic patterns, these approaches have not been widely explored in existing literature, particularly for dataset annotations in Urdu. While, applying English-centric models to translated Urdu tweets exacerbates the issue, yielding poor results due to translation quality limitations (Zhang et al., 2021).

This work pioneers Urdu ABSA by introducing a weakly supervised annotation framework that automates labeling of all core ABSA elements for the 'Budget' domain, overcoming dataset scarcity without costly manual effort. Our systematic evaluation reveals LLMs (GPT-4, DeepSeek) limited transferability to Urdu, while experiments demonstrate our method's superiority over them. To our knowledge, this constitutes the first comprehensive study of such techniques for Urdu. Baseline LSTM experiments further highlight Urdu-specific ABSA challenges, underscoring the need for advanced architectures. Our key contribution addresses Urdu's critical resource gap through scalable dataset creation methodology that eliminates manual annotation bottleneck to facilitate fine-grained Urdu ABSA.

2 Related Work

This section discusses the existing Urdu datasets developed for opinion mining tasks, analyzing their annotation methodologies, and domain applicability.

2.1 Opinion Mining Datasets in Urdu

Researchers contributed to the field of Urdu sentiment analysis by presenting annotated corpora, but most focus on document- or sentence-level sentiment classification rather than fine-grained ABSA. Early efforts, such as those by Rani and Anwar (2020), introduced a manually annotated corpus of 10,000 tweets from sports domains (cricket and football), labeling aspects, categories, and polarities. However, the absence of opinion term annotations limits applications of ABSA tasks. Similarly, ul Haq et al. (2020) presented a corpus of 8,760 political tweets with polarity and four category labels but did not annotate aspect terms or opinion expressions, restricting deeper sentiment analysis. Moreover, their dataset is not publicly available and labeled manually, hindering scalability and reproducibility.

To address the scarcity of ABSA-specific resources, Ahmad and Wan (2021) translated the SemEval-2014 ABSA dataset (2951 restaurant and 4721 laptop reviews) into Urdu, providing aspect terms, polarities, and category labels. While this enables some ABSA experimentation, the reliance on machine translation raises concerns about linguistic accuracy and cultural relevance. Other datasets, such as Ghafoor et al. (2023) introduced SentiUrdu1M dataset (1 million tweets), leveraging large-scale emoticon-based labeling but remain unsuitable for ABSA due to their document-level granularity. Similarly, Amjad et al. (2021) curated a dataset of 3,564 tweets for threat detection, but its binary classification focus makes it irrelevant for aspect-level sentiment tasks. Beyond Twitter data, researchers have collected Urdu reviews from blogs and news platforms. (Mukhtar and Khan, 2018; Mukhtar et al., 2017; Khan et al., 2021; Rehman and Bajwa, 2016) developed datasets with manually annotated sentiment labels at document-level and suffer from limited domain coverage (e.g., movies, electronics). Additionally, many of these datasets are not publicly available, and their annotation methodologies are often poorly documented, reducing their utility for ABSA research.

In short, existing Urdu sentiment analysis datasets lack fine-grained annotations, suffer from small sizes and narrow domains, and use inconsistent annotation methodologies. Most rely on either biased translations or labor-intensive manual labeling, which impedes scalability. Furthermore, existing resources neglect weakly-supervised

approaches, while available multilingual models and LLMs remain under-evaluated. Overall, these limitations underscore the dire need for comprehensively annotated Urdu ABSA datasets in several domains by combining both manual and automated annotation methods. This hybrid methodology ensures both high-quality annotations and efficient scalability, ultimately enabling advanced techniques for progress of ABSA in Urdu language.

3 Dataset

We collected approximately 13,000 tweets related to Pakistan's budgetary domain between May and July 2020 using Twitter's Standard API. Due to API constraints, tweets were gathered in daily batches, limited to a 7-day historical window, with a maximum of 100 tweets per query and 180 requests per 15-minute interval. The search queries focused on trending budgetary discourse in Pakistan, incorporating hashtags such as '#Budget2020', '#PakistanEconomy', and '#Commerce'. The dataset provides a comprehensive representation of public sentiments and economic debates surrounding Pakistan's budget during the unprecedented COVID-19 lockdown period.

3.1 Pre-processing

The collected tweets underwent an extensive three-stage pre-processing pipeline to ensure data quality and linguistic consistency.

Tweet Level: We performed Unicode normalization to address Arabic script variations, removed punctuations, and social media artifacts (emojis, hashtags, URLs, mentions) using regular expressions. We eliminated duplicate entries and truncated excessive consecutive repetitions (e.g., reducing "سلیکٹڈ بجٹ..." (selected budget selected budget...) to "سلیکٹڈ بجٹ") (selected budget) to maintain textual conciseness.

Token Level: After conducting a systematic comparison of tokenization approaches Qi et al. (2020), Ali (2020), Vasiliev (2020) and space-based methods, we preferred UrduHack for its superior performance on informal Urdu text. Use of informal language and noise on social media limit the effectiveness of language-specific tokenizers, introducing abnormal tokens. We analyzed incorrect tokens to identify the inherent patterns of their abnormalities and normalized them accordingly. This includes splitting merged stopwords (e.g., تھیاک → تھی اک), reducing character repetitions in misspelled words (e.g., پاکسسستان → پاکستان), and eliminating word repetitions (e.g., فریفریفری → فری).

Character Level: The final processing step validated individual characters against Urdu Unicode ranges and removed residual artifacts (e.g., cleaning "***" and normalizing "ہوےc002uنے کا" to "ہوے نے کا").

This hierarchical pre-processing approach, documented comprehensively in Zoya et al. (2023), resulted in dataset of approximately 10,000 tweets.

3.2 Dataset Variants

We created three versions of the dataset, introducing variation in the annotation process, as outlined below:

Bronze Standard Dataset (BS): This dataset is a raw output without manual curation from our weakly supervised annotation system.

Silver Standard Dataset (SS): This represents a refined version of the 'BS' dataset. The corpus underwent a meticulous validation process combining automated consistency checks with expert human verification to ensure higher annotation quality. This approach filtered out erroneous labels generated by our weakly supervised methods and resulted in an 13% reduction of the original dataset labels.

