From Consensus to Split Decisions: ABC-Stratified Sentiment in Holocaust Oral Histories

Daban Q. Jaff
Erfurt University, Erfurt, Germany
daban.hamad_ameen@uni-erfurt.de

Abstract

Polarity detection becomes substantially more challenging under domain shift, particularly in heterogeneous, long-form narratives with complex discourse structure, such as Holocaust oral histories. This paper presents a corpus-scale diagnostic study of off-the-shelf sentiment classifiers on long-form Holocaust oral histories, using three pretrained transformer-based polarity classifiers on a corpus of 107,305 utterances and 579,013 sentences. After assembling model outputs, we introduce an agreement-based stability taxonomy (ABC) to stratify inter-model output stability. We report pairwise percent agreement, Cohen's κ, Fleiss' κ, and row-normalized confusion matrices to localize systematic disagreement. As an auxiliary descriptive signal, a T5-based emotion classifier is applied to stratified samples from each agreement stratum to compare emotion distributions across strata. The combination of multi-model label triangulation and the ABC taxonomy provides a cautious, operational framework for characterizing where and how sentiment models diverge in sensitive historical narratives. Inter-model agreement is low to moderate overall and is driven primarily by boundary decisions around neutrality.

Keywords: Holocaust oral history, sentiment analysis, model disagreement, agreement, emotion

1. Introduction

Sentiment analysis (SA) focuses on identifying evaluative meanings, such as polarity or emotional states (Cambria et al., 2017). It is typically framed as a component of opinion mining, where the goal is to extract attitudes toward specific entities or events (Pang and Lee, 2008; Liu, 2012). Methodologically, SA has evolved from lexicon-based (Taboada et al., 2011) and rule-based (Hutto and Gilbert, 2014) to machine learning (Turney, 2002) and modern transformer architectures (Vaswani et al., 2017; Devlin et al., 2019). The contemporary models leverage deep contextual representations, though their effectiveness remains heavily dependent on the chosen unit of analysis, especially on whether it is a single sentence or an entire document (Pang and Lee, 2008; Liu, 2012).

A major obstacle for applying off-the-shelf SA systems to Holocaust oral histories is domain shift: models trained on different genres (e.g., product reviews or Twitter) face a changed input distribution when applied to long-form historical narratives, which can alter label propensities (Blitzer et al., 2007; Glorot et al., 2011; Pan and Yang, 2010). In Holocaust oral histories, evaluative meaning is often expressed indirectly (through description of the lived experience, stance taking, or moral framing), distributed across multiple sentences, and confounded by reported speech and the verbal reconstruction of experience over time. These characteristics can reduce the density of explicit sentiment cues and make polarity judgments less stable.

This paper studies the resulting phenomenon of model disagreement in Holocaust oral histories. We run three pretrained sentiment models and quantify how strongly they disagree at both sentence and utterance levels. The central methodological goal is neither to identify a single best model nor to estimate sentiment accuracy against human-annotated ground truth, but rather to harness the heterogeneous knowledge and inductive biases of these three systems simultaneously. None of the models were fine-tuned on trauma-related discourse. By treating each classifier as an independent knowledge source shaped by its training distribution, the pipeline is designed to expose genuine domain-shift behavior. Moreover, model confidences are not calibrated and are not directly comparable across architectures or training regimes; they are used here only as within-model heuristics and descriptive proxies.

2. Related Work

SA is known to degrade under domain shift and in long-form narrative settings with domain-specific linguistic phenomena. Early work shows substantial performance drops when sentiment models are transferred across domains (Blitzer et al., 2007), and transfer-learning surveys attribute this to distribution mismatch between source and target data (Pan and Yang, 2010). Domain variation in vocabulary and expression is therefore a core obstacle for opinion mining (Liu, 2012), motivating domain-adaptation approaches that jointly model diverse domains (Barnes et al., 2018).

Sentiment inference also depends on task definition and textual structure. Unit choice matters because document-level sentiment is not simply an average of sentence sentiments (Pang and Lee,

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 29–36
11 May 2026.

