Dataset-Agnostic Stance Detection via Cyclic Multi-Domain Pretraining

Divesh Basina
School of Computing and Augmented Intelligence
Arizona State University
Email: dbasina@asu.edu

***Abstract*—Aspect-based stance detection (ABSD) provides a fine-grained lens on public discourse by inferring polarity toward specific entities or aspects rather than entire documents. While recent datasets have advanced the field, they also expose strong domain and annotation biases that limit cross-dataset generalization. Building on the *Accrue and Reuse Knowledge (ARK)* paradigm of cyclic multi-task pretraining [1], we adapt a teacher–student foundation model to text, coupling a Transformer encoder (BERT) [2] with dataset-specific classification heads and an exponential-moving-average (EMA) teacher. Our approach integrates supervision across heterogeneous ABSD datasets without label harmonization and leverages an aspect-masking strategy to mitigate aspect label biases. Empirically, the proposed framework improves out-of-domain generalization and reduces dataset-specific bias relative to individually trained baselines, highlighting the benefits of multi-domain cyclic pretraining for stance detection.**

Code available at: **https://github.com/dbasina/Ark_Stance_Detection.**

Keywords: Stance Detection; Sentiment Analysis; Aspect-Based Sentiment/ Stance; Foundation Models; Cyclic Pretraining; ARK; Masked ABSA; NLP

I. Introduction

Stance detection seeks to identify an author's position (*pro*, *anti*, or *neutral*) toward a target. Compared to sentence-level sentiment analysis, stance detection yields more actionable signals by conditioning polarity on explicit aspects or entities. Consider the review: "This is the most delicious pizza I've ever had; the crust was perfectly crisp and the basil was fresh." A global sentiment label (*positive*) misses important structure: the stance toward *crust*, *basil*, and *service* can diverge. Aspect-based stance detection (ABSD) therefore enables fine-grained reasoning by predicting stance with respect to each mentioned aspect.

A major bottleneck in ABSD research is the scarcity of high-quality, large-scale stance-annotated datasets. Manual stance annotation requires domain expertise, is expensive, and often suffers from annotator disagreement, especially for politically sensitive or ideologically nuanced topics. To address this, recent work such as the Masked-ABSA framework [3] introduced a novel strategy for producing weakly supervised stance datasets. These datasets—covering U.S. political and race-related discourse—are constructed by mapping Twitter users to ideological camps and assigning stance labels based on expert-defined agreements between camps and aspect terms. Aspect mentions are masked with a `[MASK]` token, ensuring models do not learn stance merely from lexical associations. While this approach enables scalable dataset creation, it also introduces challenges: (i) many aspect labels become heavily imbalanced due to polarized ideological priors, (ii) taxonomies differ across domains, and (iii) weak supervision may inject annotation noise.

Beyond these masked-ABSA datasets, other stance detection resources across domains (e.g., product reviews, public health discourse, social movements) adopt widely varying annotation protocols. Some provide binary stances, others ternary labels; some provide explicit aspect lists, while others offer only masked spans or no aspect annotations at all. The resulting heterogeneity makes it difficult to jointly train models across datasets, and most prior work resorts to training domain-specific classifiers. Yet high-quality stance detection systems require the ability to generalize across domains, topics, and annotation styles.

To address this, we seek a framework that can *unify knowledge across heterogeneous stance datasets* without requiring label harmonization or shared taxonomies. Inspired by the *Accrue and Reuse Knowledge (ARK)* foundation model framework [1], originally developed for multi-dataset medical imaging, we adapt ARK to the NLP setting for stance detection. In medical imaging, ARK employs a teacher–student architecture with exponential moving average (EMA) updates and dataset-specific classification heads, enabling a single model to integrate supervision from tasks with incompatible label spaces (e.g., different disease taxonomies or annotation formats). Its cyclic pretraining protocol helps the model gradually accumulate generalizable features while maintaining stability during transfer across datasets.

In this work, we propose *ARK–NLP*, a text-adapted version of ARK tailored for ABSD. Our method preserves the key properties of ARK—dataset-specific output heads, a slowly updated EMA teacher, and cyclic multi-domain pretraining—while introducing two stance-specific innovations. First, we integrate the Masked-ABSA aspect-masking strategy [3] to reduce shortcut learning and encourage models to infer stance from context rather than lexical cues. Second, we show that ARK's architecture naturally accommodates datasets with different stance taxonomies (binary, ternary), missing labels, or only masked aspect terms. The result is a unified stance detection model capable of leveraging diverse, heterogeneous datasets without requiring explicit label or taxonomy align-

ment.

