DOI: 10.1111/cgf.15098
Eurographics Conference on Visualization (EuroVis) 2024
W. Aigner, D. Archambault, and R. Bujack
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 3

Visual Analytics for Fine-grained Text Classification Models and Datasets

M. Battogtokh[†1], Y. Xing[1], C. Davidescu[2], A. Abdul-Rahman[1], M. Luck[1], and R. Borgo[1]

[1]King's College London, United Kingdom [2]ContactEngine, United Kingdom

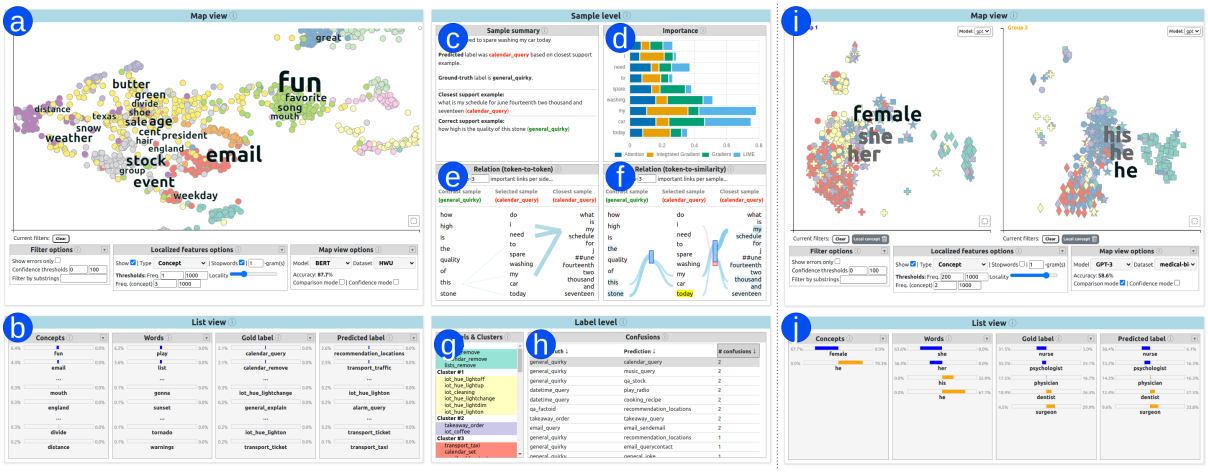


Figure 1: *User interface of SemLa (**Sem**antic **La**ndscape): Map view (a), List view (b), Sample-level consisting of natural language summary (c), Visually Integrated Feature Importance view (d), token-to-token and token-to-similarity relation graphs (e, f), Label-level comprising label-cluster list (g), confusion table (h), and the Map and List views in comparison mode (i and j).*

Abstract

In natural language processing (NLP), text classification tasks are increasingly fine-grained, as datasets are fragmented into a larger number of classes that are more difficult to differentiate from one another. As a consequence, the semantic structures of datasets have become more complex, and model decisions more difficult to explain. Existing tools, suited for coarse-grained classification, falter under these additional challenges. In response to this gap, we worked closely with NLP domain experts in an iterative design-and-evaluation process to characterize and tackle the growing requirements in their workflow of developing fine-grained text classification models. The result of this collaboration is the development of SemLa, a novel Visual Analytics system tailored for 1) dissecting complex semantic structures in a dataset when it is spatialized in model embedding space, and 2) visualizing fine-grained nuances in the meaning of text samples to faithfully explain model reasoning. This paper details the iterative design study and the resulting innovations featured in SemLa. The final design allows contrastive analysis at different levels by unearthing lexical and conceptual patterns including biases and artifacts in data. Expert feedback on our final design and case studies confirm that SemLa is a useful tool for supporting model validation and debugging as well as data annotation.

CCS Concepts

• ***Computing methodologies*** → ***Natural language processing;*** • ***Human-centered computing*** → ***Visual analytics;***

1. Introduction

In natural language processing (NLP), text classification is widely used for language understanding tasks such as sentiment analysis, intent recognition, and occupation classification [SO21, CTG*20, ECCB23]. For these tasks, NLP practitioners commonly adopt deep learning models like CNNs, LSTMs, and pre-trained large language models (LLMs) [LPL*22]. Although these models score high in performance metrics like accuracy, they are well-known to be difficult to interpret and trust. As trustworthiness is crucial to practical applications, many existing visual analytics (VA) tools and explainable AI (XAI) techniques [LWY*22, LST20, RSG16, Vig19] aim to simplify analysis of deep learning models.

† Corresponding author: munkhtulga.battogtokh@kcl.ac.uk

However, text classification tasks have grown more complex in recent years [CTG*20, MGS21] leading to what existing work refers to as fine-grained text classification, characterized by datasets with 1) *numerous*, and 2) semantically *close* labels [SO21]. For example, BANKING77, a fine-grained text classification dataset representative of those in practical applications [CTG*20], has 77 labels and LLMs like GPT-3 Davinci have difficulty distinguishing between the close labels on this dataset [SRL*22]. The many labels lead to complex semantic structure comprising intricately interconnected sample groups, and fine-grained understanding of label meaning is required to distinguish the similar labels [SO21]. Due to these challenges, existing VA tools struggle to meet the requirements of explaining how fine-grained text classification models reach their decisions (Section 2).