2008), and aggregation interacts with how opinions are expressed across discourse (Liu, 2012; Kraus and Feuerriegel, 2019). Moreover, subjectivity and attribution can confound polarity when evaluations are embedded in reported or narrative speech (Wilson and Wiebe, 2005; Wiebe et al., 2005). To address these issues in practice, ensembling polarities across models is commonly used to combine complementary sentiment systems, including confidence averaging, stacking, and neural ensembles (Hagen et al., 2015; Troncy et al., 2017; Rouvier, 2017).

There is emerging computational work on sentiment, emotion, and text classification in oral-history interviews. Recent examples include neural sentiment analysis on Holocaust interviews (Blanke et al., 2020), emotion recognition in German oral histories (Gref et al., 2022), geographic emotion modeling of Holocaust testimonies (Ezeani et al., 2024), emotion annotation in the ACT UP Oral History Project (Pessanha et al., 2025), and LLM-based classification of Japanese-American incarceration narratives (Chen et al., 2024; Cherukuri et al., 2025).

Unlike prior works, this study focuses on systematic inter-model disagreement in Holocaust oral histories. To do so, we introduce an ABC agreement taxonomy (with an A split used for polarity-specific analyses) and analyze agreement/divergence across sentence- and utterance-level predictions.

3. Method: Triangulation and ABC Taxonomy

We employ three off-the-shelf pretrained transformer sentiment classifiers: SiEBERT (Hartmann et al., 2023), CardiffNLP Twitter-RoBERTa (Barbieri et al., 2020), and NLPTown (nlptown/bert-base-multilingual-uncased-sentiment). The models are deliberately selected to capture the complementary knowledge encoded in models trained under markedly different regimes: general web text, Twitter-style conversational language, and multilingual product reviews. Moreover, SiEBERT was included intentionally despite its binary label space because its forced polarity decisions make unanimous agreement more conservative and make disagreement with the Neutral-capable models analytically useful under domain shift.

Each utterance u is segmented into sentences using NLTK's `sent_tokenize` (punkt-based). Only minimal normalization (whitespace cleanup) is applied prior to segmentation. Label harmonization follows the upstream pipelines: NLPTown's 1–5 star ratings are mapped to three-way polarity (1–2: Negative, 3: Neutral, 4–5: Positive) (confidence reflects certainty in the winning star rating).

Level	Model	Neg. (%)	Neu. (%)	Pos. (%)
Utterance 107,305				
	NLPTown	48.2	19.2	32.6
	CardiffNLP	11.3	80.4	8.3
	SiEBERT	54.2	—	45.8
Sentence 579,013				
	NLPTown	45.1	24.0	31.0
	CardiffNLP	21.2	69.3	9.4
	SiEBERT	53.9	—	46.1

Table 1: Polarity distributions across models and granularities.

For each sentence, each model outputs a label and an associated confidence score (the model's predicted probability for that label). For each model, we additionally compute an utterance-level aggregated label by assigning each polarity label ℓ the score

$$s(\ell) = \frac{n_\ell}{N} \, \mathrm{meanConf}(\ell),$$

where n_ℓ is the number of sentence-level outputs assigned label ℓ and N is the total number of sentences in the utterance. We then select the label $\arg\max_\ell s(\ell)$. Scripts and analysis materials are publicly available.[1]

Table 1 summarizes the marginal polarity distributions produced by each model at sentence and utterance levels. These distributions reveal substantial label-propensity differences across models. This motivates the agreement diagnostics and ABC stratification introduced below.

3.1. Triangulation

After obtaining sentence-level labels and confidence scores from all three models, and deriving one aggregated utterance-level label per model, we perform cross-model triangulation at two granularities (Table 2), as follows.

Sentence level. A consensus label is obtained by majority vote across the three models. If at least two models agree, that label is selected. In the case of a true three-way split (one Negative, one Neutral, one Positive), the label with the highest model-reported confidence is selected.

Utterance level. For each model separately, we first use the utterance-level aggregation procedure defined above to obtain one aggregated polarity label from that model's sentence-level outputs. Cross-model triangulation at the utterance

[1] https://github.com/dabjaff/ABC-Stratified-Sentiment-in-Holocaust-Oral-Histories.