Our contributions are: leftmargin=*, topsep=2pt

- A **text-adapted ARK** framework for ABSD that supports heterogeneous label spaces with dataset-specific output heads.
- A **cyclic multi-domain pretraining** protocol that accrues and reuses knowledge across datasets while controlling target leakage via *aspect masking*.
- An **evaluation protocol** emphasizing cross-domain generalization (train on multiple datasets, test on held-out domains) with AUROC for imbalance robustness.

II. Related Work

a) Sentiment Analysis and ABSD.: Classic sentiment analysis assigns a single polarity to text; ABSD conditions polarity on aspects, enabling multi-target reasoning. Recent work explores template-based prompting and masking to limit lexical leakage and improve zero-shot generalization (e.g., [3]).

b) Transformer Encoders.: BERT [2] and successors provide strong contextual representations for classification and sequence labeling, with the `[CLS]` token commonly used for sentence-level tasks. Encoder-decoder models (e.g., T5) can phrase stance prediction as sequence-to-sequence generation [4].

c) Foundation Models and Cyclic Pretraining.: The ARK paradigm [1] proposes a teacher–student framework with EMA updates and dataset-specific heads to integrate heterogeneous supervision without label harmonization. We adapt these principles to text for ABSD.

III. Datasets

We evaluate on the *Masked ABSA* corpora introduced by [3], which comprise multiple domains (e.g., politics, racial relations) with aspect inventories curated to support stance analysis. The datasets are weakly labelled by mapping Twitter users to ideological "camps," then assigning stance labels based on expert-defined priors for how each camp aligns with specific aspect terms. Aspect spans in the text are replaced with a `[MASK]` token to prevent stance prediction via lexical shortcuts.

A. Annotation Schema

The Masked ABSA framework produces datasets with heterogeneous stance label spaces.

- **U.S. Race Relations**: binary stance labels (*pro*, *anti*).
- **U.S. Politics**: ternary stance labels (*pro*, *neutral*, *anti*).

Due to the camp-based weak supervision, aspect-level distributions are frequently dominated by a single stance class (e.g., strongly polarized or entirely one-sided), with only a small fraction of aspects exhibiting "balanced" distributions.

B. U.S. Race Relations Dataset

The U.S. Race Relations dataset contains:

- **Samples**: 87,359
- **Aspect terms**: 239
- **Pro labels**: 14,340
- **Anti labels**: 73,019
- **Weakly balanced aspects**: 24

The distribution is highly imbalanced, with most aspect terms overwhelmingly skewed toward the anti stance. Fig. 1 visualizes the top 25 most polarized aspects, showing the difference between pro and anti label counts. A small number of aspects have mixed stance distributions (Fig. 2), but these constitute a minority. The dataset-level clustering of points near the axes in Fig. 3 further highlights extreme polarization: most aspects accumulate many more anti than pro labels, with very few aspects appearing near the diagonal.

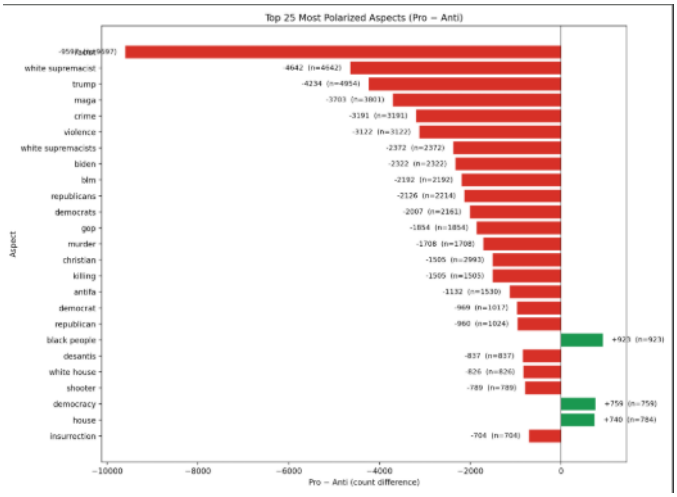

Fig. 1. Top 25 most polarized aspects (Pro – Anti) in the U.S. Race Relations dataset.

