Collapsing Distance: The Curse of Ground Truth in Computational Narrative Understanding

Junbo Huang[1,*], Ricardo Usbeck[2,*]

[1]University of Hamburg, Department of Computer Science, Bundesstraße 56b, 20146 Hamburg, Germany
[2]Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany

Abstract

Human label variation (HLV) is slowly gaining attention in natural language processing (NLP) research. It challenges the longstanding machine learning (ML) tradition of having a single ground truth via majority voting label aggregation. It influences all components of an ML pipeline, from data to modeling and evaluation. Albeit its relevance, HLV is rarely discussed in the context of computational narrative understanding. In this position paper, we bridge this gap and provide a distance-focused analysis of the ML pipeline. We examine how the single ground truth assumption shapes the interpretation of distance in annotation reliability, loss computation, and model evaluation metrics, and provide practical implications for the narrative understanding community.

Keywords

Human Label Variation, Computational Narrative Understanding, Distance Computation

1. Introduction

In the 1970s, cognitive psychologist Eleanor Rosch developed prototype theory, suggesting that human interpretation of concepts in words often relies on similarity judgments among these words and our subjective mental representation of concept prototypes [1]. Analogously, this perspective resonates with a recent line of research in machine learning (ML), known as human label variation (HLV) [2, 3, 4]. Plank [3]'s notion of HLV describes, conceptually, the problem of deriving a single ground truth through a majority voting label aggregation strategy in the machine learning pipeline, including data, modeling, and evaluation. The unified notion of HLV considers three types of variations: annotator disagreement, subjectivity, and multiple plausible answers [3]. The longstanding ML pipeline treats these variations as noise, ignoring the pluralistic nature of reading and interpreting narratives, stemming from, e.g., annotators' different socioeconomic or cultural backgrounds. In natural language processing (NLP), computational narrative understanding often deals with extracting information from text and making sense of the extracted information. HLV challenges what is extracted: *Does the ground truth label remove different readings of the same text in the dataset? Does the inter-annotator agreement score measure annotation reliability, or uniformity in opinions? Does the model extract popular opinions?*

Based on Plank [3]'s work, we provide an analysis in this work of how majority voting collapses distance measures, such as Euclidean distance, in each part of the ML pipeline. In other words, we question the measurement of annotation reliability, the expressiveness of the learned narrative embedding space with majority-voted labels, and the validity of evaluation metrics. **We encourage the computational narrative understanding community to refer to HLV when creating datasets, training, and evaluating models**. For example, it is important to have HLV in mind when interpreting annotation reliability scores (e.g., Krippendorff's $alpha$), training losses (e.g., cross entropy, metric learning loss), and task evaluation metrics (e.g., F_1 measure). Equally important is providing unaggregated datasets with meta-information to ensure that variability naturally present in human interpretation is captured in the dataset.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'26 Workshop, Delft (The Netherlands), 29-March-2026

*Corresponding author.
junbo.huang@uni-hamburg.de (J. Huang); ricardo.usbeck@leuphana.de (R. Usbeck)
0000-0002-3192-5896 (J. Huang); 0000-0002-0191-7211 (R. Usbeck)

2. HLV in Computational Narrative Understanding

Being methodologically embedded within the ML framework, computational narrative understanding does not diverge from the longstanding ML pipeline. Typically, computational narrative understanding involves various information extraction tasks [5], such as narrative identification [6, 7], and the extraction of narrative elements, such as events [8, 7], characters [9, 10], or narrative level representations [11, 12]. These tasks commonly require datasets for model training and evaluation. Datasets in computational narrative understanding mostly rely on a majority voting label aggregation strategy. To our knowledge, there are to date only five unaggregated datasets on narratives [13, 6, 14, 15, 7].

Interestingly, **the single ground truth assumption in narrative interpretation loosely resembles the classical literary criticism notion of authorial intent**, which assumes that there is a single, fixed interpretation of narratives [16]. Broadly speaking, narratives are studied as formal structures [17, 18, 19, 20, 21], encompassing structural narrative elements such as events, characters (also known as fabula), time, as well as the telling of these elements (also known as discourse). Alternatively, narratives are viewed as phenomena [22, 23, 24], describing the situated nature of how narratives are created, perceived, and the effects arising from the telling. Additional discussion on narrative definitions is provided in Appendix B. The challenge imposed by the single ground truth assumption is particularly pronounced when attempting to define narratives.