Level	Metric	Neg. (%)	Neu. (%)	Pos. (%)
Utterance 107,305				
	Count	49,133	18,162	40,009
	Percentage	45.8%	16.9%	37.3%
Sentence 579,013				
	Count	260,727	113,237	205,049
	Percentage	45.0%	19.6%	35.4%

Table 2: Triangulated polarity distribution across granularities.

level then applies majority vote over these three model-specific aggregated labels (equivalently represented as $-1, 0, +1$ for analysis); sentence-level labels and confidence scores are not consulted directly at this stage. In rare triangulation edge cases requiring deterministic tie resolution (95 sentences; 16 utterances), SiEBERT is used as a fallback to ensure reproducible label assignment, not as a claim of superior validity.

3.1.1. Stability Stratification: ABC Taxonomy

While triangulation produces an ensemble label at both granularities, it does not by itself indicate label stability, i.e., the degree of inter-model agreement or disagreement associated with that label. We therefore introduce the ABC taxonomy as a diagnostic stratification framework that tags the outputs of cross-model triangulation by inter-model agreement stability. In this framework, each category represents a different level of consensus across the three-model ensemble, and each sentence and utterance is assigned to one of three agreement categories (see Table 3).

- **Category A (Full Agreement)**: All three models assign the exact same polarity, and the shared polarity is either Positive or Negative, because SiEBERT does not produce a Neutral class.
- **Category B (Partial Agreement)**: Exactly two models agree on the label. This includes (i) cases where the agreement involves Neutral (from CardiffNLP or NLPTown) and (ii) cases where at least two models agree on Positive or Negative.
- **Category C (Maximal Conflict)**: The three models produce three distinct labels (one Negative, one Neutral, one Positive).

Because SiEBERT cannot emit Neutral, Category A should be interpreted as a conservative unanimity subset for non-neutral polarity only, not as a general high-reliability subset over the full three-way label space.

Level	Cat.	Count (n)	Share (%)
U ($N = 107{,}305$)			
	A_{-1}	8,786	8.2
	A_{+1}	6,873	6.4
	B	73,037	68.1
	C	18,609	17.3
S ($N = 579{,}013$)			
	A_{-1}	85,372	14.7
	A_{+1}	39,119	6.8
	B	365,243	63.1
	C	89,279	15.4

Table 3: ABC taxonomy prevalence by granularity.

3.2. Kappa-based Agreement Diagnostics

To quantify agreement under model variation and complement the discrete agreement strata (A/B/C), we report standard inter-rater reliability diagnostics by treating the three sentiment models as raters and each sentence/utterance as an item (Table 4).

Pair	Agr	κ	$N_{\neq 0}$	$\text{Agr}_{\neq 0}$	$\kappa_{\neq 0}$
Utterances (N=107,305)					
SiEBERT–CardiffNLP	17.8	0.088	21,045	90.9	0.816
SiEBERT–NLPTown	61.7	0.350	86,746	76.3	0.518
CardiffNLP–NLPTown	32.3	0.114	18,649	88.5	0.767
Sentences (N=579,013)					
SiEBERT–CardiffNLP	28.0	0.144	177,588	91.2	0.801
SiEBERT–NLPTown	57.5	0.308	440,159	75.6	0.504
CardiffNLP–NLPTown	42.1	0.184	150,755	87.5	0.719

Table 4: Pairwise agreement (Agr.) and κ for 3-way and polarity-only ($N_{\neq 0}$) subsets.

For each model pair, we compute percent agreement (**Agr.**) and Cohen's κ on the shared three-way label space (Negative/Neutral/Positive). Because SiEBERT is binary while CardiffNLP and NLPTown are three-class, Neutral-inclusive agreement and confusion patterns are not directly comparable across all model pairs. We therefore interpret Neutral-boundary effects primarily in the pair where both models can emit Neutral (CardiffNLP and NLPTown). In addition, we exclude any unit labeled Neutral by either model in a given pair and recompute agreement and $\kappa_{\neq 0}$ over {Negative, Positive}.

