[Work-in-Progress] Multi-Instance Learning for Social Media-Based Spatiotemporal Public Opinion Analysis

Anonymous authors
Paper under double-blind review

Abstract

This work-in-progress explores applying the framework of multi-instance learning (MIL) (Dietterich et al., 1997; Foulds & Frank, 2010) to spatiotemporally grounded public opinion analysis using social media data. While traditional surveys offer depth and precision, social media provides a scalable, cost-effective complement for real-time tracking of public sentiment (Kruspe et al., 2020; 2021; Zhu et al., 2022). However, weak supervision in data collection often results in a large volume of ambiguous or uninformative posts, complicating both prediction accuracy and interpretability (Mintz et al., 2009; Ratner et al., 2016).

We address these challenges by framing public opinion analysis as a MIL task, where social media posts (instances) are grouped into bags based on shared spatial (e.g., city, region) or temporal (e.g., daily, weekly intervals) attributes. This formulation supports learning both at the bag level (e.g., tracking how opinion shifts over time or across locations) and at the instance level (e.g., identifying specific posts that drive a shift or reflect conflicting viewpoints).

In recently completed but unpublished work, we treated geo-tagged tweets from specific buildings as instances and used non-deep MIL models to infer building functionality. That study demonstrated MIL's ability to handle noisy data and model rare or underrepresented classes (Dax & Werner, 2021; Bai et al., 2025). Building on this, we are developing a more robust MIL framework aimed at public opinion modeling. Drawing on established use cases of MIL in computer vision (e.g., tumor region identification (Barbosa et al., 2024; Vanwinckelen et al., 2016)) and NLP (e.g., document-level sentiment and relation extraction (Christopoulou et al., 2019; Ji et al., 2020; Ma et al., 2021)), we define bags by shared spatiotemporal and demographic features and pursue two core objectives:

- **Implicit Noise Handling**: MIL enables the model to learn directly from weakly labeled data by distinguishing informative from uninformative instances without explicit filtering.
- **Interpretability via Instance Scoring**: By modeling both the bag and its constituent instances, the framework reveals which posts contribute to opinion dynamics or internal disagreement in a region or time window.

While our current work focuses on developing the MIL framework and evaluating its suitability for spatiotemporal opinion modeling and interpretability, we acknowledge that selecting a specific public opinion task, dataset, and labeling strategy is essential for empirical validation. To that end, we are currently surveying existing social media datasets with geotemporal metadata (e.g., Twitter, Reddit) and exploring options for weak labeling. Our aim is to apply this framework to a real-world public opinion case study, enhancing the accountability, transparency, and actionability of models trained on noisy, weakly supervised social media data.

References

Shanshan Bai, Anna Kruspe, and Xiaoxiang Zhu. Generating synthetic oracle datasets to analyze noise impact: A study on building function classification using tweetssemantic parsing of biomedical texts. In *Proceedings of The GeoExT 2025*, volume 3969 of *CEUR Workshop Proceedings*, pp. 28–37. CEUR-WS.org, 2025. URL https://ceur-ws.org/Vol-3969/paper4.pdf.

Dalila Barbosa, Marcos Ferreira, Geraldo Braz Junior, Marta Salgado, and António Cunha. Multiple instance learning in medical images: A systematic review. *IEEE Access*, 12: 78409–78422, 2024. doi: 10.1109/ACCESS.2024.3403538. URL https://ieeexplore.ieee.org/document/10535495.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun 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)*, pp. 4925–4936, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1498. URL https://aclanthology.org/D19-1498/.

Gabriel Dax and Martin Werner. Information-optimal abstaining for reliable classification of building functions. *AGILE GIScience Ser.*, 2:1–10, 2021. ISSN 2700-8150. doi: 10.5194/agile-giss-2-1-2021. URL https://agile-giss.copernicus.org/articles/2/1/2021/. AGILE-GISS.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1–2):31–71, January 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010. doi: 10.1017/S026988890999035X. URL https://www.cambridge.org/core/journals/knowledge-engineering-review/article/review-of-multiinstance-learning-assumptions/0915098C83BF119A377015A45952247A.

Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7012–7023, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.570. URL https://aclanthology.org/2020.emnlp-main.570/.

Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. In Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace (eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.14/.

Anna Kruspe, Jens Kersten, and Friederike Klan. Review article: Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, 21(6):1825–1845, 2021. ISSN 1684-9981. doi: 10.5194/nhess-21-1825-2021. URL https://nhess.copernicus.org/articles/21/1825/2021/nhess-21-1825-2021.pdf.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. SENT: Sentence-level distant relation extraction via negative training. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6201–6213, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.484. URL https://aclanthology.org/2021.acl-long.484/.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li (eds.), *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113/.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3574–3582, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. URL https://dl.acm.org/doi/10.5555/3157382.3157497.

Gitte Vanwinckelen, Vinicius Tragante Do O, Daan Fierens, and Hendrik Blockeel. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Min. Knowl. Discov.*, 30(2):313–341, March 2016. ISSN 1384-5810. doi: 10.1007/s10618-015-0416-z. URL https://doi.org/10.1007/s10618-015-0416-z.

Xiao Xiang Zhu, Yuanyuan Wang, Mrinalini Kochupillai, Martin Werner, Matthias Häberle, Eike Jens Hoffmann, Hannes Taubenböck, Devis Tuia, Alex Levering, Nathan Jacobs, Anna Kruspe, and Karam Abdulahhad. Geoinformation harvesting from social media data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):150–180, 2022. doi: 10.1109/MGRS.2022.3219584. URL https://ieeexplore.ieee.org/document/10049246.