Gold Standard Dataset (GS): The GS dataset was constructed through rigorous manual annotation by three native Urdu speakers with expertise in NLP. From the SS corpus, we selected a representative subset of 3000 tweets for fine-grained annotations. Three annotators followed strict annotation guidelines of Pontiki et al. (2014) standards, with only labels receiving consensus from at least two annotators being retained. The GS corpus serves as a reliable ground truth for evaluating model performance on Urdu ABSA tasks, while also revealing additional linguistic patterns not captured in the initial SS annotations. The statistics about these datasets have been described in Table 1.

Dataset	Tweets	Asp_Cat.	Asp_Terms	Opinion_Terms
Bronze	9693	14	5179	5456
Silver	8949	14	4247	5364
Gold	3000	14	1126	1410

Table 1: Statistics of Datasets with Distinct Values

4 Methodology

We present our methodology for annotating Urdu datasets for ABSA. First, we highlight the limita-

tions of LLMs for this task, followed by our custom framework designed to address these challenges.

4.1 LLMs Limitations for Dataset Annotation

The utilization of the GPT 4.0 and DeepSeek models for dataset annotation in Urdu revealed several challenges. Firstly, the model encountered challenges in thoroughly capturing all aspects and sentiment words present in tweets. Secondly, an inherent instability in labeling responses was observed, as the model exhibited varying results for the same query when executed multiple times. Thirdly, the issue of selecting irrelevant words alongside sentiment and aspect terms introduced a lack of uniformity, necessitating post-processing efforts for pruning. Fourthly, the model tended to repeat sentiment terms within aspect terms or vice versa. Fifthly, breaking down tweets into shorter chunks did not significantly improve their response quality. Sixthly, the model demonstrated a tendency to ignore rare words and occasionally overlook crucial aspects. In conclusion, LLMs exhibited limitations in fully grasping the context. A representative case of tweet annotation generated by the LLM in Figure 4 (see appendix).

4.2 Dataset Annotation Framework for ABSA in Urdu

Our dataset labeling approach encompasses two fundamental phases: ACD and the annotation of Aspect-Opinion-Sentiment (AOS) triplet. Initially, ACD was completed through topic modeling and clustering techniques. Subsequently, the identification of triplet components within tweets was carried out through methods like pattern mining and a bidirectional window-based labeling strategy.

4.2.1 Aspect Category Detection

We used pre-trained sentence transformers Reimers and Gurevych (2019) to generate embeddings and applied both Top2Vec (Angelov, 2020) and traditional clustering algorithms (Ackermann et al., 2014; Frey and Dueck, 2007) to identify nuanced subtopics. Our analysis revealed optimal cluster counts (39) based on cosine similarity metrics and cluster validation techniques (Kaoungku et al., 2018; Yuan and Yang, 2019). Notably, Top2Vec initially predicted 54 topics, but these were ultimately clustered within the same range.

To reduce cluster overlap, we performed graph-based analysis, where edges represented cosine similarities between embeddings. Edge weights were set to 0 for similarities below a threshold of 0.7, ensuring that only highly similar tweets were grouped together while preserving distinct terms across clusters. However, some topics (clusters) exhibited irrelevance like synonymous terms or polysemy of less substantial words. Such problematic topics featuring highly coherent terms could form distinct clusters, leading to favourable scores in standard metrics. Conversely, some significant topics might be overlooked due to lower coherence or similarity scores between words, particularly if such topics cover diverse perspectives not covered well in the coherence metrics reference corpus. To address these limitations, we incorporated a manual curation step to refine and consolidate topics. From the generated clusters, we selected distinct categories (Table 3 in section 8) and further subdivided broad topics by analyzing top topic words (Table 4 in section 8). For example, 'Social Welfare' was divided into 'Education', 'Agriculture', 'Health', and 'Social Programs'.

4.2.2 Aspect-Opinion-Sentiment Triplet Annotation

This triplet annotation process is divided into four fundamental stages, which include word classification, seed enrichment, tweet labeling, and evaluation, as discussed below:

Words Classification: The selected topics consist of a mixture of terms related to aspect and opinion that require further classification. To systematically categorize these terms (seed terms), we adopted a straightforward yet effective approach: nouns were designated as aspect, while adjectives were treated as indicators of opinion. The selection process for seed terms emphasized domain relevance, frequency, and diversity, ensuring the chosen nouns were explicit and closely related to core topics. To enhance relevance, we excluded irrelevant or ambiguous terms, rare occurrences, verbs, adverbs, and any expressions introducing sentiment bias or lacking clear aspect association. This rigorous selection process resulted in a refined lexicon of aspect and opinion seeds, with the complete workflow detailed in Figure 2.

Seeds Enrichment: Given the limited coverage of initial seed words for annotating all tweets within each subtopic, we employed multiple strategies to expand and refine our seed term collection. As illustrated in Figure 3, our enrichment approach incorporated sentiment lexicons, active learning,

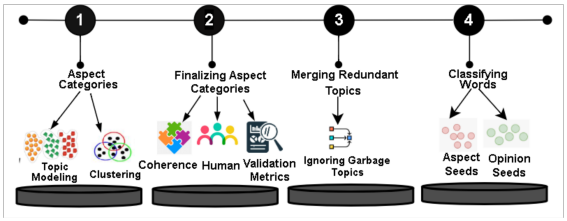

Figure 2: Aspect Categories and Words Classification Process

pattern mining, and embedding-based methods.

Sentiment Lexicon: We used an existing Urdu

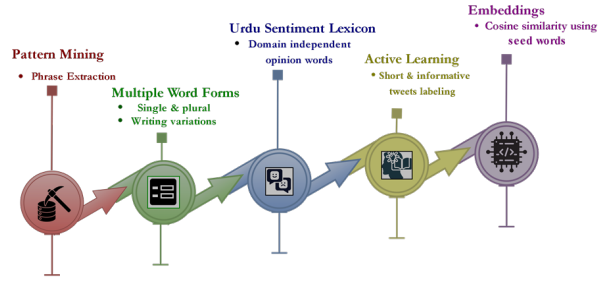

Figure 3: Sentiment Classification From Predicted Topics

sentiment lexicon [42] to identify polar expressions within our tweet corpus. This process yielded 1,322 positive and 1,395 negative terms that overlapped between the lexicon and our budget-related tweets. However, the lexicon exhibited notable limitations: its coverage of domain-specific fiscal terminology was incomplete, and its formal vocabulary often mismatched the informal expressions and morphological variations prevalent in social media. Similar challenges emerged when we attempted to use translated lexicons intended for sentiment analysis in the English language.