Motivated by this gap, we introduce our novel VA system SemLa, which is designed in an iterative design-and-evaluation process in close collaboration with NLP experts from both industry and academia. Building on our previous project on interpretable fine-grained text classification [BLDB24], we started our collaboration on this project in late 2022 and worked together to tackle the challenges in the workflow of developing fine-grained text classification models in practice. We developed and evaluated our system iteratively through multiple rounds of expert feedback each followed by improvements to the system.

Our final system streamlines various tasks in model development workflow, as we demonstrate through case studies and validate via expert feedback. The capabilities of our system include showing discrepancies between ground-truth data distribution and what the model has learned, unearthing lexical and conceptual patterns including biases from data, sample-level explanations that explicitly show fine-grained label semantics, and label-level insights that help understand relationships between different classes or within the same class. Our contributions in this paper are as follows: (i) The design of our visual analytics system SemLa (**Sem**antic **La**ndscape) (and its components) for fine-grained text classification; (ii) Documentation of the iterative design study, including reflections and discussion; and (iii) Detailed evaluation of the final design based on expert feedback and case studies.

2. Related Work

2.1. Visual Analytics for Deep Learning Models in NLP

An extensive body of existing work leverages VA systems and techniques to understand, assess, and debug deep-learning models in various domains [HKPC19, LRBB*23, GZL*21, SSSEA20]. Earlier works in NLP include sample-level (local) explanation techniques e.g., saliency visualization of encodings [LCHJ16], bipartite-graph attention visualization in LLMs [Vig19], and feature importance saliency visualizations based on LIME [RSG16] and SHAP [LL17]. Others enable label-level insights. For example, FeatureInsight [BAL*15] allows users to inspect model errors of a specific label by comparing two groups of samples by the words that are most unique to each group. Similarly, FIND [LST20] uses word clouds to assess what words a model associates with a class.

Most similar to our work is DeepNLPVis [LWY*22], a VA tool for analyzing deep learning models at multiple (corpus, word, and sample) levels. However, since its corpus-level view allows selecting only two labels at a time, understanding all key interrelationships between many labels is infeasible due to combinatorial explosion, which means a large number of classes remains a challenge. The sample-level visualization does not consider fine-grained label semantics, essential for understanding similarities and differences between analogous labels [MGS21,LLLZ21,BLDB24]. In contrast, our system has a corpus-level view that visualizes any number of labels at the same time and a sample-level view that explicates fine-grained label semantics to differentiate similar labels.

Another similar work is LabelVizier, a human-in-the-loop workflow for validating and relabeling text annotations [ZXD*23], in which a surrogate machine learning model, in combination with LIME, is used to explain why the existing labels may have been assigned to text records so as to identify labeling issues such as duplicate or incorrect labels. LabelVizier focuses on annotation errors made by humans and overlooks errors made by the model itself, whereas our system supports the debugging of model errors.

2.2. Text Corpora Visualization

As our work aims for multi-level analysis including corpus level, text corpora visualization is also relevant to us. The most relevant method in this category is topic modeling (the task of identifying high level topics in a corpus), including Non-negative Matrix Factorization (NMF) [LS99] and Latent Dirichlet Allocation (LDA) [BNJ03]. An existing VA system employs LDA-based method for incremental labeling and classification [YTJ*19]. Its known limitations include the need for users to determine the number of topics in the beginning and poor suitability for visualizing many topics at the same time. *Hierarchical* topic modeling, in which topics are identified at multiple levels, is an alternative approach suitable for addressing the above limitations relating to the number of topics. These include Semantic Concept Spaces [EAKC*20], ArchiText [KDEP21] and TopicBubbler [FWC23], which are however designed for analyzing text corpora rather than model outputs.

Some works use topic modeling to analyze *generative* models. A tool for detecting AI-generated text, unCover, employs it to explain the change of topics in (potentially AI-generated) news articles [LBS*23]. Another approach uses it to analyze sentiment in online public discussions about ChatGPT [OUR*23]. Our work focuses on explaining the outputs of a *discriminative* model.

3. Domain Background

In this section, we detail related XAI literature and the workflow of the collaborating NLP experts, essential for anticipating the practical challenges and requirements described in Section 4.

3.1. Explainable AI

Contrastive explanations answer "Why P rather than Q?" [JSR*21, Mil19], a natural question when a model makes an error. Contrastive explanations can reveal the "distinguishing factors" that separate one outcome from the other, and are more suitable for explaining differences between similar labels than simple feature importance estimates [RSG16, STY17]. Related work

[JSR*21, RMP21] often adopts a minimal approach, which aims to identify the most distinguishing factor. However, this approach does not elaborate *why* a word is considered to be distinguishing, as it fails to show how the word is related (or unrelated) to each outcome (label). In an ideal explanation, the labels should be represented by fine-grained aspects such that the explanation shows how (through which aspects) a word is related to a label [BLDB24].

Contrastive explanations are distinct from counterfactual ones, as the former answers "Why outcome P rather than Q?" while the latter concerns alternative antecedents (counterfactuals) that change the outcome from P to another (but not necessarily Q in particular) [SACPF21]. In our setting with many labels (outcomes), the former is more suitable for targeted debugging of specific confusions. Furthermore, answering "Why P rather than Q" does not require the use of counterfactuals. Examples of VA tools that rely heavily on counterfactuals include AdViCE [GHYB21], DECE [CMQ21], and ViCE [GHYB20]. We employ non-counterfactual *example-based* explanations that are suitable for explaining fine-grained labels (see Section 6.4).