Firstly, defining narratives as formal structures does not make the annotation task less subjective. Annotators disagree not only on subjective tasks, such as hate speech detection [25], but also on tasks that are often considered objective, such as part-of-speech tagging [26]. With text as the modality, these formal structures of narratives can include events, characters, time, or places, referenced by surface-form mentions in text. Annotators may disagree when marking the same event spans, and they may also disagree on character coreference. Previous data annotation work has shown the existence of HLV in narrative identification and extraction tasks. Huang et al. [7] deployed qualitative content analysis as their annotation methodology for narrative identification and extraction. They observed human variation in judging whether a news article primarily discusses the cause of inflation. Antoniak et al. [6] and Mire et al. [14] observed human variation in both sequence classification tasks (i.e., *Does this text contain a story?*) and sequence tagging tasks (i.e., *identifying event spans in text*).

Secondly, defining narratives as phenomena inherently invites multiple plausible readings of the same textual material, but majority voting suppresses minority voices. In this setting, disagreement among annotators should not be interpreted as annotation noise, but rather as reflecting the pluralistic nature of reading and interpreting narratives. Aggregating this variation into a single label collapses this plurality of readings into a single representation, implicitly enforcing a single notion of narrative similarity, violating its own definition.

3. Distance Computation in The ML Pipeline

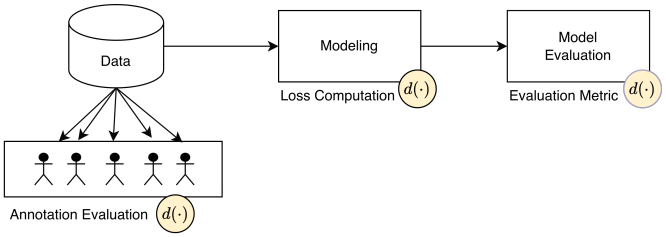

Figure 1: Diagram of the machine learning pipeline. Each component involves distance computation.

Distance is a fundamental concept in machine learning and plays a crucial role in all components of the ML pipeline as in Fig. 1, including data, modeling, and model evaluation. For the data component, measuring annotation reliability (e.g., Krippendorff's $alpha$) requires distance computation between

individual annotations. For the modeling and evaluation component, computing training loss (e.g., metric learning loss) and evaluation metric (e.g., F_1 measure) requires distance computation between model predictions and ground truths. Similarity between two data points is computed with a distance function $d(\cdot)$, formally known as a distance metric[1], such as Euclidean distance. Distance computation for narratives typically involves several stages.

Following a neural approach, we start with two surface-form narrative instances (i.e., raw text inputs). An encoder maps these instances into a latent space, where numerical representations of narratives are used for distance computation. An effective method to learn this latent space is distance metric learning. **Distance metric learning is powerful because one can explicitly define what is considered to be similar or dissimilar**. It learns to structure the latent space such that semantically similar pairs, or *positives*, are pulled together while dissimilar pairs, or *negatives*, are pushed apart. The effectiveness of distance metric learning is exemplified by architectures for learning representation, such as FaceNet [27] and SimCLR [28] in computer vision, and SBERT [29] and SimCSE [30] in NLP, as well as CLIP [31] for both modalities. In the context of HLV, these successful learning architecture learns to represent the majority opinion, because the distance between prediction and ground truth is minimized during training. In practice, aggregation of annotations based on majority voting suppresses minority interpretations, resulting in biased datasets and models. Details on how this biased distance computation influence each part of the ML pipeline will be provided below.

4. Data

The data component includes the construction of narrative corpora (e.g., news articles, microblogs or fictions), annotation and annotation evaluation. We focus on annotation evaluation by means of the computation of inter-annotator agreement (IAA) as a reliability measure. Computing IAA involves applying a distance function over annotated labels among annotators. Majority voting favors a single, fixed ground truth reading of narratives, steering what IAA measures: uniformity of narrative interpretations among annotators of different socioeconomic and cultural backgrounds.