Multiform words: We took into account various word forms in our seed terms, including singular and plural forms, such as "قیمت" (price) in singular and "قیمتیں" or "قیمتوں" (prices) in plural. Urdu's rich inflectional system, where words vary by tense, gender, number, and loanword integration, renders conventional lemmatization and stemming ineffective. For instance, contextual variants (e.g., verb conjugations or gendered forms) lack reliable root-mapping rules. Moreover, Urdu has a diverse vocabulary with numerous loanwords and context-dependent variations that further complicate such tasks. Consequently, we excluded this step to preserve semantic precision given the absence of robust Urdu-specific linguistic tools.

Active Learning:. We utilized an active learning approach and created a preliminary dataset consisting of hundred short tweets (5-10 words each). Subsequently, we conducted manual labelling with a specific focus on AOS triplets. This step provided valuable insights for various types of words beyond the seed words and their contextual relationships within the domain. Additionally, we noted the prevalence of multi-word phrases as opposed to single words for seed terms. We quantified phrase frequencies and their sentiment associations, iteratively expanding the seed lexicon to include high-impact multi-word terms. This comprehensive examination not only enriched our seed inventory but also our understanding of the multifaceted language used in the dataset.

Pattern Mining: Based upon the manually labeled data from active learning, we identified recurring patterns that encompassed consecutive domain-specific words and Urdu case markers (کا (ka), کے(kay), کی(ki), کو(ko), میں(mein), پر(par), سے(se), نے(nay)). We developed a hybrid pattern mining approach combining rule-based and statistical techniques. This integrated approach revealed important multi-word expressions that served as more precise indicators of aspects and opinions compared to conventional single-word seeds. We first analyzed recurring syntactic structures involving domain-specific terms paired with Urdu case markers. Matching these patterns against tweets is depicted in Algorithm 1 and a comprehensive list of extracted patterns is provided in Table 5 (Section 8). We then implemented a sequential pattern mining algorithm with minimum support thresholds to discover statistically significant co-occurring word sequences, prioritizing longer phrases that captured more nuanced meanings. The extracted patterns enabled us to automatically identify aspect-opinion pairs in new tweets. For instance, in the structure "[X] ka [Y]", X was classified as the aspect term and Y as the opinion term. Sentiment polarity was then assigned to these newly discovered opinion terms through contextual analysis, leading to the formation of AOS triplets (detailed in Appendix Algorithm 2).

Embeddings: We utilized the pre-trained embedding model FastText to identify the top 10 words that exhibited the highest cosine similarity with our seed terms, particularly focusing on expanding our set of opinion words. Additionally, we considered terms returned by the Top2vec model that exhib-

Algorithm 1 Patterns Matching Algorithm

```
procedure EXTRACTPATTERNS(budget_tweets)
    pattern ← (\w+\sکا\sبجٹ)
    extracted_patterns ← EmptyList()
    for each tweet in budget_tweets["tweet_text"] do

        matches ← FindAllMatches(pattern, tweet)
        for each match in matches do
            pattern_text ← match.group(1)

            AddPatternToExtractionList(extracted_patterns, pattern_text)
        end for
    end for
    return extracted_patterns
end procedure
```

ited similarity with seed terms by surpassing the 0.5 threshold in similarity score. We selectively kept words that fell within the categories of aspect or opinion-related terms. Any words failing to meet these criteria were excluded from further consideration. Exemplary instances have been presented in the Table 6 (Appendix).

Labeling Tweets: Initially, we annotated using mined patterns with analogous structures, resulting in the creation of AOS triplets while accounting for sentiment reversals caused by negators (e.g., no, not). Annotated example can be seen in Figure 5 (in Appendix), proved effective for contiguous word patterns but limited for non-adjacent term relationships. To address this limitation, we introduced a window-based annotation strategy consisting of two main steps: seeds cartesian product with sentiment polarity assignment and seeds co-occurrence analysis.

Seeds Cartesian Product with Sentiment Polarity Assignment: In this phase, we performed a Cartesian product operation between the aspect seeds and sentiment seeds to form their pairs (a, o). Despite that sentiment polarity was already predefined in lexicons for numerous opinion words, several pairs underwent cross-validation due to domain-specific variations or informal language use. As the sentiment of the same opinion word may vary based on its association with different aspect words. For example, the term increment is considered positive when associated with salary but negative when linked with poverty.

Window-based Strategy: We implemented a dynamic window-based approach to detect co-occurring aspect-opinion (a, o) pairs within tweets. The tweet segmentation process entailed setting a token threshold of 15 words from both the beginning and end. Subsequently, we systematically examined the co-existence of (a, o) pairs within these segments. Meanwhile, we addressed negators when they co-occurred within a segment in the context of the (a, o) pair, and selectively inverted the sentiment for that particular (a, o) pair occurrence in the given tweet. Likewise, we advanced to the next segment by adjusting the window after every five words. Eventually, we reconstructed the original tweet and gathered all unique AOS triplets from every segment of a tweet. This comprehensive methodology allowed us to discern nuanced sentiment variations associated with (a, o) pairs within the dynamic context of tweets. The entire process is summarized below and labeled tweet result is presented in Figure 6 (Appendix).

$$segment\_length = 15$$

$$window\_size = 5$$

$$S_i = \text{Segment}(t\_i, segment\_length)$$

$$C_{i,j} = \text{CoExistence}(S_{i,j}, (a, o))$$

$$N_{i,j} = \text{NegatorHandling}(C_{i,j}, \text{negators})$$

$$S_i = \text{NextSegment}(S_i, window\_size)$$

$$R_i = \text{Reconstruct}(S_i, N_i)$$

$$AOS_i = \text{GatherTriplets}(R_i)$$

Evaluation: To assess label quality, we employed dual evaluation approaches as given below:
Automated Evaluation: We compared our methodology's outputs against both LLMs annotations and Gold-standard (GS) labels on identical tweet samples. Our analysis revealed that our proposed approach mostly outperformed LLMs in AOS triplet accuracy.