Evaluating explanations When incorporating multiple XAI methods into a single VA framework, "trust in the explanation methods themselves" remains a significant issue [SSSEA20], as "How to evaluate explanations?" is still an open question. One important factor for the trustworthiness of explanations is faithfulness, which refers to how accurately an explanation represents the actual reasoning process of a model [JG20]. Existing faithfulness measures include calculating the impact of perturbing or erasing important words [DJR*20,JW19], or guessing back the model predictions based on explanations [LYW19, Ngu18, BLDB24]. However, there is yet no consensus on which method is the best. As long as the question of "How to evaluate explanations?" remains open, any explanation method on their own cannot be fully trusted, especially if multiple competing methods are in disagreement with each other.

3.2. Model development cycle

To understand the problems experts face in their work, we characterized their *model development cycle* based on detailed discussions and preliminary questions in our evaluation rounds (Section 7). The simplified cycle (see Fig. 2) consists of three main phases: data preparation, model preparation, and entering production.

The data preparation phase in turn consists of three stages: data collection, analysis, and annotation. User data is first collected from conversations in the *data collection* stage and then manually analyzed for patterns so as to design the label set in the *data analysis* stage, which involves the experts using visualization techniques like topic modeling and dimension reduction. Each sample is then assigned a label in the last *data annotation* stage of this phase.

In the next *model validation and selection* stage of the model preparation phase, the experts evaluate the models using validation data and standard performance metrics (e.g., accuracy), and select a model. This takes substantial manual effort involving ad-hoc applications of visualization and XAI techniques (e.g., feature importance estimation, dimension reduction) using tools like Jupyter Notebook, Holoviews, and spreadsheets. As the red dotted arrows in Fig. 2 show, the experts often go back to data preparation to do more data collection or *re-annotation* (e.g., to correct mislabels).

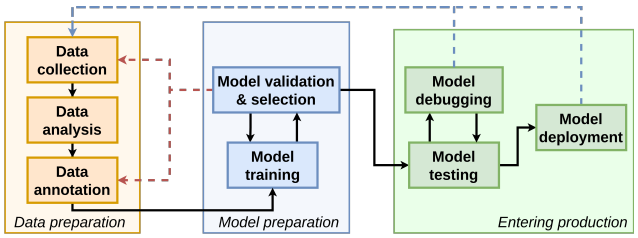


Figure 2: *Model development cycle.*

Finally, once a model is trained, the experts debug the model before deployment, which entails similar manual efforts as during validation. As the blue dotted arrows in Fig. 2 show, it is normal practice to collect data after model debugging or even deployment to continually improve the model.

4. Requirements Analysis

This section establishes the domain requirements and corresponding visual tasks, distilled from analysis of existing visualization literature (Section 2), domain background in XAI (Section 3.1), and iterative discussions with the domain experts and understanding of their workflow (Sections 3.2 and 5).

4.1. Domain requirements

Five key requirements were identified in our investigation:

- R1 Hierarchical understanding of model reasoning. Achieving a multi-level understanding of model predictions and errors, using high-level (general) and low-level (specific) insights.
- R2 Fine-grained explanation of individual predictions. Explaining the fine-grained aspects that distinguish similar labels.
- R3 Revealing model weaknesses. Uncovering weaknesses like spurious features, biases, and root cause of frequent confusions.
- R4 Visualizing semantic characteristics of labels. Understanding the inter-relationships between labels, dissecting the fine-grained concepts, and identifying sub-clusters within a label.
- R5 Safeguarding against false impressions. Faithfully representing model reasoning with minimal built-in assumptions and helping users in critically assessing explanations.

4.2. Tasks

Abstracting tasks is crucial in turning domain requirements into actionable elements. We have identified a set of key tasks based on the domain requirements. To satisfy the requirement of enhancing the hierarchical understanding of model reasoning (R1), the tasks are categorized based on the level of investigation they cater to.

1. **Global-level:**
   - T1 Identifying high-level patterns in encoding space: Tied to the requirement for hierarchical understanding (R1), this task focuses on model validation and identifying systemic patterns.
   - T2 Identifying areas of weakness: Directly addressing the requirement to reveal model weaknesses (R3), this task involves pinpointing weaknesses for model validation and debugging.

T3 Comparing models: Aligned with the requirement for hierarchical understanding (R1), this task involves high-level comparisons between models.

2. **Label-level:**
   T4 Explaining and highlighting decision boundaries between labels: Tied to the requirement for understanding the characteristics of labels (R4), this task is vital for debugging and understanding how labels relate to each other.
   T5 Explaining how a model (mis)understands a certain label: This task supports revealing model weaknesses by identifying misconceptions in label understanding (R3).
   T6 Identifying similarities between label groups: Aligned with the requirement to visualize semantic characteristics of labels (R4), this task aids in debugging and data re-annotation.
   T7 Identifying sub-clusters within one label group: Relating to the same requirement (R4) as the previous task, this task aids in debugging and data re-annotation.
3. **Sample-level:**
   T8 Explaining the importance of each word through multiple metrics: This task, crucial for model debugging, aligns with explaining individual samples (R2) and safeguarding against potential false impressions from a single metric (R5).
   T9 Providing fine-grained contrastive explanations: Supporting the requirement for fine-grained sample-level explanations (R2) and safeguarding against potential false impressions from the coarse-grained feature importance explanation (R5), this task focuses on analysis for debugging and validation.