4.1. Formalism

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be a dataset with N data points $x_i \in \mathbb{R}^d$. In a supervised learning setting, dataset $\mathcal{X}$ requires a corresponding set of labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$, which is typically created with data annotation. In a self-supervised learning setting, $\mathcal{Y}$ can be derived from $\mathcal{X}$ itself automatically. Depending on the goal of the task, different teachniques to create input label pairs can be applied. For instance, in masked language modeling, $\mathcal{Y}$ is constructed by masking a ratio of randomly-selected tokens for each instance in $\mathcal{X}$. In next token prediction, $\mathcal{Y}$ is constructed by sequentially shifting instances in $\mathcal{X}$ by one token. In Siamese networks-based classification, $\mathcal{Y}$ is constructed by defining positive (e.g., instances with the same label) and negative pairs (e.g., instances with different labels) from $\mathcal{X}$, where positive pairs denote instances with the same class label. Data annotation is the process of labeling each data point in the dataset $\mathcal{X}$ with K annotators. Given an input data point X_i, K labels are created $\{y_{i1}, y_{i2}, \ldots, y_{iK}\}$. Following the ML tradition, the final label y_i is obtained by aggregating individual annotation by majority voting, leading to a single ground truth for each data point.

4.2. Distance in Reliability Computation

To evaluate data annotation quality, IAA measures the degree of agreement among annotators. Common IAA metrics include Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha [32]. These metrics relies on a predefined distance function $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$ over the unaggregated label space $\mathcal{Y}$. Krippendorff's alpha is defined as

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

[1]Definitions of a distance metric is provided in Appendix A.

where D_o and D_e denote the observed and expected disagreement, respectively, both computed as empirical frequency-weighted averages of pairwise label distances. Consequently, the choice of distance metrics depend on the data type in $\mathcal{Y}$.

For categorical labels, a nominal distance can be used:

$$d(y_a, y_b) = \begin{cases} 0 & \text{if } y_a = y_b \\ 1 & \text{if } y_a \neq y_b \end{cases} \tag{2}$$

For continuous labels, the squared distance function is commonly adopted,

$$d(y_a, y_b) = (y_a - y_b)^2 \tag{3}$$

For graph-structured labels, viewed as a set of triples[2], graph edit distance can be used, defined as

$$d(y_a, y_b) = \min_{(e_1, \ldots e_k) \in \mathcal{P}(y_a, y_b)} \sum_{i=1} c(e_i) \tag{4}$$

where $\mathcal{P}(y_a, y_b)$ denotes the set of edit operations transforming y_a into y_b. Each edit operation e_i can be a node or edge insertion, deletion, or substitution. The cost function $c(e_i)$ assigns a constant weight to each edit, and the total distance corresponds to the minimal number of edits needed to make the two graphs identical.

4.3. Implications for Computational Narrative Understanding

First of all, **a rigorous annotation methodology is required to reduce ambiguity in the annotation guideline and class labels**. Since narrative annotation is itself a task of manual narrative understanding, it involves interpreting both the narratives and the annotation guidelines when making labeling decisions. Ambiguous guidelines and class labels hinder annotation quality and lead to annotation errors [33]. Therefore, pilot studies and rigorous annotation methodologies are essential to ensure a shared understanding among annotators. Additionally, qualitative content analysis [34], a popular qualitative methodology in the humanities, can be applied as an annotation methodology, despite being time-consuming and expensive [7].

Secondly, **distances between distinct discrete labels do not always depict uniformity.** Annotators' perceived boundaries between different discrete labels are fuzzy. Passonneau [35] assigned weights to different types of edit operations for set distance, applied in coreference resolution and document summarization. This same idea can be applied in Eq. 4, where the constant cost $c(e_i)$ signals a uniform distance between each edit. Assigning weights to different label pairs in distance computation for IAA can induce further constraints on the learnt metric space.

Last but not least, **ignoring HLV leads to ambiguous ground truth labels**. Label aggregation via majority voting collapses HLV into a single ground truth label, eliminating information carried by variation in labeling decisions. Treating instances with high and low HLV equivalently introduces unmeasurable noise into the labels and hinders model learning. For example, the popular microblog classification benchmark TweetEval [36] includes an emotion detection subtask with low inter-annotator agreement (Fleiss' kappa < 0.30) [37]. Its post-aggregation strategy assigns the same label to instances with high degree of disagreement, resulting in **ambiguous ground truth** labels. As shown in the ablation study by Huang and Usbeck [38], posts such as "*Binge watching #revenge im obsessed.*" are labeled as `anger` but predicted as `joy`, while "*Don't grieve over things so badly..*" is labeled as `sadness` but predicted as `optimism`.

5. Modeling

The ability of deep learning to automatically learn feature representations has accelerated the evolution of NLP tasks from modeling structural aspects of text (e.g., syntax, entities) to capturing abstract human

[2]An example of a triple with two events and their causal relation: (*War*, *Increases*, *Inflation*).

phenomena (e.g., emotion, irony). This raises an important question: **can ML algorithms, such as distance metric learning, capture such abstract and complex phenomena?** In this section, we examine supervised contrastive learning (SCL) [39] as an instance of distance metric learning. While effective in many settings, its reliance on a similarity measure between predictions and majority-voted labels limits the expressiveness of narrative embedding representations.