Human Evaluation: We developed a weakly-supervised validation protocol addressing multi-aspect tweets where window-based strategies occasionally produced spurious aspect-opinion associations in case of multiple aspects. The validation process involved: (1) categorizing tweets by presence of aspect complexity (single/multiple), (2) cross-referencing novel multi-aspect pairs with pre-labeled single-aspect examples and pattern-mined results, and (3) manual verification of unmatched pairs on a sample representing at least 2% of the tweets containing each such pair. If more than 50% of the labels were deemed accurate in the chosen sample, we retained them as final labels.

Finally, to ensure label consistency across datasets, we performed comparative analysis by identifying tweet overlaps between all dataset variants and discrepancies were compared against the GS labels. Then the F1 measure and accuracy were computed (Table 2) as defined in (Pontiki et al., 2014) and expressed below:

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

where precision (P) and recall (R) were determined as:

$$P = \frac{|SS \cap GS|}{|SS|} \tag{2}$$

$$R = \frac{|SS \cap GS|}{|GS|} \tag{3}$$

$$Acc. = \frac{|GS \cap SS|}{|SS \cup GS|} \tag{4}$$

Label	Acc.	P	R	F1
Aspect	69.9	91.7	74.6	82.3
Opinion	71.4	89	80.9	84.8
Polarity	73.6	88.5	74.4	80.8
Category	86.3	93.1	89.8	91.4

Table 2: Scores of evaluation measures on annotated dataset labels.

5 Experimental Set-Up

5.1 Tasks

We performed experiments on three key ABSA tasks, as given below:

Aspect Category Detection (ACD): Identifying the categories for each tweet from a set of predefined aspect categories.

Aspect Category Sentiment (ACS): Sentiment polarity classification (positive/negative/neutral) toward detected aspect categories.

Aspect Sentiment Classification (ASC): Sentiment polarity analysis targeting explicit aspect terms.

5.2 Model

We implemented LSTM as our baseline model initialized with 300-dimensional FastText embeddings. The model was trained with a batch size of 32, hidden state dimension of 300, and the adam optimizer (learning rate = 0.001) for 100 epochs. To ensure robustness, we ran five training repetitions using categorical cross-entropy loss. Hyperparameters were tuned via Grid search, testing epochs [10, 50, 100, 300], embedding dimensions [100, 300], learning rates [0.001, 0.01, 0.0001], batch sizes [16, 32, 64, 128], and dropout rates [0.2, 0.3, 0.5], with early stopping (patience = 5) and stratified 5-fold cross-validation. The hyperparameter grid values are chosen based on optimal LSTM performance observed in sentiment analysis-related studies (Kumar et al., 2021; Naqvi et al., 2021).

5.3 Dataset Distribution

We implemented a rigorous train-test split (Table 7 in Section 8) on the SS dataset to maintain proportional representation of both aspect categories and sentiment polarities. The partitioning preserved identical distributions of positive, negative, and neutral sentiment labels across training (75%) and testing (25%) subsets for each aspect category. The equal percentage distribution provides a balanced representation for classifier training and fosters robust model development by minimizing biases through learning from comparable instances across various aspect categories.

5.4 Results Analysis

The LSTM baseline results reveal a consistent performance trend across datasets (results in Appendix 8). For ACD, the GS achieves strong performance at 100 epochs, while SS and BS show gradual improvements, peaking at 0.596 and 0.562 accuracy, respectively. This aligns with the high F1 scores (91.4 for Category, 82.3 for Aspect) in Table 2, confirming that our annotation framework produces usable labels. In ACS, the GS reached near-ceiling macro-F1 (0.877) by 50 epochs, whereas SS and BS plateau at $\sim$0.490.59 F1. This reflects the challenge of sentiment polarity prediction. The SS consistent lead over BS dataset justifies our refinement step, though both trail Gold due to inherent noise. For ASC, all datasets struggle (F1 < 0.35), mirroring the difficulty of fine-grained sentiment analysis. The marginal gains with more epochs suggest the LSTMs limited capacity to resolve ambiguities. Traditional LSTM is viable for coarse tasks (ACD) but face limitations in sentiment-related tasks. However, the LSTM model was intentionally selected as a lower-bound baseline to assess the discriminative strength of annotation quality and task difficulty, without the confounding influence of pretraining or large-scale parameters in advanced architectures. Despite balanced splits,

macro-F1 scores highlight challenges from label imbalance, multi-label learning, and Urdu's morphological complexity. Progressive performance gains from (Bronze→Silver→Gold) highlight annotation quality as a stronger factor than model complexity.

6 Discussion

The proposed weakly-supervised framework demonstrates significant advancements in Urdu ABSA by overcoming the critical bottleneck of manual annotation in dataset creation. The multidimensional annotation requirements, encompassing all ABSA elements, render fully manual annotation impractical for scalable model development due to its time-consuming nature and human labor requirements. Our framework automates this process, starting with a seed-based approach for high-precision in noisy, code-mixed Urdu social media text and mitigate limitation of domain coverage through iterative enrichment using lexicon expansion, syntactic patterns, and contextual embedding strategies. This dynamic refinement transforms static seeds into a robust, domain-adaptive seeds inventory suited for low-resource and informal text settings. Thus, the core strength lies in the novel integration of context-aware seed expansion and morphologically-sensitive preprocessing, which collectively reduce annotation costs.