5. Iterative Design Study

The SemLa VA system, emerged from an intensive one-year design study involving a total of six NLP experts. Table 1 details their diverse expertise and their specific contributions to the design study. Adhering to the structured Nine-Stage Design Study Framework [SMM12], we ensured a methodical construction of the system, which encompassed iterative cycles of design, implementation, and evaluation. In this section, we delineate the conduct of our design study, detailing the chronological progression of the development process as depicted in Fig. 3.

5.1. Precondition Stages

In Nov. 2022, our journey began with an initial consultation with industry NLP expert E1, who oversees the entire development cycle of their organization's AI models as the leader of the research and development team. During the meeting, upon recognizing the significant potential of visual analytics in enhancing their model development workflow, we started our dedicated exploration in this direction. Previously, we had collaborated with the same leading expert on research about interpretability of fine-grained text classification [BLDB24]. Leveraging the domain knowledge from this prior engagement, we expanded our research to visual analytics, guided by thorough and targeted literature review. Motivated by the aim to overcome the challenges posed by existing VA systems, particularly their inadequacies in managing the combinatorial complexity associated with numerous labels (motivation behind R1), we conceptualized an approach. The strategy involved spatializing the samples of all labels within one model embedding space, illustrating the relationships among labels through the spatial distribution and proximity of sample points. This approach can make relationships between similar labels and samples visually discernible, and in combination with a high-level of interactivity, enable users to navigate and explore this space– comprising diverse neighborhoods and areas– at various levels of granularity. Driven by this interactive spatialization approach, we implemented an initial proof-of-concept system, which consisted of the early forms of two key components: the ***Map*** view (Fig 1a) prototype and the *label-cluster list* (Fig 1g) within the ***Label-level*** view (see Section 6.1).

Table 1: *Summary of the interview participants' background.*

#	Domain expertise	Role	Involvement
E1	5 years in NLP	AI team leader	Entire study
E2	7 years in dialogue system	Client liaison & Model development	Evaluation 1 & 2
E3	7 years in NLP	Model development & data analysis	Evaluation 1
E4	3 years in NLP & dialogue system	Configuring generic models to client specifications	Evaluation 2
E5	2 years in NLP	PhD in abusive language detection	Evaluation 2
E6	2 years in medical NLP	PhD in verification	Evaluation 2

5.2. Core Stages: Iteration 1

Discussion: In Jan. 2023, the preliminary version of the system was showcased to the lead expert during a meeting. The consensus was that the spatialization approach indeed showed promise in tackling R1 and could serve as a foundational element in the system. However, it was also recognized that the current design fell short of fully meeting the domain experts' needs, indicating the necessity for additional features. Following an in-depth discussion about the domain-specific challenges and the expert's vision for the tool's functionality, a set of requirements was delineated, primarily focusing on enhancing the explanation of individual predictions (R2) and the identification of model weaknesses (R3).

Co-design & Implementation: Prioritizing the crucial need to expose model weaknesses (R3), we introduced features such as error filtering and a confusion table to the leading expert. We opted for presenting the confusions in a table format (Fig. 1h), which offered compactness and clarity as well as the ability to sort confusions based on frequency through simple interactions with the table header. Following this, we developed the ***Sample-level*** view (Fig. 1c-f), directly addressing the requirements for elucidating individual predictions (R2) and pinpointing model weaknesses (R3). Initially, our focus was on integrating a straightforward feature importance functionality. However, given the imperative to avoid false impressions (R5) and taking into account insights from the domain background (Section 3.1), it became evident that incorporating results from various metrics was essential. Consequently, we crafted our Visually Integrated Feature Importance (VIFI) view (Fig. 1d), which provides a user-friendly interface to investigate the significance of words according to multiple diverse metrics. We soon reacknowledged that while feature importance is informative, it inherently offers a coarse-grained perspective. To address this, we

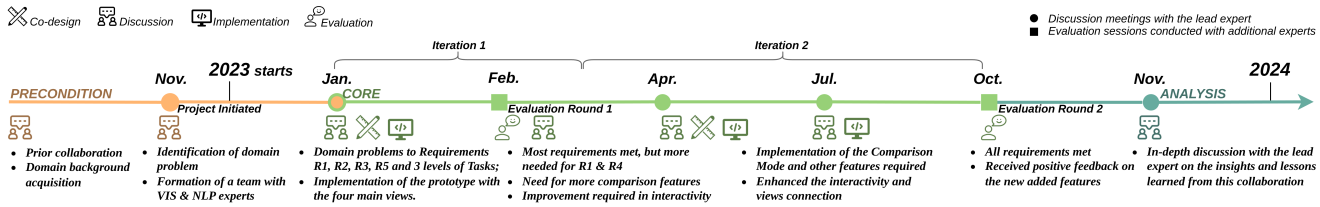


Figure 3: *Timeline of the iterative design process.*

turned our attention to contrastive explanations. Based on our prior insight that explaining fine-grained labels requires making label semantics explicit [BLDB24], we conceptualized an example-based contrastive explanation approach, incorporating three distinct visualizations (Fig. 1c,e,f), detailed in Section 6.4.

Evaluation: The current prototype, designed to meet requirements R2, R3, and R5, was evaluated by three industrial experts E1, E2, and E3 through semi-structured interviews, detailed in Section 7.1. While receiving positive feedback, experts also highlighted areas for enhancement, including the need for 1) a model comparison feature, 2) clearer headers in sample view relation charts, and 3) improved bidirectional and hierarchical sorting of the confusion list.