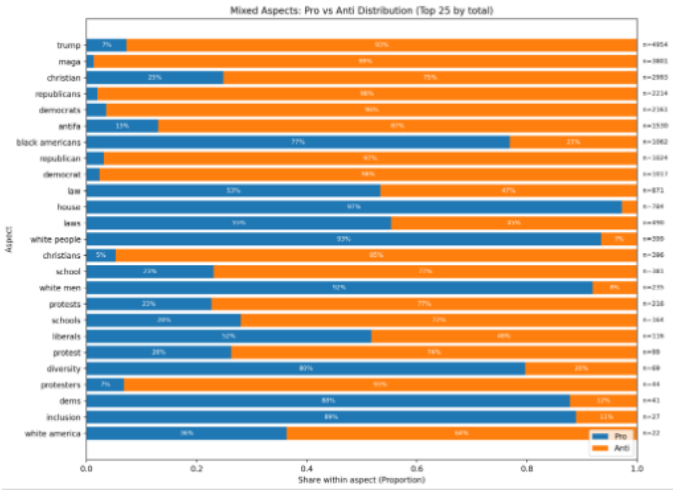

Fig. 2. Pro vs. Anti proportion for the 25 aspects with the largest total counts.

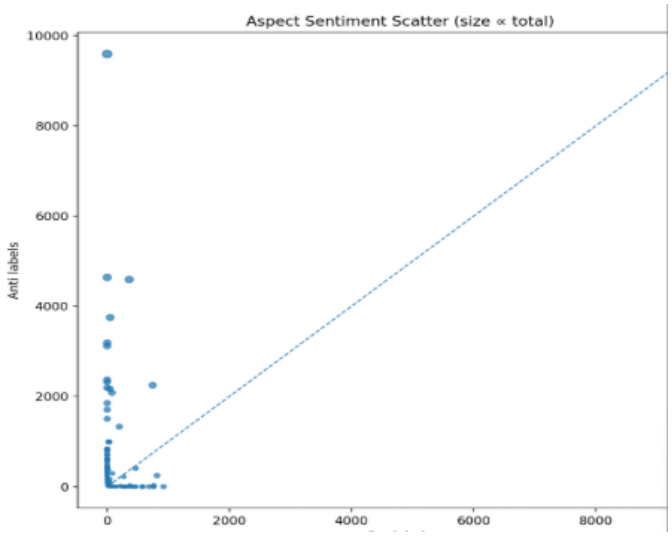

Fig. 3. Scatter plot of aspect-level Pro vs. Anti label counts (marker size indicates total samples).

C. Splits and Preprocessing

We adopt official train/validation/test splits when available and create stratified splits otherwise. Text is minimally normalized: URLs and usernames are removed, casing is preserved where meaningful, and aspect spans are replaced by a `[MASK]` token following [3].

IV. Methodology and Architecture

A. Baselines

We establish strong baselines with **BERT-base** [2] fine-tuning using the `[CLS]` embedding and a linear classifier. The models were initialized using English-Wikipedia and Book corpus pre-trained weights and finetuned on individual Masked ABSA [3] datasets. Two baselines utilizing AUC were established using this method for the U.S race relations and the U.S political datasets.

B. Text-Adapted ARK (ARK–NLP)

Let $\mathcal{D} = \{D_1, \ldots, D_K\}$ denote ABSD datasets with heterogeneous label sets $\mathcal{Y}_k$. ARK–NLP comprises a *student* encoder E_θ and an *EMA teacher* encoder $E_{\bar{\theta}}$, which share the **same architecture and initialization**. In the original ARK framework for medical imaging, both models use identical Swin-b backbones [1], [5]. For our NLP adaptation, we replace this backbone with a **BERT encoder** [2], allowing the teacher and student to operate directly on masked textual inputs. A shared projector g_ϕ maps encoder outputs into a common representation space, and dataset-specific classification heads h_k handle the heterogeneous label sets found across ABSD datasets.

For an input x with masked aspect, the student and teacher produce embeddings $z = g_\phi(E_\theta(x))$ and $\bar{z} = g_\phi(E_{\bar{\theta}}(x))$. The loss for dataset k is

$$\mathcal{L}_k = \underbrace{\ell_{\text{cls}}\big(h_k(z), y\big)}_{\text{dataset-specific classification}} + \lambda \underbrace{\|\text{sg}[\bar{z}] - z\|_2^2}_{\text{consistency}}, \qquad (1)$$

where $\text{sg}[\cdot]$ stops gradients through the teacher and λ balances the consistency regularization. After each dataset cycle, the teacher parameters are updated via EMA: $\bar{\theta} \leftarrow \alpha\,\bar{\theta} + (1-\alpha)\,\theta$.