Furthermore, the method demonstrates robust capability in handling Urdu's linguistic complexities through its hybrid approach combining n-gram pattern matching with dynamic window labeling. This approach effectively identifies multi-word aspects, such as "پٹرول کی قیمت" (petrol price), and successfully resolves polarity inversion cases by incorporating negation scope detection. Additionally, an automated validation pipeline was introduced that minimize human effort to maintain label quality. The limited variation with Gold-Standard dataset underscores the significance of high-quality annotations from our proposed method. Comparative analysis with prevailing LLMs reveals the proposed framework achieves substantially better performance for annotation task in Urdu, especially for aspect and opinion terms extraction tasks. These advancements establish a practical foundation for Urdu ABSA where fully supervised approaches remain infeasible due to resource constraints. Regarding classification results, the performance of conventional models like LSTM across ABSA tasks and datasets are emphasized. Despite extended training, the baseline LSTM's limited improvement reveals its inability to capture Urdu's linguistic nuances in ABSA tasks. There were instances where additional epochs do not yield significant gains, suggesting a potential saturation point in the models learning curve. Although our evaluation is constrained to the budget domain due to the availability of gold-standard annotations, framework's core components such as linguistic and syntactic pattern rules, clustering mechanism, and seed augmentation are domain-independent and easily adaptable to other domains.

6.1 Limitations

The proposed dataset annotation methodology endeavors to address many challenges, yet certain issues persist. Sentiments occurring beyond segment window lengths are occasionally overlooked, although this is mitigated by considering segments from multiple positions within tweets. The exclusion of tweets lacking seed terms may inadvertently dismiss relevant sentiment expressions. Overlapping labels or spurious associations may emerge occasionally when a sentiment word applies in multiple perspectives or simultaneously relates to multiple aspects within a tweet. In cases like sarcasm, where the same word is employed in diverse contexts (positively or neutrally), priority is determined based on its frequency of occurrence. Polysemous terms labeling (e.g., for budget pass vs. approach) risks errors, highlighting needs for context-aware rules.

7 Conclusion

Our research introduced a novel weak supervision methodology for creating a benchmark dataset in Urdu ABSA. We shed light on the inherent challenges in ABSA for under-resourced languages and made a significant contribution to addressing the resource scarcity in Urdu ABSA. Our dataset encompasses tweets within the budget domain and was annotated at four distinct levels: aspect, opinion, sentiment, and category levels. The consistently high F1 scores across all label annotations demonstrate the proposed method's effectiveness in producing high-quality. Through a detailed comparative analysis involving LLMs and human annotations based on expertly curated datasets, we illuminated the intricate nature of our proposed dataset. Empirical evaluations utilizing LSTM model showed limita-

tions of conventional methods for various ABSA subtasks and laid the groundwork for future advancements in ABSA techniques for Urdu.

8 Future Work

We aim to generalize our methodology to expand ABSA dataset annotations into other domains. Our focus will extend to advanced deep learning techniques, moving beyond basic LSTM models for diverse ABSA tasks in Urdu. We plan to conduct fine-tuning pre-trained models on an extended dataset across various domains for a comprehensive understanding of Urdu sentiment expressions. In summary, our future trajectory involves leveraging advanced techniques, annotating diverse datasets, and refining models for domain-specific applications, ultimately enhancing Urdu ABSA tools.

References

Marcel R Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. 2014. Analysis of agglomerative clustering. *Algorithmica*, 69:184–215.

Naveed Ahmad and Jing Wan. 2021. Aspect based sentiment analysis for urdu. In *2021 6th International Conference on Computational Intelligence and Applications (ICCIA)*, pages 309–313.

Ikram Ali. 2020. Urduhack: A python library for Urdu language processing. *CoRR*, abs/2010.06810. ArXiv: 2010.06810.

Maaz Amjad, Noman Ashraf, Alisa Zhila, Grigori Sidorov, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. Threatening language detection and target identification in Urdu tweets. *IEEE Access*, 9:128302–128313.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *CORR*, abs/2008.09470. ArXiv: 2008.09470.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Sarang Shaikh, and Rakhi Batra. 2023. Sentiurdu-1m: A large-scale tweet dataset for urdu text sentiment analysis using weakly supervised learning. *PLOS ONE*, 18(8):e0290779.

Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-label few-shot learning for aspect category detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6330–6340, Online. Association for Computational Linguistics.

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop. 2018. The silhouette width criterion for clustering and association mining to select image features. *International journal of machine learning and computing*, 8(1):69–73.

Lal Khan, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, and Alexander Gelbukh. 2021. Urdu sentiment analysis with deep learning methods. *IEEE Access*, 9:97803–97812.

Asad Khattak, Muhammad Zubair Asghar, Anam Saeed, Ibrahim A Hameed, Syed Asif Hassan, and Shakeel Ahmad. 2021. A survey on sentiment analysis in urdu: A resource-poor language. *Egyptian Informatics Journal*, 22(1):53–74.

Avinash Kumar, Aditya Srikanth Veerubhotla, Vishnu Teja Narapareddy, Vamshi Aruru, Lalita Bhanu Murthy Neti, and Aruna Malapati. 2021. Aspect term extraction for opinion mining using a hierarchical self-attention network. *Neurocomputing*, 465:195–204.

Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

Zoya Maqsood. 2023. Weakly supervised learning for aspect based sentiment analysis of urdu tweets. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 78–86.

Neelam Mukhtar and Mohammad Abid Khan. 2018. Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02):1851001.

Neelam Mukhtar, Mohammad Abid Khan, and Nadia Chiragh. 2017. Effective use of evaluation measures for the validation of best classifier in urdu sentiment analysis. *Cognitive Computation*, 9:446–456.

Uzma Naqvi, Abdul Majid, and Syed Ali Abbas. 2021. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082. ArXiv: 2003.07082.

Sadaf Rani and Waqas Anwar. 2020. Resource creation and evaluation of aspect based sentiment analysis in urdu. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 79–84.

Zia Ul Rehman and Imran Sarwar Bajwa. 2016. Lexicon-based sentiment analysis for urdu language. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 497–501.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ramya Tekumalla and Juan M Banda. 2023. Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Computing and Applications*, 35(25):18161–18169.

Ehsan ul Haq, Sahar Rauf, Sarmad Hussain, and Kashif Javed. 2020. Corpus of aspect-based sentiment for urdu political data. In *Mexican International Conference on Artificial Intelligence*, pages 37–40. Springer.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press.

Chunhui Yuan and Haitao Yang. 2019. Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *CoRR*, abs/2202.05433. ArXiv: 2202.05433.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230.