5.3. Core Stages: Iteration 2

Discussion: After reviewing the feedback so far obtained, we identified two additional key requirements: 1) the need for a hierarchical understanding including high-level model comparisons (R1), and 2) the necessity for more in-depth comparative analysis of label similarities and differences (R4) both between and within groups.

Co-design & Implementation: In response to the feedback from Iteration 1, several improvements were integrated, e.g., headers were added to the relation charts for better clarity and the confusion table received hierarchical sorting capabilities. To meet the requirement of R1 & R4, a new abstraction layer was introduced for local words, enabling the visualization of localized commonsense concepts. Furthermore, a *comparison mode* feature (Fig 1i-j) was implemented, providing the capability to contrast any two groups of samples. This mode was compatible with the system's existing functionalities, offering options such as concept-based filtering, ground-truth vs. prediction comparisons, label differentiation, a lasso tool, and sample-level analysis for longer texts.

Evaluation: In Oct. 2023, upon the completion of the system, a second round of evaluation was initiated. This phase involved five domain experts: two returning from the initial round (E1 & E2), a new member from the same industrial AI team (E4), and two additional evaluators (E5 & E6) with an academic background encompassing NLP, XAI, and VIS. Employing a similar approach as before, semi-structured interviews followed by surveys were conducted with each evaluator. Section 7.2 details the procedure and results of this evaluation. The follow-up survey revealed a unanimous endorsement of the system by the evaluators. Feedback was uniformly positive across all components, validating the fulfillment of all predefined requirements. Particularly, the *comparison mode* garnered strong recognition for its practicality and effectiveness.

5.4. Analysis Stages

Post the second evaluation round, we gathered comprehensive feedback, leading to in-depth discussions and valuable reflections on the lessons learned, which we discuss in Section 9.

6. SemLa: System Description

This section describes our VA system **SemLa** and its components.

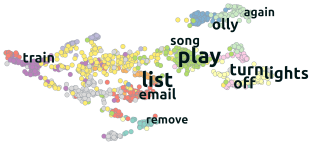
6.1. Overview

The system interface consists of four coordinated views. The **Map** view projects a corpus to a 2D scatter plot using sample embeddings by a selected model and dimension reduction (Fig. 1a). We primarily used t-SNE [MH08] as the dimension reduction method to calculate positions, but the users can also switch to UMAP positions [MHSG18]. These two methods are known to consistently produce high-quality projections across different parameter settings [EMK*21]. For t-SNE, we used the scikit-learn implementation with the parameters perplexity $p = 40$, number of iterations $n = 1000$, and otherwise default parameters. For UMAP, we used the Python UMAP library and default parameters (neighborhood size of 15 and minimum distance of 0.1). In both cases, we used inner product as the distance metric. Each sample is represented by a circle (or another shape at low-level to differentiate labels). Upon hovering over a circle, a tooltip shows the underlying raw text, the ground-truth label, and the predicted label of the associated sample. At the high level, as there are too many labels to visually encode, which often mix and overlap without forming clear clusters [JKA*23], labels are grouped into clusters and the clusters (not the labels) are encoded in the colors of the samples. The Map view dissects the visual space through various features, including zooming, panning, filtering, and a novel interactive *local word* visualization (see Section 6.2), which helps users navigate and understand patterns in different neighborhoods. The **List** view (Fig. 1b) synchronously summarizes the concepts, words, and labels present in the current samples on the Map view as the user interacts with the system. Both views have *comparison modes* (Fig. 1i-j and Fig. 4c) for contrastive analysis, in which the Map view presents two separate scatter plots (showing two different groups of samples, or the same corpus through two different models) and the List view shows which concepts, words, and labels are more likely to be in one group than in the other. The **Sample-level** view explains a single prediction upon the user selecting an individual sample (Fig. 1c-f) through four sub-views (see Sections 6.3 and 6.4). Lastly, the **Label-level** view consists of *label-cluster list* (Fig. 1g), which lists labels grouped by similarity, and *confusion table* (Fig. 1h), which

is a ranked table showing pairs of labels and how frequently they were confused with one another. The user can sort this table by its columns (e.g., confusion frequency) to easily find which labels were most confused with one other. Upon selecting certain labels from the confusion table or label-cluster list, the Map and List views update by showing only samples of those labels.

6.2. Local Words

For semantic structures comprising many neighborhoods with hierarchical relations, a method for identifying patterns at multiple levels is necessary (R1). The closest related work is BERT-based topic modeling [ACT*24], which can not only explain datasets but also models. However, it relies on clustering to identify topics in a top-down manner by grouping samples into clusters and then extracting the top keywords (using class-based *tf-idf* analysis [Gro22]), which has two problems: 1) clustering is computationally expensive and 2) the top-down approach forces the data to be seen through an extra lens (by introducing arbitrary assumptions in the form of hyperparameters, e.g., the number of clusters or their density).



Therefore, we propose our simple and fast Localized Word Clouds (LWC) algorithm (Algorithm 1), which finds patterns directly from model embedding space in a bottom-up manner without relying on clustering (R5). LWC identifies words *localized* to a neighborhood, i.e., occurring only there and nowhere else. It computes the locality of a word as the area enclosing all of its occurrences, allowing users to filter words by locality size. The result of LWC, which we refer to as *local words* (see above), resembles a word cloud but it differs in that LWC results can be overlaid meaningfully on a corpus' 2D projection. Over generating multiple word clouds to explain multiple labels (as in FIND [LST20]), our approach has the advantages of being space efficient and avoiding word repetitions.