5.1. Formalism

Let a neural network be a nonlinear mapping $f : \mathcal{X} \rightarrow \mathcal{H}$, which encodes an input instance $x_i \in \mathcal{X}$ onto an embedding space H. As universal function approximators, neural networks can, in principle, learn arbitrarily complex transformations such that $\mathcal{H}$ captures semantic structure in $\mathcal{X}$, given a well-defined objective function. Typically, as shown in Fig. 2, enforcing a pairwise similarity constraint on $\mathcal{H}$ allows $f(\cdot)$ to learn an embedding space optimized for similarity comparison, as in Siamese networks.

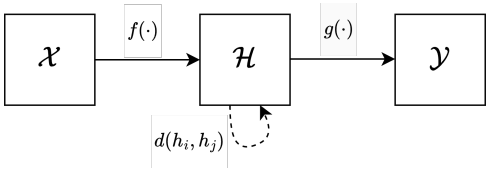

Figure 2: Diagram of distance metric learning components. $\mathcal{H}$ is optimized for similarity comparison.

During training, a linear transformation $g : \mathcal{H} \rightarrow \mathcal{Y}$ is implemented to indicate task-related information in $\mathcal{Y}$. At inference time, g can be discarded and a distance function can be applied on $\mathcal{H}$ to compute similarity between input pairs, such that $d(h_i, h_j)$ is small if $d(y_i, y_j)$ is small, and $d(h_i, h_j)$ is large if $d(y_i, y_j)$ is large.

5.2. Distance in Supervised Contrastive Loss Computation

SCL follows the above principle and encodes positive (x_j) and negative (x_k) instances in relation to an anchor x_i, where x_i shares the same label as x_j, and x_k has a different label. Instead of considering a single positive or negative pair at a time as in SBERT [29], SCL jointly optimizes over multiple positive and negative instances within a batch, encouraging representations of all samples sharing the same label to form compact clusters in $\mathcal{H}$ while separating samples from different classes. The supervised contrastive loss function is defined as:

$$\mathcal{L}_{SCL} = \sum_{i \in \mathcal{X}} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(d(h_i, h_j))}{\sum\limits_{k \in K(i)} \exp(d(h_i, h_k))} \tag{5}$$

, where $P(i)$ and $K(i)$ denote a set of positve and negative pairs and

$$d(h_i, h_j) = \frac{h_i \cdot h_j}{\tau} \tag{6}$$

, where $\tau \in \mathbb{R}^+$ denotes the temperature parameter. $h_i \cdot h_j$ denotes the dot product between h_i and h_j, which resembles cosine similarity as h is L_2-normalized. Minimizing $\mathcal{L}_{SCL}$ indicates that the distance between anchor-positive pairs ($h_i \cdot h_j$) learns to be small and the distance between anchor-negative pairs ($h_i \cdot h_k$) learns to be large.

5.3. Implications for Computational Narrative Understanding

Firstly, **majority voting collapses HLV in $\mathcal{H}$.** SCL learns an embedding space $\mathcal{H}$ that best resembles similarity relations between instances in $\mathcal{X}$, defined by $\mathcal{Y}$. Since $\mathcal{Y}$ is aggregated with majority voting, the latent embedding space $\mathcal{H}$ learns to capture the semantic structure in $\mathcal{X}$ with respect to the dominant

perspective in $\mathcal{Y}$. Therefore, the expressive power of the learnt embedding space $\mathcal{H}$ is limited. In other words, narrative similarity measures distances between dominant interpretations.

Furthermore, to effectively learn a narrative embedding space, **more unaggregated narrative datasets with meta-information are needed.** Existing unaggregated datasets on narrative understanding tasks are scarce [13, 6, 14]. These datasets provide unaggregated individual annotations and meta-information about annotators or narrative instances. Occasionally, meta-annotation is conducted for disagreed items. With limited unaggregated narrative datasets, sample efficiency should be considered in developing learning models. An example is to condition a distance function $d(h_i, h_j \mid Z)$, compared to 6, with meta-information Z on annotators (e.g., age, gender, or country of residence). This allows the model to capture more fine-granular differences between annotators and to generalize better to unseen data.