Finally, we compute Fleiss' κ (three raters) on the three-way space and on the polarity-only subset to summarize overall agreement, and we produce row-normalized confusion matrices for each classifier pair to localize which labels drive disagreement.

3.3. Auxiliary Emotion Profiling

As an auxiliary descriptive signal, we apply a T5-based emotion classifier (`mrm8488/t5-base-`

Level	N (3-way)	Fleiss' κ	$N_{\neq 0}$	Fleiss' $\kappa_{\neq 0}$
Utterance	107,305	0.0535	18,649	0.7835
Sentence	579,013	0.1287	150,755	0.7398

Table 5: Overall three-model agreement (Fleiss' κ) on the full three-way label space and on the polarity-only subset ($N_{\neq 0}$), where units labeled NEUTRAL by CARDIFFNLP or NLPTOWN are excluded.

`finetuned-emotion`; Raffel et al., 2020) to assess whether ABC strata exhibit coherent affective profiles. We use T5 as a descriptive probe because it predicts discrete emotions within a different model family/objective (text-to-text generation), reducing the risk of simply reproducing the same polarity decision boundary. Like the sentiment models, it remains out-of-domain for Holocaust testimony discourse, so its outputs are interpreted descriptively only.

Because Category A splits into two polarity-consistent subsets, we define four groups for affective triangulation: A_{+1} (full tri-model agreement on POSITIVE), A_{-1} (full tri-model agreement on NEGATIVE), and Categories B and C. We randomly sample $2{,}000$ utterances (500 per group) and $4{,}000$ sentences ($1{,}000$ per group), restricting inputs to 10–350 words. For each group, we compute (i) emotion-label distributions at sentence and utterance levels and (ii) mean confidence for the predicted emotion label (reported as a relative proxy, not a calibrated probability). We then compare these profiles across groups to assess whether agreement strata align with distinct affective signatures.

3.4. Data

The pipeline is applied to CORHOH (Jaff, 2025) (see Table 6), and only survivor utterances are analyzed (107,305 utterances, segmented into 579,013 sentences).

4. Results

4.1. Model-wise Polarity Distributions

Before turning to agreement metrics, Table 1 shows that the three models produce sharply different marginal polarity distributions across both granularities, indicating strong label-propensity mismatch under domain shift. In particular, CARDIFFNLP is strongly NEUTRAL-dominant, NLPTOWN is substantially more polar, and SIEBERT is strictly binary; these model-specific output profiles motivate the inter-model diagnostics reported next.

Category	Attribute	Count	%
Gender	Female	270	54.0
	Male	230	46.0
Birth cohort	1890s–1910s	120	24.0
	1920s	320	64.0
	1930s	60	12.0
Top birthplaces	Poland	181	36.2
	Germany	115	23.0
	Other (23 loc.)	204	40.8
Migration era	1930s–1940s	273	54.6
	1950s	111	22.2
	Other/Unknown	116	23.2

Table 6: Corpus demographics and background variables (*N=500*).

4.2. Inter-model Classification

Inter-model classification is examined using pairwise κ-based diagnostics (Table 4), overall three-model agreement via Fleiss' κ (Table 5), and row-normalized confusion matrices (Table 7).

Sentence-level (N=579,013)

	Negative	Neutral	Positive
Negative	35.1	63.3	1.7
Neutral	12.2	80.7	7.1
Positive	8.1	69.4	22.5

Utterance-level (N=107,305)

	Negative	Neutral	Positive
Negative	18.1	80.5	1.3
Neutral	6.5	88.3	5.2
Positive	4.1	75.5	20.4

Table 7: Row-normalized confusion matrices (rows: NLPTown, columns: CardiffNLP). Values are percentages (rounded).