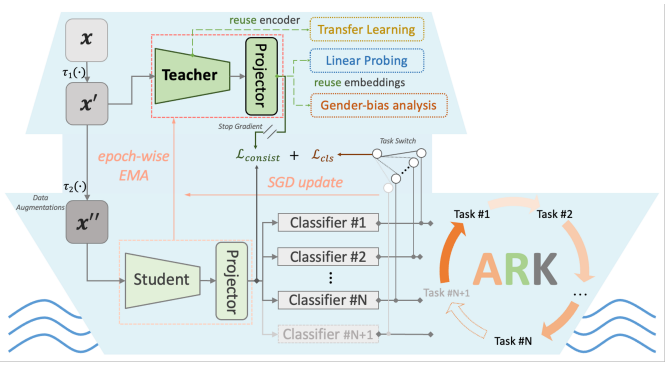

Fig. 4. ARK foundation model architecture [1]. Both the student and teacher share the same backbone architecture; in ARK–NLP this backbone is replaced with a BERT encoder. Cyclic task switching and EMA updates allow the teacher to accumulate knowledge across heterogeneous datasets without label harmonization.

a) Cyclic schedule.: Each *cycle* visits all datasets once (one epoch per dataset). The teacher stabilizes supervision and accrues cross-domain knowledge that benefits subsequent datasets.

b) Label heterogeneity.: Heads h_k match $|\mathcal{Y}_k|$ and can implement binary or ternary softmax. This avoids label harmonization and preserves dataset semantics.

C. Training Details

Tokenization uses the respective model's WordPiece/SentencePiece vocabulary with a maximum length of L (default $L = 256$). We optimize with AdamW, linear warmup over the first 10% of steps, cosine decay, batch size B, and EMA decay $\alpha \in [0.95, 0.999]$.

V. Evaluation and Results

Stance detection datasets—particularly the Masked ABSA race and political corpora [3]—exhibit substantial class imbalance, with many aspects dominated by a single stance class. In such settings, accuracy alone can be misleading: a classifier that predicts only the majority class may achieve high accuracy despite poor discriminatory power. To mitigate this, we adopt **Area Under the ROC Curve (AUC)** as our primary evaluation metric. AUC evaluates ranking quality independently of classification thresholds, making it more robust to skewed label distributions and more appropriate for weakly supervised stance datasets.

Although AUC is our primary metric, we also report accuracy to permit direct comparison with prior work, particularly the Masking the Bias baseline [3], which evaluates models using accuracy. Accuracy remains informative for in-domain comparisons but should be interpreted cautiously in the presence of heavy label imbalance.

A. Political Domain Results

Figures 5 and 6 show the epoch-wise accuracy and AUC for the U.S. political dataset. The ARK framework consistently outperforms the BERT baseline in both accuracy and AUC. Notably, the ARK teacher model achieves smoother and more stable performance than the student or baseline, demonstrating the benefits of EMA updates and cross-domain knowledge accumulation.

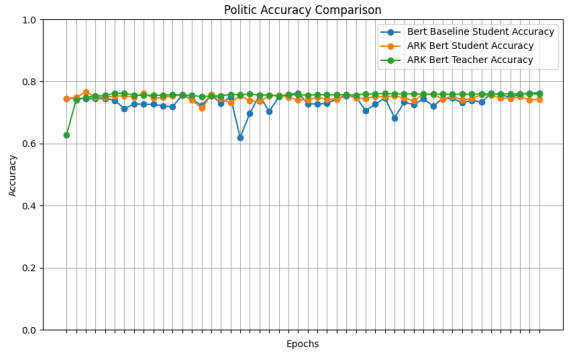

Fig. 5. Comparison of accuracy for the U.S. political dataset across epochs. ARK (student and teacher) outperforms the BERT baseline with improved stability.