Jie Zhou, Jimmy Xiangji Huang, Qin Chen, Qinmin Vivian Hu, Tingting Wang, and Liang He. 2019. Deep learning for aspect-level sentiment classification: survey, vision, and challenges. *IEEE access*, 7:78454–78483.

Zoya, Seemab Latif, Rabia Latif, Hammad Majeed, and Nor Shahida Mohd Jamail. 2023. Assessing urdu language processing tools via statistical and outlier detection methods on urdu tweets. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(10):1–31.

Appendix

Prompt: *Extract all aspect expressions with their corresponding opinion expressions and sentiment polarity (positive/ negative/neutral) in given text:*

Aspect-Opinion-Sentiment Analysis

Text:

"ملک معاشی طور پر دیوالیہ ہو چکا ہے اور بجٹ بنانے میں دقت پیش آ رہی ہے۔"

Extracted Pairs:

Aspect Term (پہلو)	**Opinion Term** (رائے)	**Sentiment Polarity** (جذباتی کیفیت)
ملک کی معاشی حالت (Economic condition)	دیوالیہ ہو چکا ہے (Bankrupt)	**Strong Negative** (انتہائی منفی)
بجٹ سازی (Budget preparation)	دقت پیش آ رہی ہے (Facing difficulty)	**Negative** (منفی)

Figure 4: DeepSeek response to Annotate Tweet

S#	Label	Top 15 Words
1	سماجی_پروگرام Social Welfare	ہیلتھ, روزگار, وایرس, کسان, وباء, زراعت, اصلاحات, ایجوکیشن, تعلیم, ڈاکٹر, فنکاروں, صحت, طلبہ, شعبہ, محکمہ 'Health', 'Employment', 'Virus', 'Farmer', 'Epidemic', 'Agriculture', 'Reforms', 'Education', 'Learning', 'Doctor', 'Artists', 'Health', 'Students', 'Department'
2	معیشت Economy	معاشی, مالی, مہنگا, پالیسی, سود, سالانہ, بحران, خسارے, معیشت, ڈالر, سود, معاشیات, اقتصادی, ترقیاتی, خزانہ 'Economic', 'Financial', 'Expensive', 'Policy', 'Interest', 'Annual', 'Crisis', 'Losses', 'Economy', 'Dollar', 'Interest', 'Economics', 'Economic', 'Development', 'Treasury'
3	میڈیا Media	تفصیلات, اطلاعات, پریس, رپورٹ, معلومات, میڈیا, صحافی, نیوز, خبر, احتجاج, تقریر, اعلان, عوامی, اداروں, حکومتیں 'Details', 'Information', 'Press', 'Report', 'Knowledge', 'Media', 'Journalist', 'News', 'Protest', 'speech','Report', 'Announcement', 'Public', 'Institutions', 'Governments'
4	سیاست Politics	سرکاری, وفاقی, صدر, حکمران, اپوزیشن, پاکستان, وزراء, سیاسی, ملک, ریاست, ایوان, اسمبلی, کابینہ, سیاست, پارٹی 'Government', 'Federal', 'President', 'Rulers', 'Opposition', 'Pakistan', 'Ministers', 'Political', 'Country', 'State', 'Assembly', 'Cabinet', 'Politics', 'Party'
5	مذہب Religion	اسلام, محمد, اسلامی, مسلم, گلوکار, اللہ, شریف, شیخ, علما, مدارس, دین, بجٹ, حکومت, سود, مدینہ 'Islam', 'Muhammad', 'Islamic', 'Muslim', 'Singer', 'Allah', 'Sharif', 'Scholar', 'Scholars', 'Schools', 'Religion', 'Budget', 'Government', 'Interest', 'Medina'
6	دفاع Defense	بم, پنشن, خطرہ, جنگ, افواج, پاک, دفاع, دشمن, دھشت, پولیس, ٹیکنالوجی, اصلاحات, اداروں, مراعات, عالمی 'Bomb', 'Pension', 'Threat', 'War', 'Forces', 'Pakistan', 'Defense', 'Enemy', 'Terror', 'Police', 'Technology', 'Reforms', 'Institutions', 'Consideration', 'Global'
7	منصوبہ Project	سبسڈی, پروجیکٹ, منصوبے, مختص, فنڈ, اخراجات, روپے, پیسہ, خرچ, قرضوں, کرپشن, فری, معاشرے, ترقیاتی, پالیسیوں 'Subsidy', 'Project', 'Projects', 'Specialized', 'Fund', 'Expenditure', 'Rupees', 'Money', 'Expense', 'Loans', 'Corruption', 'Free', 'Society', 'Development', 'Policy'

Table 3: Selected subtopics derived from predicted clusters and topic modeling

S#	Aspect Category	Attributes
1	معیشت Economy	بجٹ، عمومی، اخراجات، مہنگائی, قیمت، محصول، قرض، سود budget, general, expenditure, inflation, price, revenue, debt, interest
2	وفاق Federal	بجٹ، جنرل، وزیراعظم، صدر، کابینہ، اسمبلی، حکومت budget, general, prime minister, president, cabinet, assembly, government
3	تعلیم Education	بجٹ، جنرل، ادارے، استاد، طالب علم budget, general, institutions, teacher, student
4	صحت Health	بجٹ، جنرل، ڈاکٹر، صحت، ہسپتال، دیکھ بھال، وبا budget, general, doctor, health, hospital, care, epidemic
5	زراعت Agriculture	بجٹ، جنرل، زراعت، کسان، فصل، ٹڈی budget, general, Agriculture, farmer, crop, locust
6	سماجی_پروگرام Social Welfare Program	بجٹ، جنرل، اصلاحات، بینظیر انکم ٹیکس ترقیاتی پروگرام, اصلاحاتی پروگرام، انکم ٹیکس سپورٹ پروگرام budget, general, reforms, Benazir income tax, development program, reform program, income tax support program
7	دفاع Defense	بجٹ، جنرل، فوجی، تحفظ، حملہ budget, general, military, protection, attack
8	مذہب Religion	بجٹ، جنرل، مذہب، مومن، علماء، مقدس مقامات budget, general, religion, believers, scholars, holy places
9	سیاسی_جماعت Political Party	بجٹ، جنرل، پارٹی، پالیسی، کانفرنس، اپوزیشن budget, general, party, policy, conference, opposition
10	قیادت Leadership	بجٹ، جنرل، لیڈر، چیئرمین، کرپشن budget, general, leader, chairman, corruption
11	صوبائی Provincial	بجٹ، جنرل، صوبہ (پنجاب، سندھ، بلوچستان، پختونخوان) حکومت، کابینہ، اسمبلی budget, general, provinces (Punjab, Sindh, Balochistan, Pakhunkhawan) govt., cabinet, assembly
12	عوام Public Dynamics	بجٹ، جنرل، امیر، غریب، روزگار، تنخواہ، پنشن budget, general, rich, poor, employment, salary, pension
13	میڈیا Media	بجٹ، جنرل، صحافی، میڈیا، خبریں، چینل، رپورٹ، آرٹسٹ budget, general, journalist, media, news, channel, report, artist
14	جنرل Miscellaneous	بجٹ budget