Standard agreement statistics contextualize inter-model classification. On the full three-way label space (NEGATIVE/NEUTRAL/POSITIVE), pairwise percent agreement and Cohen's κ are low to moderate (Table 4), and overall three-model agreement measured by Fleiss' κ is low for both granularities (Table 5). When Neutral labels are included, Fleiss' κ is 0.1287 at sentence level and 0.0535 at utterance level, confirming that inter-model disagreement is dominated by boundary decisions around neutrality rather than outright polarity reversal. When Neutral cases are excluded (polarity-only subset), Fleiss' κ rises sharply to 0.7398 for sentences and 0.7835 for utterances (Table 5), indicating that the models align much more strongly once the task is reduced to Positive-vs.-Negative polarity.

To localize the disagreement mechanism identified above, we highlight the NLPTOWN–CARDIFFNLP row-normalized confusion matrix (NLPTown rows, CardiffNLP columns), because

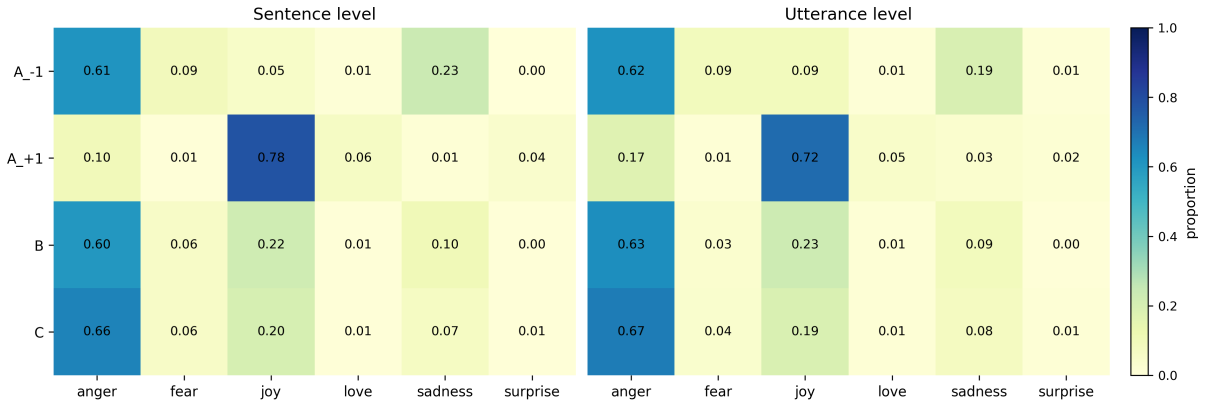

Figure 1: T5 emotion distributions (within-group percentages) across analysis groups (A_{-1}, A_{+1}, B, C) at sentence (left) and utterance (right) levels.

this pair directly exposes Neutral-boundary behavior between the two three-class models. When NLPTown predicts Positive, CardiffNLP predicts Neutral in 69.4% of sentences and 75.5% of utterances; likewise, NLPTown Negative maps to CardiffNLP Neutral in 63.3% (sentences) and 80.5% (utterances) (Table 7). Even when NLPTown predicts Neutral, CardiffNLP remains Neutral in 80.7% of sentences and 88.3% of utterances (Table 7). Conversely, NLPTown's polarity predictions frequently map to CardiffNLP's Neutral label, directly showing that disagreement is concentrated at the Neutral boundary rather than in systematic Positive/Negative reversal.

4.3. Stratification

Agreement patterns are further summarized using the ABC strata (Table 3). Unanimous polarity agreement (Category A) is more common for sentences than utterances, whereas Category B dominates at both granularities and Category C remains non-trivial, indicating persistent disagreement under aggregation. Full agreement (A_{-1}+A_{+1}) covers 21.5% of sentences but only 14.6% of utterances, while B remains the majority at both levels (Table 3). This makes the ABC taxonomy a practical agreement-based stability stratification for sentiment outputs in Holocaust oral histories: Category A isolates a conservative high-consensus subset suitable for polarity-stratified sampling when higher inter-model stability is desired. Accordingly, A_{+1} and A_{-1} can be used as conservative polarity-specific subsets for downstream analysis, while Categories B and C capture the dominant disagreement region. Because A is polarity-skewed ($A_{-1} > A_{+1}$), polarity-stratified sampling from A should preserve this split explicitly rather than treating A as a single homogeneous "high-agreement" set.