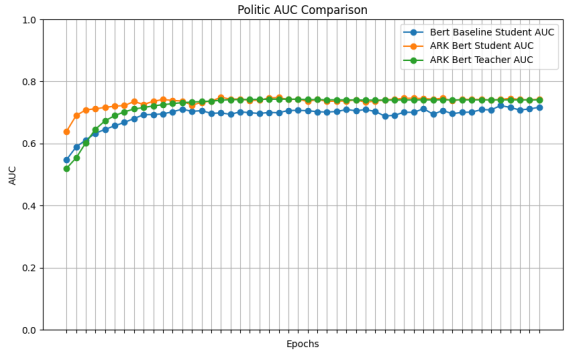

Fig. 6. AUC comparison for the U.S. political dataset. ARK demonstrates superior ranking ability, especially in early epochs, where teacher and student models converge faster than the baseline.

B. Race Domain Results

Figures 7 and 8 illustrate the results for the U.S. race relations dataset. Here, the imbalance is even more severe (73k anti vs. 14k pro). ARK again surpasses the baseline: the teacher model reaches high AUC rapidly and maintains performance throughout training. The baseline BERT model exhibits much more volatile performance, particularly in early epochs.

C. Summary of Findings

Across both domains, ARK consistently provides:

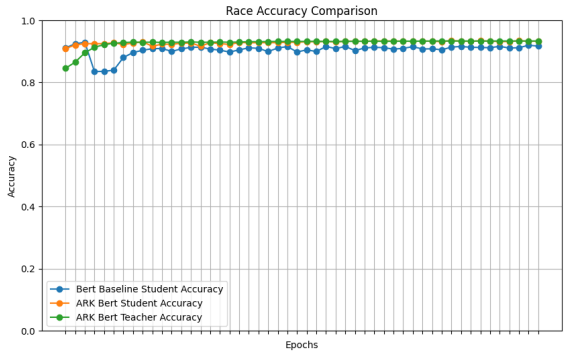

Fig. 7. Accuracy comparison for the U.S. race dataset. ARK models achieve higher and more stable accuracy than the baseline despite the strong label imbalance.

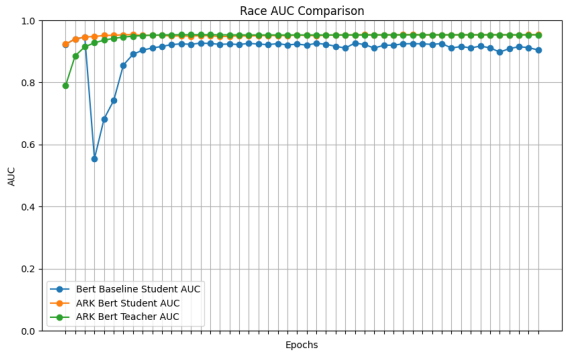

Fig. 8. AUC comparison for the U.S. race dataset. ARK significantly outperforms the baseline and converges more rapidly, highlighting the advantage of cyclic multi-domain pretraining.

- **Higher AUC**, indicating better discriminative ability under label imbalance compared to the individual baselines.
- **More stable training dynamics**, attributed to the EMA teacher and cyclic knowledge reuse.
- **Improved accuracy**, enabling fair comparison with earlier work and confirming general performance gains.

These results validate that dataset-agnostic cyclic pretraining enables models to leverage heterogeneous stance datasets more effectively than independently trained baselines.

VI. Discussion

Cyclic pretraining encourages a curriculum over datasets that stabilizes optimization and enables the teacher encoder to accrue reusable knowledge across domains. Prior work on ARK [1] demonstrates empirically that cyclic training consistently outperforms naïve aggregation—i.e., simply merging all datasets and jointly training on them. While the mathematical foundations for this effect are not yet fully established, the hypothesis is that cyclic task exposure helps maintain domain-specific distinctions while still encouraging transfer of generalizable representations.

Dataset-specific heads further preserve label semantics without requiring cross-dataset label harmonization, allowing ARK to accommodate heterogeneous stance taxonomies. Aspect masking also plays a crucial role: by preventing lexical shortcuts, the model is encouraged to rely on contextual cues, improving generalization to unseen domains and reducing reliance on surface-level correlations.