6. Evaluation

Model evaluation relies on evaluation metrics to provide quantitative signals on how well the model learns. With majority-voted labels, evaluation metrics are not neutral: they are biased towards the dominant perspective. Evaluation metrics compare model output $\hat{\mathcal{Y}}$ against $\mathcal{Y}$. The evaluation result indicates how well models resemble these dominant perspectives.

6.1. Formalism

Given the trained model $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{H} \to \mathcal{Y}$, predictions are obtained either directly in the label space, $\hat{y}_i \in \mathcal{Y}$. An evaluation metric can be viewed as a distance function $d(y, \hat{y})$ which quantifes the discrepancy between labels y and predictions $\hat{y}$.

6.2. Distance in Computation of Evaluation Metrics

For classification, ranking, and retrieval, count-based evaluation metrics are typically accuracy, precision, recall, F_1 measure, or their variants (i.e., Precision@K). They are particularly useful when discrete labels are hard labels. For HLV-driven computational narrative understanding, labels most likely represent a distribution of opinions. Soft labels are often adopted, and therefore count-based metrics cannot be applied.

For regression, evaluation metrics are typically mean squared error (MSE) and mean absolute error (MAE). MSE corresponds to the averaged squared Euclidean distance between predictions $\hat{y}$ and ground truths y, denoted as

$$d(y, \hat{y}) = (y - \hat{y})^2$$

. MAE corresponds to the ℓ_1 distance denoted as

$$d(y, \hat{y}) = |y - \hat{y}|$$

.

For evaluating similarity between texts, BLEU [40] and ROUGE [41] measures are used. They rely on surface-level overlap between generated predictions (y) and ground truth texts ($\hat{y}$). From a distance function perspective, both BLEU and ROUGE induce a distance over the output space by treating lexical overlap as a proxy for semantic similarity, which can be expressed abstractly as

$$d(y, \hat{y}) = 1 - \text{lexical\_overlap}(y, \hat{y}) \tag{7}$$

. Both enforces a single, surface-form-based notion of similarity between two text sequences.

6.3. Implications for Computational Narrative Understanding

Primarily, it is important to evaluate computational narrative understanding models on datasets that are carefully created and annotated. **Evaluating a highly performant model optimized to a biased ground truth against this biased ground truth can lead to misleading conclusions about model quality.** For instance, even if a model achieves an F_1 score of 0.99, such a result may simply indicate that the model has learned to reproduce the dominant interpretation encoded in the aggregated labels, rather than capturing the plurality of opinions.

Related work on evaluation with soft labels proposes comparing model predictions against label distributions using measures such as cross entropy, Jensen–Shannon divergence (a symmetrized version of KL divergence), or other distributional distance measures [42]. These methods treat annotations as distributions rather than descret labels, allowing models to account for the obsered variation in the annotation sample. The aim of label aggregation in this setting is not to collapse disagreement, but to preserve the empirical distribution of annotations. Evaluation then compares two distributions: the model prediction and the soft label. This approach becomes particularly effective when the annotator pool is sufficiently large to reliably estimate the underlying distribution of interpretations.

7. Related Work

Related work at the intersection of HLV and Computational Narrative Understanding is scarce, and most of it focuses on the data side. Narrative datasets with unaggregated labels are being released [13, 6, 14, 15, 7]. Narrative annotation methodology and a variant of Krippendorff's α for graph annotations is proposed in [7].

Additionally, we include related work in HLV that can be adopted for the above narrative datasets. Hovy et al. [43] proposed Multi-Annotator Competence Estimation (MACE), a variational Bayesian inference-based system for estimating the reliability of individual annotators, which can be helpful for creating soft labels. Uma et al. [44] proposed a soft loss, which is essentially cross-entropy loss, and normalized unaggregated labels with a softmax function over the empirical frequency of annotators choosing a specific class.

8. Conclusions

In this work, we addressed the problem of the single ground truth assumption in computational narrative understanding. We provided a distance-focused analysis of the effect of label aggregation via majority voting on the computation of annotation reliability, distance metric learning loss, and evaluation metrics. Finally, we suggested the computational narrative understanding community to have human label variation in mind when interpreting reliability score, training loss and evaluation metrics.

Acknowledgments

The authors acknowledge the financial support by the Hub of Computing and Data Science (HCDS) of University of Hamburg within the Cross-Disciplinary Lab programme.

