SemLa: A Visual Analysis System for Fine-Grained Text Classification

Munkhtulga Battogtokh[1], **Cosmin Davidescu**[2], **Michael Luck**[1], **Rita Borgo**[1]

[1]King's College London, United Kingdom
[2]ContactEngine, United Kingdom
{munkhtulga.battogtokh, michael.luck, rita.borgo}@kcl.ac.uk
cosmin.davidescu@nice.com

Abstract

Fine-grained text classification requires models to distinguish between many fine-grained classes that are hard to tell apart. However, despite the increased risk of models relying on confounding features and predictions being especially difficult to interpret in this context, existing work on the interpretability of fine-grained text classification is severely limited. Therefore, we introduce our visual analysis system, SemLa, which incorporates novel visualization techniques that are tailored to this challenge. Our evaluation based on case studies and expert feedback shows that SemLa can be a powerful tool for identifying model weaknesses, making decisions about data annotation, and understanding the root cause of errors.

Introduction

Text classification is a standard problem in NLP used for a wide range of tasks like sentiment analysis and intent recognition. An important development in recent years driven by real-life application needs and availability of better models is *fine-grained* text classification, which involves distinguishing between many similar classes (Suresh and Ong 2021). This has exacerbated the interpretability limitation of the state-of-the-art black-box models and has reduced the usability of those models in real-life tasks where trustworthiness is necessary. Unfortunately, existing explanation methods and tools have not kept pace with this development. Therefore, we propose novel visualization techniques for explaining fine-grained text classification and incorporate them into our own visual analysis (VA) system SemLa[1].

The key novelty of SemLa is addressing the aforementioned challenges: 1) many and 2) fine-grained classes. Also, SemLa is unique in allowing the user to safeguard against false impressions from the explanations themselves. Though existing VA frameworks for text classification like DeepNLPVis (Li et al. 2022) and FIND (Lertvittayakumjorn, Specia, and Toni 2020) were shown to be effective for simpler tasks with few well-defined classes (e.g., binary sentiment analysis), they are laborious or simply incompatible to use with many fine-grained classes. Furthermore, previous tools (Spinner et al. 2020; Wallace et al. 2019) incorporate various instance-level explanation techniques such as LIME (Ribeiro, Singh, and Guestrin 2016) or integrated gradients (Sundararajan, Taly, and Yan 2017) as alternatives to each other but do not offer users the ability to spot potential disagreements. In contrast to these previous works, SemLa enables users to intuitively and thoroughly analyze the complex semantic landscape of a model's embedding space while also explaining individual predictions with fine-grained and multi-perspective visualizations.

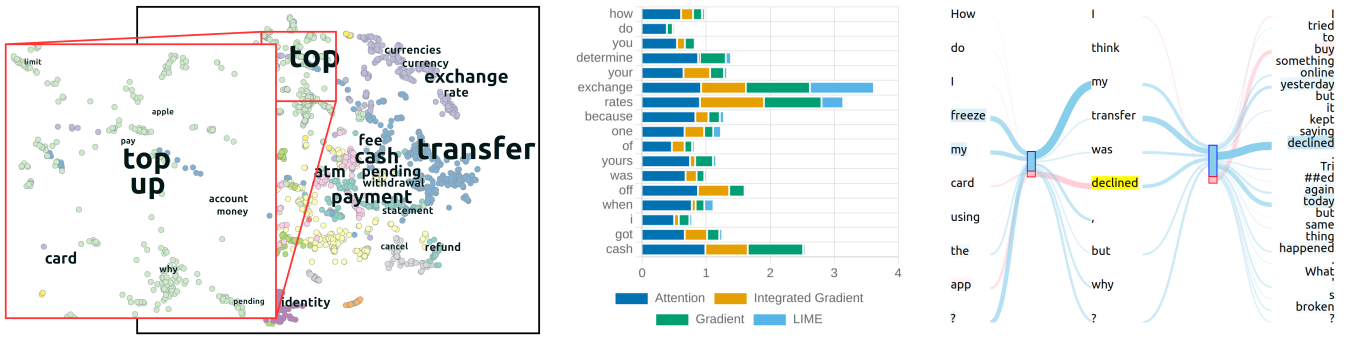
The significance of SemLa lies in its unique focus on fine-grained text classification and trustworthiness. Expert feedback shows that our system is highly useful overall for a variety of use cases. Furthermore, the novel techniques we contribute are highly extensible, effective yet elegantly minimal, and easy to adopt into existing systems.

Overview of SemLa

Interactive Embedding Space Visualization with a Guide Embedding space visualization using dimension reduction is commonly used in previous works to provide global-level insights (El-Assady et al. 2019; Li et al. 2022). However, the high number of classes in fine-grained classification causes the embedding space to be fragmented, which makes understanding such a visualization very difficult. We tackle this challenge by 1) allowing the user to freely navigate the embedding space via a suite of interaction options like zooming, panning, and filtering, and 2) introducing a simple and efficient algorithm, which we use to visualize local patterns that intuitively guide the user as they navigate. Our algorithm computes the locality of a feature w as the minimal hypercube that contains all (or most) occurrences of w.[2] Based on this, we visualize interactive *local words* whose placements show where they occur and sizes show how frequently they occur. We allow the users to filter these by frequency and how big their localities are, which control how localized or widespread the visualized words are. The local words update following user interactions and act as a versatile tool with diverse use cases (e.g., become more fine-grained when a user zooms in like in Fig. 1a, or show the commonalities when a user filters by certain classes, and so on).