Table 4: Conclusive sub-categories of budget topic

Patterns	'کا بجٹ' (Budget of)	'دوست بجٹ' (Friend's Budget)	'دشمن بجٹ' (Enemy's Budget)	'میں کمی' (Decrease in)
Phrases	امراء کا بجٹ (Budget of Aristocrats) تباہی کا بجٹ (Budget of Destruction) خسارے کا بجٹ (Budget of Loss) مافیا کا بجٹ (Mafia's Budget) تعلیم دوست بجٹ (Education-Friendly Budget)	انسان دوست بجٹ (Human-Friendly Budget) عوام دوست بجٹ (Public-Friendly Budget) تعلیم دوست بجٹ (Education-Friendly Budget) غریب دوست بجٹ (Poor-Friendly Budget) نوجوان دوست بجٹ (Youth-Friendly Budget)	انسانیت دشمن بجٹ (Inhumane Budget) برآمدات دشمن بجٹ (Incomes-Enemy Budget) صحت دشمن بجٹ (Health-Enemy Budget) مزدور دشمن بجٹ (Labor-Enemy Budget) مسلم دشمن بجٹ (Muslim-Enemy Budget)	حکومتی اخراجات میں کمی (Reduction in Government Expenditure) تعلیمی اخراجات میں کمی (Reduction in Educational Expenditure) بجٹ خسارہ میں کمی (Reduction in Budget Loss) فیسوں میں کمی (Reduction in Fee) پٹرولیم قیمتوں میں کمی (Reduction in in Petroleum Prices)
Total	240	20	39	59

Table 5: Phrases extracted by the Pattern Mining

Algorithm 2 : MineSequentialPatterns

```
procedure MINESEQUENTIALPATTERNS(budget_tweets)
    stopwords ← LoadStopwords()                                  ▷ Load stopwords
    ps ← PrefixSpanAlgo(data)                                    ▷ Initialize pattern mining algorithm
    min_support ← 20                                             ▷ Set minimum support
                                                                 ▷ Mine frequent patterns with minimum support
    result ← ps.Frequent(min_support)
                                                                 ▷ Filter patterns
    filtered_patterns ← FILTERPATTERNS(result, stopwords)
                                                                 ▷ Display and store patterns
    obt_patterns ← DISPLAYANDSTOREPATTERNS(filtered_patterns)
    return obt_patterns                                          ▷ Return the obtained patterns
end procedure
```

```
function FILTERPATTERNS(result, stopwords)
    filtered_patterns ← []                                       ▷ List for filtered patterns
    for each (support, pattern) in result do
                                                                 ▷ Check if pattern is valid
        if ISPATTERNVALID(pattern, stopwords) then
                                                                 ▷ Keep valid pattern to list
            filtered_patterns.append((pattern, support))
        end if
    end for
    return filtered_patterns                                     ▷ Return the filtered patterns
end function
```

```
function ISPATTERNVALID(pattern, stopwords)
                                                                 ▷ Check length of pattern
    if Length(pattern) > 1 then
                                                                 ▷ Count stopwords in pattern
        stopwords_count ← COUNTSTOPWORDS(pattern, stopwords)
        if stopwords_count ≤ 1 then
            is_subpattern ← False                                ▷ Initialize flag for subpattern
            for each (_, other_pattern) in result do
                                                                 ▷ Check if pattern is subset of other pattern
                if pattern ≠ other_pattern & ISSUBSET(pattern, other_pattern) then
                    is_subpattern ← True
                    break                                        ▷ Exit loop if subpattern is found
                end if
            end for
            if not is_subpattern then
                return True
            else
                return False
            end if
        else
            return False
        end if
    else
        return False
    end if
end function
```

Seeds	Top 10 similar words
قرض (Loan)	'اُدھار', 'پرقرض', 'مقروض', 'ادھار', 'قرضدار', 'کاپیس', 'قرضوں', 'قرضے', 'قرضہ', 'Risky' 'Loan', 'Loans', 'Debts', 'Cabinet', 'Debtor', 'Interest', 'Indebted', 'Owing', 'Debt', 'Risky'
حکومت (Govt.)	'وزارت', 'نوازحکومت', 'حکومتیں', 'بشارحکومت', 'کوصوبے', 'میحکومت', 'ہیحکومت', 'ےھےحکومت', 'حکومتوں', 'حکومتی' 'Governments', 'Governmental', 'Nawaz Govt.', 'Governments', 'Bashar Govt.', 'In Govt.', 'In Govt.', 'Are in Govt.', 'Is in Govt.', 'Ministry'
بجٹ (Budget)	'0084ارب', 'روڈنئی', 'شیڈوبجٹ', 'کابجٹ', 'ےکریلیف', '05ارب', '04ارب', '74کھرب57ارب', '005ارب', '006ارب' '4800 Billion', 'Rodney', 'Shadow Budget', 'Cabinet', 'Relief', '50 Billion', '40 Billion', '47 Billion 75 Million', '500 Billion', '600 Billion'
مہنگائی (Inflation)	'اورغربت', 'قیمتوں', 'قیمتیں', 'ہوشرباء', 'مہنگائی', 'مہنگائ', 'پھرمہنگائی', 'اورمہنگائی', 'کومہنگائی', 'پرمہنگائی' 'Hyperinflation', 'And Inflation', 'And Inflation', 'Then Inflation', 'Inflation', 'Inflation', 'Hosharba', 'Prices', 'Prices', 'And Poverty'
تنخواہ (Salary)	'تنخوا', 'تنخواہیں', 'ےتنخواہ', 'تنخواہ', 'تنخواہیں', 'تنخواہ', 'تنخواہیں', 'تنخواہوں', 'تنخوائیں', 'تنخواہو' 'Salary', 'Salaries', 'From Salary', 'Salary', 'Salaries', 'Salaries', 'Salary', 'Salaries', 'Salary', 'Salaries'