Declaration on Generative AI

The author(s) have employed ChatGPT for grammar/spelling check and formatting assistance for instantiating latex syntax placeholder for inserting e.g., equations and figures. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] E. Rosch, Cognitive representations of semantic categories, Journal of Experimental Psychology: General 104 (1975) 192–233.

[2] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to consider disagreement in evaluation, in: K. Church, M. Liberman, V. Kordoni (Eds.), Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. URL: https://aclanthology.org/2021.bppf-1.3/. doi:10.18653/v1/2021.bppf-1.3.

[3] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 10671–10682. URL: https://doi.org/10.18653/v1/2022.emnlp-main.731. doi:10.18653/V1/2022.EMNLP-MAIN.731.

[4] F. Cabitza, A. Campagner, V. Basile, Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. doi:10.1609/aaai.v37i6.25840.

[5] A. Piper, R. J. So, D. Bamman, Narrative theory for computational narrative understanding, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 298–311. URL: https://doi.org/10.18653/v1/2021.emnlp-main.26. doi:10.18653/V1/2021.EMNLP-MAIN.26.

[6] M. Antoniak, J. Mire, M. Sap, E. Ash, A. Piper, Where do people tell stories online? story detection across online communities, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 7104–7130. URL: https://doi.org/10.18653/v1/2024.acl-long.383. doi:10.18653/V1/2024.ACL-LONG.383.

[7] J. Huang, M. Weinig, U. Fritsche, R. Usbeck, From variance to invariance: Qualitative content analysis for narrative graph annotation, arXiv preprint arXiv:2603.01930 (2026).

[8] S. M. Mousavi, S. Tanaka, G. Roccabruna, K. Yoshino, S. Nakamura, G. Riccardi, What's new? identifying the unfolding of new events in a narrative, in: N. Akoury, E. Clark, M. Iyyer, S. Chaturvedi, F. Brahman, K. Chandu (Eds.), Proceedings of the 5th Workshop on Narrative Understanding, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1–10. URL: https://aclanthology.org/2023.wnu-1.1/. doi:10.18653/v1/2023.wnu-1.1.

[9] D. Stammbach, M. Antoniak, E. Ash, Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data, in: E. Clark, F. Brahman, M. Iyyer (Eds.), Proceedings of the 4th Workshop of Narrative Understanding (WNU2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 47–56. URL: https://aclanthology.org/2022.wnu-1.6/. doi:10.18653/v1/2022.wnu-1.6.

[10] P. Gervás, G. Méndez, Tagging narrative with propp's character functions using large language models, in: R. Campos, A. M. Jorge, A. Jatowt, S. Bhatia, M. Litvak (Eds.), Proceedings of Text2Story - Seventh Workshop on Narrative Extraction From Texts held in conjunction with the 46th European Conference on Information Retrieval (ECIR 2024), Glasgow, Scotland, UK, March 24, 2024, volume 3671 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 137–148. URL: https://ceur-ws.org/Vol-3671/paper12.pdf.

[11] P. Gervás, J. L. L. Calle, Representing complex relative chronology across narrative levels in movie plots, in: R. Campos, A. M. Jorge, A. Jatowt, S. Bhatia, M. Litvak (Eds.), Proceedings of Text2Story - Seventh Workshop on Narrative Extraction From Texts held in conjunction with the 46th European Conference on Information Retrieval (ECIR 2024), Glasgow, Scotland, UK, March 24, 2024, volume 3671 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 65–76. URL:

https://ceur-ws.org/Vol-3671/paper6.pdf.

[12] P. Gervás, Interpreting narrations of events witnessed: Relying on location data to help place embedded stories, in: R. Campos, A. M. Jorge, A. Jatowt, S. Bhatia, M. Litvak (Eds.), Proceedings of Text2Story - Eighth Workshop on Narrative Extraction From Texts held in conjunction with the 47th European Conference on Information Retrieval (ECIR 2025), Lucca, Italy, April 10, 2025, volume 3964 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 127–135. URL: https://ceur-ws.org/Vol-3964/paper11.pdf.

[13] L. Frermann, J. Li, S. Khanehzar, G. Mikolajczak, Conflicts, villains, resolutions: Towards models of narrative media framing, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 8712–8732. URL: https://doi.org/10.18653/v1/2023.acl-long.486. doi:10.18653/V1/2023.ACL-LONG.486.

[14] J. Mire, M. Antoniak, E. Ash, A. Piper, M. Sap, The empirical variability of narrative perceptions of social media texts, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 19940–19968. URL: https://doi.org/10.18653/v1/2024.emnlp-main.1113. doi:10.18653/V1/2024.EMNLP-MAIN.1113.