The low latency of LWC allows the local words to update as the user zooms, pans, or filters to analyze different sample groups. For example, as shown on the left, when the user zooms into a certain area (the area containing "train" in the previous image), more fine-grained patterns appear (R1). Moreover, patterns within errors (R3) and label groups (R4) can be unearthed by filtering the samples by error or label with filtering options and the Label-level view respectively.

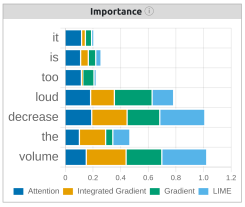
Furthermore, LWC can be applied recursively to extract abstract concepts from local words (see left). Such concepts can assist users in identifying potential biases, spurious features (R3), and hidden relationships between labels (R4). LWC can also be applied to positions with any number of dimensions but applying it to 2D sample projections takes advantage of the dimension reduction algorithms' existing ability to generate high-quality visualization layouts.

Algorithm 1: Localized Word Clouds (LWC)

LWC outputs a set of l *local words* $L = \{w_1, w_2, \ldots, w_l\}$ and their corresponding positions $P_L = \{p_1, p_2, \ldots, p_l\}$ in a space S (of arbitrarily many dimensions) given inputs that include M samples $D = \{x_1, x_2, ..., x_M\}$ and their positions $P_D = \{p_1, p_2, \ldots, p_M\}$ in S. The output words are filtered by their frequency with parameter T and by locality size with function $R(\cdot)$. A function $C(\cdot)$ computes the center of the locality of word w, on which w is to be visualized. Our choices of functions R and C are described in supplementary materials.

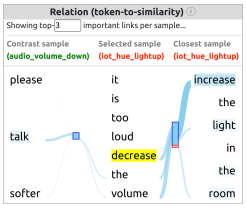
input : $D = \{x_1, x_2, ..., x_M\}$, $P_D = \{p_1, p_2, ..., p_M\}$, R, T, C
output: $L = \{w_1, w_2, ..., w_l\}$, $P_L = \{p_1, p_2, ..., p_l\}$
$L \leftarrow \{\}$;
$P_L \leftarrow \{\}$;
$W \leftarrow$ an empty map;
for $x_m \in D$ **do**
  for $w_j \in x_m$ **do**
    if $w_j \notin W.keys()$ **then**
      $W[w_j] \leftarrow$ an empty list;
    end
    $p \leftarrow p_m \in P_D$;
    $W[w_j].add(p)$;
  end
end
for $w_i \in W.keys()$ **do**
  $(p_1, p_2, ..., p_F) \leftarrow W[w_i]$;
  if $R(p_1, p_2, ..., p_F)$ *and* $F > T$ **then**
    $p_i \leftarrow C(p_1, p_2, ..., p_F)$;
    $L.add(w_i)$;
    $P_L.add(p_i)$;
  end
end
return L, P_L

6.3. Visually Integrated Feature Importance (VIFI)



Our Visually Integrated Feature Importance (VIFI) view employs a stacked bar chart to merge various feature importance metrics into a unified visual representation (Fig. 1d). Each segment of a bar (see left) is allocated to a distinct feature importance metric, which enables users to see the cumulative importance of a word as well as the individual contribution of each metric. VIFI facilitates critical analysis of feature importance and helps users avoid false impressions (R5).

6.4. Example-based contrastive explanations



We propose novel *example-based* contrastive explanations, which explicate why a sample relates more to one label than it does to another (R2, R3) by representing the labels with their respective in-distribution samples exemplifying their fine-grained aspects (R4). SemLa incorporates three such visualizations, each explaining a query sample with respect to two other samples: the closest sample, which shares the same label as the query sample, and a contrast sample, which has a different label. Out of those three, one shows natural language summary (Fig. 1c), and the other two (which we refer to as *relation graphs* collectively) show token-to-token links between the three samples (Fig. 1e and Fig. 4b) and the contribution of each token to the similarities between the three samples (see image directly above and Fig. 1f).

7. Evaluation

In this section, we provide a detailed report of the methodology followed in our evaluation. These include conditions, sample size, data exclusions (if any), any statistical method used in the analysis, and all relevant measures.

Our evaluation involved two structured evaluation rounds (each comprising semi-structured interviews and follow-up surveys) with a total of six experts (four from industry and two from academia). Table 1 describes the profile and involvement of each expert. The first round was in Feb 2022 with three industry experts, assessing the tool's alignment with the primary requirements, followed by the second round in Oct 2023 with five experts (two returning from the previous round), which focused on evaluating the overall usability of the final tool and specifically examining the enhancements made in response to the first round's feedback.

7.1. Evaluation Round 1: Feb. 2022

7.1.1. Protocol

Time and Participants. This round in Feb. 2022 involved three industry dialogue system experts, who work on delivering conversational AI solutions to client companies. Experts E1 and E3 are deeply involved in model development, while E2 has a more client-centric role. A common aspect of their daily responsibilities is explaining model reasoning and weaknesses.