4.3.1. Descriptive Emotion Profiles

To profile the ABC agreement strata, we apply a T5-based emotion classifier at both sentence and utterance levels. T5 is also out-of-domain in this setting; it is used only descriptively (not as validation or ground truth) to assess whether the strata exhibit coherent external affective profiles (Figures 1–2).

4.3.2. Affective Distribution Patterns

The emotion heatmaps (Figure 1) show polarity-consistent affective profiles in the high-agreement strata. A_{+1} is dominated by joy (78% sentence; 72% utterance), while A_{-1} is dominated by anger (61–62%) with sadness as a substantial secondary emotion (19–23%). In contrast, B and C display more blended profiles (still anger-forward, with anger at 60–67%, but with larger secondary shares of joy at 19–23% and sadness at 7–10%), consistent with affective heterogeneity that may contribute to cross-model polarity disagreement.

4.3.3. T5 Confidence Proxy

The certainty heatmaps (Figure 2) partially mirror these patterns: A_{+1} shows the highest certainty for joy (0.88 sentence; 0.83 utterance), while A_{-1} shows relatively high certainty for anger (0.73 sentence; 0.74 utterance) and sadness (0.76 at both granularities). In B and C, certainty is generally less concentrated and varies more across labels, consistent with affective ambiguity in long-form oral-history discourse, although some sparse label cells show high confidence (e.g., B–surprise at the sentence level).

5. Conclusion

This study shows that sentiment-classifier disagreement in Holocaust oral histories is not merely a technical nuisance but an analytically meaningful

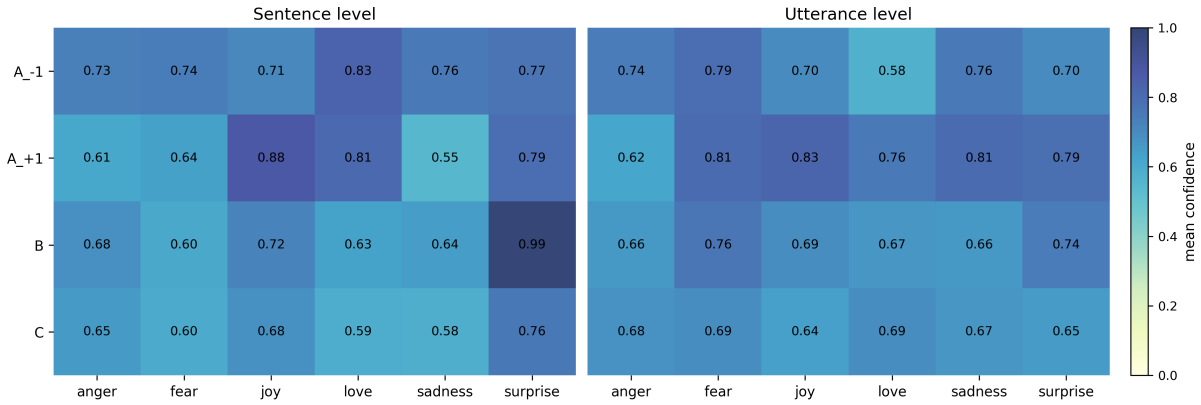

Figure 2: T5 mean confidence (uncalibrated certainty proxy) across analysis groups (A_{-1}, A_{+1}, B, C) at sentence (left) and utterance (right) levels.

signal of domain-shift sensitivity. Rather than converging on a single sentiment profile, off-the-shelf sentiment models produce different polarity distributions at both sentence and utterance levels, with disagreement concentrated especially around the NEUTRAL boundary. However, the present study is diagnostic rather than interpretive.

Triangulation and ABC provide an operational map of model behavior under domain shift: complemented by a T5-based descriptive affective probe, the framework identifies a conservative non-neutral consensus subset for downstream analysis and broader disagreement regions that can be flagged, filtered, or analyzed separately in future work.