[15] S. Hamilton, M. Wilkens, A. Piper, Narrabench: A comprehensive framework for narrative benchmarking, CoRR abs/2510.09869 (2025). URL: https://doi.org/10.48550/arXiv.2510.09869. doi:10.48550/ARXIV.2510.09869. arXiv:2510.09869.

[16] J. Huang, R. Usbeck, Narration as functions: from events to narratives, in: Y. K. Lal, E. Clark, M. Iyyer, S. Chaturvedi, A. Brei, F. Brahman, K. R. Chandu (Eds.), Proceedings of the 6th Workshop on Narrative Understanding, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1–7. URL: https://aclanthology.org/2024.wnu-1.1/. doi:10.18653/v1/2024.wnu-1.1.

[17] V. Propp, Morphology of the Folktale, University of Texas press, 1968.

[18] V. Shklovskiĭ, Literature and Cinematography, Dalkey Archive Press, 2008.

[19] T. Todorov, A. Weinstein, Structural analysis of narrative, NOVEL: A Forum on Fiction 3 (1969) 70–76. URL: http://www.jstor.org/stable/1345003.

[20] R. Barthes, Introduction to the structural analysis of the narrative (1966).

[21] G. Genette, Narrative discourse: An essay in method, volume 3, Cornell University Press, 1980.

[22] M. Foucault, What is an author?, Screen 20 (1979) 13–34. doi:10.1093/screen/20.1.13, first published in 1969.

[23] R. Barthes, The death of the author, in: The Rustle of Language, Collins Publishers, 1986, pp. 49–55.

[24] H. Cixous, K. Cohen, P. Cohen, The laugh of the medusa, Signs 1 (1976) 875–893. URL: http://www.jstor.org/stable/3173239.

[25] X. Yu, E. Blanco, L. Hong, Hate speech and counter speech detection: Conversational context does matter, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5918–5930. URL: https://aclanthology.org/2022.naacl-main.433/. doi:10.18653/v1/2022.naacl-main.433.

[26] M. Sandri, E. Leonardelli, S. Tonelli, E. Jezek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2428–2441. URL: https://aclanthology.org/2023.eacl-main.178/. doi:10.18653/v1/2023.eacl-main.178.

[27] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 815–823. URL: https://doi.org/10.1109/CVPR.2015.7298682. doi:10.1109/CVPR.2015.7298682.

[28] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607. URL: http://proceedings.mlr.press/v119/chen20j.html.

[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[30] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. URL: https://aclanthology.org/2021.emnlp-main.552/. doi:10.18653/v1/2021.emnlp-main.552.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[32] K. Krippendorff, Computing krippendorff's alpha-reliability, 2011.

[33] L. Weber-Genzel, S. Peng, M. de Marneffe, B. Plank, Varierr NLI: separating annotation error from human label variation, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 2256–2269. URL: https://doi.org/10.18653/v1/2024.acl-long.123. doi:10.18653/V1/2024.ACL-LONG.123.

[34] P. Mayring, Qualitative content analysis: theoretical foundation, basic procedures and software solution, Klagenfurt, 2014.

[35] R. J. Passonneau, Computing reliability for coreference annotation, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal, European Language Resources Association, 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/summaries/752.htm.

[36] F. Barbieri, J. Camacho-Collados, L. E. Anke, L. Neves, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 1644–1650. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.148. doi:10.18653/V1/2020.FINDINGS-EMNLP.148.

[37] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, M. Carpuat (Eds.), Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, 2018, pp. 1–17. URL: https://doi.org/10.18653/v1/s18-1001. doi:10.18653/V1/S18-1001.

[38] J. Huang, R. Usbeck, Revisiting supervised contrastive learning for microblog classification, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 15644–15653. URL: https://doi.org/10.18653/v1/2024.emnlp-main.876. doi:10.18653/V1/2024.EMNLP-MAIN.876.