Activities and Duration. The evaluation encompassed three individual semi-structured interviews, each tailored to assess the system at its initial stage. Key components under review were the ***Map***, ***List***, ***Sample-level***, and ***Label-level*** views (Fig. 1). Following a standardized format outlined in Table 2, each interview spanned approximately ninety minutes. Sessions started with a preliminary preparation during which objectives were clarified. A succinct overview of the tool, highlighting its main features and a demonstration using the BANKING77 public dataset [CTG*20] known for its fine-grained intents, was provided. We used the benchmark dataset due to privacy concerns of using actual user data. The demo system incorporates a BERT model [DCLT19], trained via metric-based learning method [CZMX22] for 5-way 1-shot classification, which facilitates explaining predictions via distances (from query sample to a support set comprising samples from five distinct labels, with one sample per label). Subsequently, experts engaged with SemLa, experimenting with its functionalities across four predefined tasks (see supplementary material). The tasks, designed to mirror specific requirements and objectives, provided insights while experts verbalized their thoughts. The evaluation concluded with a reflective survey with open-ended questions, alongside eight five-point Likert scale questions (with options ranging from "strongly disagree" to "strongly agree"). The survey aimed to gather holistic feedback about the system's effectiveness with respect to domain needs, its overall usability, and suggestions for improvements.

7.1.2. Results

SemLa received highly positive feedback. All experts *strongly* agreed on the system's overall usefulness and its ability to clarify individual predictions and identify label sub-clusters (R2). For aspects like identifying model weaknesses (R3), high-level understanding of models, explaining decision boundaries within labels, and identifying semantic overlaps between labels, the majority *strongly* agreed on the system's effectiveness, with one expert *somewhat* agreeing. As for the visualizations' intuitiveness, one expert *strongly* agreed, while the others *somewhat* agreed. A detailed summary of the experts' responses to open-ended questions aligning with these ratings will be presented in the following paragraphs.

Use Cases and System Utility: The experts expressed significant enthusiasm about the system's potential, describing it as "*immensely valuable*" for tasks like model debugging, which involves pinpointing weaknesses and understanding the root causes of errors. Additionally, they highlighted its utility in client communications. One expert noted the system's capability to "*intuitively spot weaknesses*" by enabling users to 1) filter errors using confidence thresholds (referred to as "*top and tail*" the errors), and 2) comprehend these errors for "*targeted intervention*" rather than relying on "*trial-and-error*" approaches for model improvement.

Most Effective Visualizations: The experts varied in their preferences for the system's visualizations. One expert highlighted the *sample-level* visualizations, particularly the contrastive explanations, as "*novel and very useful*." Another expert emphasized the utility of integrating the confusion table with token-to-similarity relations, finding it instrumental in grasping the model's primary errors. The third expert valued the combination of Local Words visualization with the interactive features of the *map*, appreciating its versatility in offering insights at various levels.

System Novelty and Capabilities: Responses to the system's novelty largely focused on its *sample-level* explanations. All experts agreed that the system provides "*deeper insights*" compared to existing tools at their disposal. The two experts (E1 & E3) deeply involved in model development provided specific insights: one remarked they had "*never seen anything like this*" that digs deep into the root causes of errors despite being experienced with explanation and visualization techniques (e.g., LIME, topic modeling); the other noted that, unlike current tools, our system enables a clearer understanding of error causation. Particularly, they highlighted the relation graphs (Fig. 1e-f) that offer contrastive explanations as the most innovative and useful visualizations.

Visualization Understandability and Learnability: Overall, all experts agreed that the visualizations were intuitive and easy to grasp with minimal learning required. They did, however, suggest some enhancements for clarity, such as adding *column headers* to the relation graphs for easier sample identification, and tooltips providing further meta-details about the visualizations (e.g., explanation of value calculations). One expert particularly noted the system's compatibility with their existing workflow, stating it would integrate seamlessly and enhance their work process. They emphasized the system's ability to easily reveal insights that are "*very difficult to find from spreadsheets*" (raw tabular data), thereby significantly improving their workflow experience.

Recommendations for System Enhancement: Experts suggested that our system could be more effective in *facilitating comparisons between different models*, e.g., checkpoints or models trained with different hyperparameters. While acknowledging the system's current capacity to support model comparison by loading

Table 2: *Interview Procedure and Duration.*

Order of Procedure	Activities	Duration
Preliminary Preparation	1) Introductory questioning 2) Tool walkthrough	10-15 min
Task Scenarios	1) Test via predefined tasks	45-60 min
Follow-up Survey	1) Reflection on the tool 2) Likert-scale questions	15-20 min

different models, they recommended features specifically tailored for this task. Suggestions included a sequential time-lapse animation across checkpoints, or a parallel view comparing two models. Other suggestions include the earlier ones like adding tooltips to provide meta-details about the visualizations and implementing column headers in the relation graphs for enhanced clarity.

7.2. Evaluation Round 2: Nov. 2023

SemLa was further refined based on the insights from the first evaluation round, adhering to the iterative process. The usefulness and usability of the tool were then evaluated again with the one-on-one interviews and follow-up surveys.

7.2.1. Protocol

Time and Participants. In Nov. 2023, we conducted the second evaluation round, broadening our reach to more experts from both industry and academia to gather extensive and unbiased feedback about the system's usability across diverse domain backgrounds. Five participants were involved: E1 and E2, who had participated in the previous evaluation, were joined by E4, a colleague from their industrial AI team. E5 and E6 were from academia and both had two years of NLP experience in academic and industrial settings. The evaluators' familiarity with NLP, XAI, and VIS was gauged through background questions in the survey. Each of the five participants had NLP experience, averaging 3.8 years in the field. All five rated their familiarity with XAI as moderate to high, and similarly, they were well-versed in visualization techniques.