[1]A demo is available at: https://munkhtulgab.github.io/SemLa

[2]Detailed description and implementation code of this algorithm as well the whole system is available at https://github.com/MunkhtulgaB/SemLa



(a) Interactive local words (b) Safeguarded explanation (c) Fine-grained contrastive explanation

Figure 1: The key features of SemLa. Best viewed in electronic format.

Safeguarded Explanation Feature importance methods that quantify the importance of input features to model output are predominant in explainable AI (XAI) but there are multiple alternatives with different theoretical backgrounds. It is unclear which one of them should be trusted when they disagree since how to evaluate explanations is still an open question in XAI. Therefore, we support the user in thinking critically about the outputs of a single explanation method by making it simple to see 1) how much importance in total a word receives from different feature importance metrics and 2) how much each metric contributes. This visualization, which we refer to as Visually Integrated Feature Importance (VIFI), is implemented with a stacked bar chart (see Fig. 1b) and results in a novel and elegant visualization design that offers more information while maintaining minimal difference from the familiar bar chart visualization of LIME.

Fine-grained Contrastive Explanation Despite their usefulness, feature importance explanations are too coarse-grained for explaining fine-grained classification in which similar classes share commonalities and differ with subtle distinguishing factors (Jacovi et al. 2021). Therefore, we complement our VIFI visualization with two example-based contrastive explanations that discern the commonalities from the nuances. These visualizations explain why a model chooses its predicted class label (fact) over another similar label (foil) and enable the user to see which features in the input are common for the two classes and which features are unique to the predicted class. Fig. 1c shows an example of the first visualization (the second one is similar but draws links directly from word to word without the rectangular glyphs; see it from our demo). The middle column corresponds to the instance x being explained. The instances in the right and left columns correspond to the fact and foil labels respectively. Sankey-style links indicate the fine-grained relation between x and the two other instances. The thicknesses of the links indicate the strength of the relation whereas the blue and red colors indicate positive and negative relations respectively. By looking at the links on both sides of a word w in x, one can determine whether w is a commonality or a distinguishing factor. To aid this, we highlight the most contrastive word in yellow (e.g., "declined" in Fig. 1c). The rectangular glyphs between the instance pairs express the total strength of the positive and negative relations with size, which is why the blue glyph on the right is larger in Fig. 1c.

Use Cases & Expert Feedback

We demonstrated our system on intent recognition, an especially fine-grained text classification task used in dialogue systems, to three experts from the industry and obtained feedback through an interview involving open-ended questions and a questionnaire with eight close-ended 5-point Likert scale questions. The experts were generally *excited* about the system for how it could be "*immensely valuable*" in model debugging and communicating with clients. The closed-question responses were highly positive with all experts *strongly* agreeing that SemLa was useful overall. They also *strongly* agreed that SemLa helps with understanding individual predictions and identifying sub-clusters within classes. They further noted that our example-based contrastive explanations provide a deeper understanding than existing explanation methods like LIME by clarifying the "*root cause*" of model predictions. Regarding the desirable use cases of identifying model weaknesses, understanding a model at a high level, understanding decision boundaries between classes, and identifying semantic overlap between classes, two out of the three experts *strongly* agreed that the system was helpful and the remaining one expert somewhat agreed. As for how easy the visualizations were to understand, one expert *strongly* agreed that they indeed were intuitive, while the other two somewhat agreed.

Conclusion & Future Work

In this work, we have demonstrated SemLa, our visual analysis system for analyzing and interpreting fine-grained text classification models. It incorporates novel visualization techniques that are simple and intuitive by design yet powerful when they work together interactively. Expert feedback shows that SemLa is effective as a system for various use cases. In future work, we intend to extend our system to new scenarios and make our local word visualization even more informative by experimenting with other feature types (e.g., abstract concepts or named entities) than lexical.

Acknowledgements

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

References

El-Assady, M.; Kehlbeck, R.; Collins, C.; Keim, D.; and Deussen, O. 2019. Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. *IEEE Transactions on Visualization and Computer Graphics*, PP: 1–1.

Jacovi, A.; Swayamdipta, S.; Ravfogel, S.; Elazar, Y.; Choi, Y.; and Goldberg, Y. 2021. Contrastive Explanations for Model Interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1597–1611. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lertvittayakumjorn, P.; Specia, L.; and Toni, F. 2020. FIND: Human-in-the-Loop Debugging Deep Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 332–348. Stroudsburg, PA, USA: Association for Computational Linguistics.

Li, Z.; Wang, X.; Yang, W.; Wu, J.; Zhang, Z.; Liu, Z.; Sun, M.; Zhang, H.; and Liu, S. 2022. A Unified Understanding of Deep NLP Models for Text Classification. *IEEE Transactions on Visualization and Computer Graphics*, 28(12): 4980–4994.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Spinner, T.; Schlegel, U.; Schäfer, H.; and El-Assady, M. 2020. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 1064–1074.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Suresh, V.; and Ong, D. 2021. Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4381–4394. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Wallace, E.; Tuyls, J.; Wang, J.; Subramanian, S.; Gardner, M.; and Singh, S. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations)*, 7–12. Association for Computational Linguistics.