[39] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https:

//proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.
[40] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, ACL, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.
[41] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[42] G. Rizzi, E. Leonardelli, M. Poesio, A. Uma, M. Pavlovic, S. Paun, P. Rosso, E. Fersini, Soft metrics for evaluation with disagreements: an assessment, in: 3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives 2024, European Language Resources Association (ELRA), 2024, pp. 84–94.
[43] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. H. Hovy, Learning whom to trust with MACE, in: L. Vanderwende, H. D. III, K. Kirchhoff (Eds.), Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, The Association for Computational Linguistics, 2013, pp. 1120–1130. URL: https://aclanthology.org/N13-1132/.
[44] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A case for soft loss functions, in: L. Aroyo, E. Simperl (Eds.), Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2020, Hilversum, The Netherlands (virtual), October 25-29, 2020, AAAI Press, 2020, pp. 173–177. URL: https://doi.org/10.1609/hcomp.v8i1.7478. doi:10.1609/HCOMP.V8I1.7478.
[45] W. K. Wimsatt, M. C. Beardsley, et al., The intentional fallacy, Universityof the South, 1946.
[46] E. W. Said, Orientalism, Pantheon Books, New York, 1978.
[47] M.-L. Ryan, Possible Worlds, Artificial Intelligence, and Narrative Theory, Indiana University Press, USA, 1991.

A. Distance Metric

Strictly speaking, a distance metric is a mapping $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_0^+$ over the metric space $\mathcal{M}$, if the following properties for any data points $m_i, m_j, m_k \in \mathcal{M}$ are satisfied:

- Triangular inequality: $d(m_i, m_j) + d(m_j, m_k) \geq d(m_i, m_k)$
- Non-negativity: $d(m_i, m_j) \geq 0$
- Symmetry: $d(m_i, m_j) = d(m_j, m_i)$
- Distinguishability: $d(m_i, m_j) = 0$ if and only if $m_i = m_j$

In ML, the mapping d does not always satisfy all four properties, also named distance measure. Cosine distance is a typical example. It violates two properties: triangular inequality and distinguishability. For practical reasons in ML, the term distance metric and distance measure is often used interchangeably. In this paper, we unify the both notions and use the term distance function.

B. Narratives Definitions

Piper et al. [5] offered a comprehensive overview of the concept of narratives as discussed by scholars across different historical periods and schools of thought. Broadly, narratives are studied either as formal structures (e.g., plots composed of sequences of events and characters, as well as the telling of the plot) or as phenomena (e.g., the situated nature of narrative perception and the effects arising from the telling of events). While narrative formalism views narratives as possessed objects that can be decomposed into formal structures [17, 18], narrative phenomenology emphasizes that narratives are constructed through readers' interpretations, shaped by social, cultural, and other contextual backgrounds [23].

B.1. Narratives as Formal Structures

Abandoning classical literary criticism's notion of authorial intent, which assumes that authors dictate a single, fixed interpretation of narratives, a series of movements emerged to define what constitutes formal narrative structures. In the early 20th century, Russian Formalism, represented by Propp [17], Shklovskiĭ [18], emphasized the fundamental distinction between plot/fabula/historie (the chronological order of events in a narrative) and discourse/syuzhet (the manner in which these events are told). Later developments in New Criticism, exemplified by Wimsatt et al. [45]'s *The Intentional Fallacy*, further reinforced the idea that these formal structures reside within the text itself. In the mid-20th century, literary structuralism generalized insights from literary formalism to posit universal narrative structures. Key structuralist theorists include Todorov and Weinstein [19], Barthes [20], Genette [21].

B.2. Narrative as Phenomenology

Post-structuralist scholars challenged the idea of universal narrative structures and emphasized that meaning emerges from the act of reading narratives. In his essay *The Death of the Author*, Barthes [23] highlighted readers' subjective interpretations, which arise from differing ideological and cultural backgrounds. Similarly, Foucault [22] argued that meaning emerges within systems of discourse, power, and knowledge. Additionally, feminist critics Cixous et al. [24] challenged claims of global narrative structures by foregrounding gender, embodiment, and power in both narrative production and interpretation. Postcolonial critics Said [46] showed how Western narratives construct the east as an object of knowledge and control. Finally, Ryan [47]'s *Possible Worlds, Artificial Intelligence, and Narrative Theory* bridged structural and reader-oriented approaches by providing a formal logical account of modalities and story worlds.

B.3. Takeaway for Computational Narrative Understanding

In the context of computational narrative understanding, regardless of which definition of narrative is adopted by machine learning researchers or engineers, it is crucial to acknowledge a pluralistic view on narrative interpretation. Even narratives of the same formal structures can be read differently. Additionally, narrative representations can encode forms of control by privileging certain voices, viewpoints, and interpretations while marginalizing others. Therefore, the consideration of human label variation in data annotation, modeling, and evaluation is of great importance for developing computational models that more accurately reflect the subjective nature of narrative interpretations, and for developing socially-responsible and inclusive systems.