Furthermore, these results provide a principled starting point for future work by indicating where future efforts could focus. The utterance-level aggregation rule is an operational heuristic combining within-model label frequency and confidence; alternative aggregation rules such as unweighted majority vote are left for future work. Future extensions may include domain-adaptive fine-tuning to improve polarity detection in Holocaust oral histories, introducing human-annotated evaluation subsets, and extending the ABC framework to other sensitive oral-history corpora.

6. Ethics Statement

This work analyzes publicly available Holocaust oral histories with respect for the survivors, their families, and the historical record. All analyses are strictly descriptive and exploratory. We do not claim that sentiment or emotion labels reflect ground-truth psychological states, nor do we present them as clinical or therapeutic interpretations. The models were used off-the-shelf (without fine-tuning on this corpus) to examine domain-shift behavior in a sensitive historical setting. Our goal is analytical: to identify where and why current NLP tools diverge on Holocaust oral histories.

7. Acknowledgements

I gratefully acknowledge the support of the ***Deutscher Akademischer Austauschdienst (DAAD)*** through a PhD research grant (Grant No. 57645448) for my doctoral studies at **Erfurt University** (Host: **Language and Its Structure**, Prof. Dr. Beate Hampe). I am also grateful to Beate Hampe for reading the manuscript and providing valuable comments. I thank the anonymous reviewers for their valuable comments. Last but certainly not least, this work was carried out using **Prince**, the computational resource of the **Language and Its Structure** professorship, for which I am grateful.

8. References

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tobias Blanke, Michael Bryant, and Mark Hedges. 2020. Understanding memories of the

Holocaust—a new approach to neural networks in the digital humanities. *Digital Scholarship in the Humanities*, 35(1):17–33.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Haihua Chen, Jeonghyun (Annie) Kim, Jiangping Chen, and Aisa Sakata. 2024. Demystifying oral history with natural language processing and data analytics: a case study of the Densho digital collection. *The Electronic Library*, 42(4):643–663.

Komala Subramanyam Cherukuri, Pranav Abishai Moses, Aisa Sakata, Jiangping Chen, and Haihua Chen. 2025. Large language models for oral history understanding with text classification and sentiment analysis. arXiv preprint arXiv:2508.06729.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ian N. Gregory, Tim Cole, Erik Steiner, and Zephyr Frank. 2024. The geography of 'fear', 'sadness', 'anger' and 'joy': Exploring the emotional landscapes in the Holocaust survivors' testimonies. In *Proceedings of the Seventh Workshop on Narrative Extraction From Texts (Text2Story 2024)*, volume 3671 of *CEUR Workshop Proceedings*, pages 93–103.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 513–520. Omnipress.

Michael Gref, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke, and Joachim Köhler. 2022. A study on the ambiguity in human annotation of German oral history interviews for perceived emotion recognition and sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2022–2031, Marseille, France. European Language Resources Association.

Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. 2015. Twitter sentiment detection via ensemble classification using averaged confidence scores. In *Advances in Information Retrieval (ECIR 2015)*, volume 9022 of *Lecture Notes in Computer Science*, pages 741–754. Springer, Cham.

Jochen Hartmann, Mark Heitmann, Christina Siebert, and Bram Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–97.

C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Daban Q. Jaff. 2025. Corhoh: Text corpus of holocaust oral histories. *Data in Brief*, 59:111426.

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:332–343.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Francisca Pessanha, Ian Padovani, Justus van Klaveren, Heysem Kaya, Almila Akdag, and Judith Masthoff. 2025. Listening to oral history: Emotion annotation and recognition in the ACT UP oral history project. In *SUMAC '25: Proceedings of the 7th International Workshop on analySis, Understanding and proMotion of heritAge Contents*, pages 41–50, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena,

Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mickael Rouvier. 2017. LIA at SemEval-2017 task 4: An ensemble of neural networks for sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 760–765, Vancouver, Canada. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Raphaël Troncy, Enrico Palumbo, Efstratios Sygkounas, and Giuseppe Rizzo. 2017. SentiME++ at SemEval-2017 task 4: Stacking state-of-the-art classifiers to enhance sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 648–652, Vancouver, Canada. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan. Association for Computational Linguistics.