A. Error Analysis and Sarcasm as a Failure Mode

In our work, we release the full error sets for both datasets. The authors of [3] hypothesized that sarcasm is a likely source of mis-classification in weakly supervised stance datasets. Through an initial qualitative review of our released error sets, we observe evidence consistent with their hypothesis: many incorrect predictions appear to arise from sarcastic or irony-laden text. Sarcasm is particularly problematic because weak labeling assumes that a user's stance aligns with the ideological camp they belong to, while sarcastic language frequently inverts the literal sentiment expressed in the text. Consider the following example from the error set:

Aspect: Trump
User Camp: BLM
Weakly Assigned Stance: Anti
Model Prediction: Pro

> *"Supreme irony is that [MASK] is very likely the least racist president we have ever had. Even Bill Clinton had a couple of black friends that I believe were really his friends. Joe Biden has never had a single person of color as a friend. He must've resented Obama being his boss."*

Although the weak annotation marks this sample as anti (based on the user camp), the text itself expresses a positive stance toward the masked aspect. The model prediction therefore appears incorrect relative to the dataset label, but is arguably semantically accurate when interpreting the text. This illustrates a fundamental weakness of weakly supervised stance datasets: sarcasm may invert the apparent polarity.

B. Establishing Sarcasm as the Primary Error Pattern

To validate that sarcasm is indeed the dominant failure mode, we propose the following methodology:

1) **Baseline sarcasm understanding.** Evaluate a pretrained BERT model (without fine-tuning) on the Reddit Sarcasm Dataset [6] to measure its baseline sarcasm detection capability.
2) **Label transformation of Masked ABSA error set.** Convert error-set samples into sarcasm-labeled examples using the following rule:

Ground Truth	Model Prediction	Sarcasm?
Anti	Pro	Yes
Pro	Anti	Yes
Pro	Pro	No
Anti	Anti	No

Instances where model predictions invert the weakly supervised label are treated as sarcasm candidates.

3) **Fine-tuning with error-set sarcasm signals.** Fine-tune BERT using the transformed Masked ABSA error set and re-evaluate on the Reddit Sarcasm Dataset [6].
4) **Interpretation.** If performance on Reddit Sarcasm improves after fine-tuning, this provides empirical evidence that the Masked ABSA error set contains systematic sarcasm patterns that the stance model fails to capture.

C. Future Work: Integrating Sarcasm Modeling into ARK

If sarcasm is confirmed as a key error source, ARK provides a natural mechanism to incorporate sarcasm detection as an additional task within its multi-head, cyclic pretraining structure. Specifically:

- A new **sarcasm detection head** can be introduced alongside stance heads.
- Sarcasm labels from external datasets or inferred from error patterns can be included in the cyclic training loop.
- The EMA teacher may help propagate sarcasm-aware representations across stance domains.

This integration could potentially resolve misclassifications arising from sarcasm, improving both in-domain accuracy and out-of-domain generalization, especially for datasets that rely on weak supervision.

VII. Conclusion

We present ARK–NLP, a dataset-agnostic stance detection framework that unifies heterogeneous ABSD datasets through cyclic multi-domain pretraining. By integrating dataset-specific heads, an EMA-updated teacher, and aspect masking, the approach improves cross-domain generalization and yields more stable training dynamics than strong BERT baselines. Our error analysis highlights sarcasm as a prominent failure mode in weakly supervised stance datasets, suggesting that stance misclassification often arises from limitations in the underlying annotation process rather than purely from model capacity.

Future work will focus on incorporating explicit sarcasm modeling within the ARK framework—potentially through an additional sarcasm detection head or auxiliary task—to better capture pragmatic nuances in text. More broadly, extending ARK–NLP to handle other latent linguistic factors such as figurative language, domain-specific discourse patterns, and annotator bias offers a promising direction for building more robust and semantically grounded stance detection models.

References

[1] D. Ma, J. Pang, M. B. Gotway, and J. Liang, "Foundation ark: Accruing and reusing knowledge for superior and robust performance," *arXiv preprint arXiv:2310.09507*, 2023. [Online]. Available: https://arxiv.org/abs/2310.09507

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] Y. Lee, Y. M. Çetinkaya, E. Külah, H. Toroslu, and H. Davulcu, "Masking the bias: From echo chambers to large scale aspect-based sentiment analysis," in *Social Networks Analysis and Mining. ASONAM 2024*, ser. Lecture Notes in Computer Science, L. M. Aiello, T. Chakraborty, and S. Gaito, Eds., vol. 15212. Springer, Cham, 2025, pp. 285–300. [Online]. Available: https://doi.org/10.1007/978-3-031-78538-2_19

[4] C. Raffel, N. Shazeer *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.
[5] Z. Liu, Y. Lin, Y. Cao, H. Hu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.
[6] M. Khodak, N. Saunshi, K. Liang, T. Ma, S. Kumar *et al.*, "A large self-annotated corpus for sarcasm," in *LREC*, 2018.