Table 6: Most Similar words by FastText model

"Urdu Tweet": " غریب دشمن بجٹ نامنظور دھاندلی کی حکومت نامنظور پی ٹی آئی ایم ایف بجٹ نامنظور مزدور دشمن
بجٹ نامنظور عوام دشمن بجٹ نامنظور"

"Translated Tweet": "Anti-Poor budget disapproved. The government of rigging disapproved.
PTIMF budget disapproved. Anti-labor budget disapproved. Anti-public budget disapproved.


```
"entries": [
    {
      "Aspect": "بجٹ" (budget),
      "Opinion": "عوام دشمن" (anti-public),
      "Category": "بجٹ" (budget),
      "Polarity": "Negative"
    },
    {
      "Aspect": "بجٹ" (budget),
      "Opinion": "نامنظور (disapprove)",
      "Category": "بجٹ" (budget),
      "Polarity": "Negative"
    },
    {
      "Aspect": "بجٹ" (budget),
      "Opinion": " مزدور دشمن" (anti-labor),
      "Category": "بجٹ" (budget),
      "Polarity": "Negative"
    },
    {
      "Aspect": "بجٹ" (budget),
      "Opinion": " غریب دشمن " (anti-public),
      "Category": "بجٹ" ( budget),
      "Polarity": "Negative"
    },
    {
      "Aspect": " حکومت " (government),
      "Opinion": " دھاندلی " (rigged),
      "Category": "وفاق " (Federal),
      "Polarity": "Negative"
    },
    {
      "Aspect": " حکومت " (government),
      "Opinion": "نامنظور" (disapprove),
      "Category": "وفاق " (Federal),
      "Polarity": "Negative"
    }
]
```


Figure 5: Pattern Mining: Aspect-Sentiment Labels Division Based on Identified Phrases-Similar color shows single pattern

"Urdu Tweet": " آئی ایم ایف کی ہدایت پر کٹھ پتلی وزیراعظم عمران نیازی کا بجٹ پاکستان کے غریبوں کے خلاف اعلان جنگ ہے لعنت ایسی تبدیلی پر کامریڈ افتخار"

Translated Tweet: The puppet Prime Minister Imran Niazi's budget on the instructions of IMF is a declaration of war against the poor of Pakistan. Curse such a change, Comrade Iftikhar.


```
"entries": [
    {
      "aspect": "وزیراعظم عمران نیازی کا بجٹ" (Budget of Prime Minister Imran Niazi),
      "sentiment": "غریبوں کے خلاف" (against the poor),
      "category": "بجٹ" (Budget),
      "polarity": "Negative"
    },
    {
      "aspect": "وزیراعظم عمران نیازی کا بجٹ" (Budget of Prime Minister Imran Niazi),
      "sentiment": "لعنت" (Curse),
      "category": "بجٹ" (Budget),
      "polarity": "Negative"
    }
  ]
```


Figure 6: Unique Labels obtained by bidirectional Window-based Strategy

Aspect Categories	**Train**					**Test**				
	Positive	Negative	Neutral	Total	P(%)	Positive	Negative	Neutral	Total	P(%)
Education	42	78	20	139	10%	9	12	2	47	10%
Agriculture	5	10	3	18	1%	1	1	0	6	1%
Federal	246	459	114	820	57%	52	68	14	274	57%
Media	2	4	1	7	0%	1	2	0	3	1%
Economy	13	23	6	42	3%	3	3	1	14	3%
Provincial	18	34	8	60	4%	4	5	1	21	4%
Political party	11	21	5	37	3%	2	3	1	13	3%
Health	38	70	18	125	9%	8	10	2	42	9%
Project	4	7	2	13	1%	1	1	0	5	1%
Social Welfare Programs	20	36	9	65	5%	4	5	1	22	5%
Leadership	9	16	4	29	2%	2	3	0	10	2%
Defense	19	35	9	63	4%	4	5	1	21	4%
Religion	2	4	1	8	1%	1	1	0	3	1%
Miscellaneous	2	3	1	4	0%	0	1	0	2	0%
Total	429	800	201	1430	100%	91	120	24	483	100%
Percentage(%)	30%	56%	14%	75%	–	29%	59%	12%	25%	–

Table 7: Polarity-Specific Aspects Categories Distribution in Train-Test Split

Task	Epoch	Gold-Standard		Bronze-Standard		Silver-Standard	
		Acc.	**macro-F1**	**Acc.**	**macro-F1**	**Acc.**	**macro-F1**
Aspect Category Detection (ACD)	10	0.660	0.681	0.437	0.417	0.449	0.439
	50	0.711	0.722	0.505	0.526	0.580	0.526
	100	0.732	0.733	0.528	0.506	0.596	0.545
	300	0.717	0.723	0.562	0.525	0.590	0.549
Aspect Category Sentiment (ACS)	10	0.669	0.802	0.531	0.456	0.524	0.536
	50	0.687	0.877	0.566	0.487	0.531	0.590
	100	0.706	0.877	0.571	0.493	0.556	0.571
	300	0.706	0.877	0.575	0.494	0.524	0.590
Aspect Sentiment Classification (ASC)	10	0.575	0.296	0.561	0.243	0.582	0.258
	50	0.577	0.304	0.577	0.304	0.583	0.281
	100	0.578	0.318	0.581	0.313	0.589	0.302
	300	0.583	0.321	0.583	0.318	0.591	0.318

Table 8: LSTM Average Results for five-runs on ABSA Tasks