Activities and Duration. Mirroring the methodology of the previous evaluation, we continued with semi-structured interviews, adhering to the specific process and timing detailed in Table 2, concluding with the follow-up survey. To illustrate the generalizability of the system, we demonstrated it on more datasets for fine-grained text classification, including Medical Bios [ECCB23], GoEmotions [DMAK*20], and HWU64 [LESR21]. This round, while still assessing the overall usability of the system, emphasized evaluating the impact of the newly integrated modes and functionalities through the evaluation tasks and survey questions.

7.2.2. Results

Reflection on Overall Usability: The follow-up survey results showed the system had a good level of overall usability. According to the Likert-scale question responses, all five participants agreed on the system's utility, indicating that the tool was indeed beneficial. Regarding how easy the visualizations were to understand, there was a strong alignment among participants, two answering *agree* and three answering *strongly agree*. The system excelled in meeting the requirements of offering a high-level model understanding (R1), facilitating the understanding of individual predictions (R2), identifying model weaknesses (R3), and aiding users to discern sub-clusters within labels (R4), with all five participants answering *strongly agree*. The system's effectiveness in clarifying the fine-grained aspects of individual labels (R2) and distinguishing between similar labels (R4) received positive feedback, with four out of five participants answering *strongly agree* and one answering *agree*.

Reflection on Usefulness: In addition to the above, by analyzing the interview transcripts and the open-ended question responses, we found the common theme that the participants consistently praised the system's ability to "massively" reduce the complexity of many difficult analysis tasks in one tool, e.g., error analysis, model validation, bias detection, and data exploration, at different hierarchical levels. Regarding the efficacy of specific visualizations, the *comparison mode* emerged as a standout feature, receiving accolades from all five participants for its effectiveness. The Map view was commended by four participants, making these two features the most praised for their usefulness.

Participants provided valuable insights into unanticipated use cases for the tool, revealing potential areas for feature expansion. Notably, E6 highlighted its capacity to discern adversarial samples while E1, E2, and E4 recognized its utility in demystifying complex AI systems for non-expert audiences, such as elucidating why solutions like ChatGPT do not supplant task-oriented dialogue systems. These unexpected applications suggest promising directions for further refinement and enhancement of the tool's capabilities.

Expectations on Improvements Some experts commented that SemLa is ready to be polished for production and as such they expect minor improvements including UI enhancements and better system guidance, which we discuss in Section 9.

8. Case Studies

Throughout our design study, we conducted in-depth case studies on various fine-grained text classification datasets. We illustrate through two case studies how SemLa can be used to tackle important requirements and streamline the model development workflow.

8.1. Identifying Root Cause of Model Errors on BANKING77

BANKING77 is a popular public benchmark dataset for intent recognition [CTG*20]. This dataset has 77 labels related to the same banking domain, which makes it very fine-grained even among other fine-grained text classification datasets and representative of how complex datasets are in practical applications. We analyzed a BERT-base model (fine-tuned on the training split) on the test split with 3080 samples.

We began our analysis by filtering errors on the Map view and inspecting the label distributions in the List view. From the *gold label list* and *predicted label list* respectively, we saw the labels ranked by their shares of the false negative and false positives predictions. For example, 3.7% and 3.4% of the false negatives corresponded respectively to the labels *compromised_card* and *supported_cards_and_currencies*, whereas 3.2% and 3.2% of the false

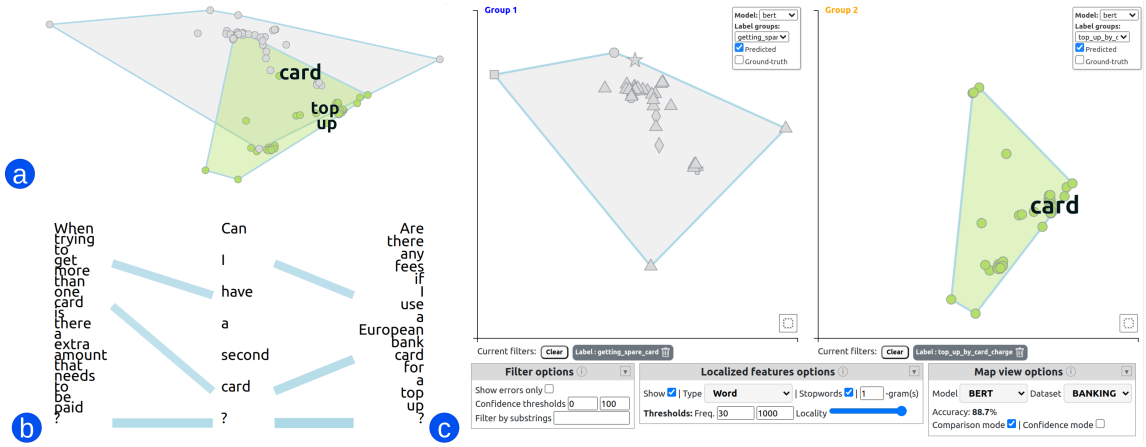


Figure 4: *Multi-level analysis results: a) local words in the two top-confused labels getting_spare_card and top_up_by_card_charge, b) token-to-token links in a false positive case of getting_spare_card (represented by an example on the left) being mistaken for topping_up_by_card_charge (represented by an example on the right) confirms that the word "card" was a confounding feature, and c) label topping_up_by_card_charge has the word "card" more often than the label getting_spare_card in the model predictions, as the word "card" appears over the former but not the latter at frequency threshold of 30.*

positives corresponded respectively to *top_up_by_card_charge* and *reverted_card_payment?*. The *confusion table* in the Label-level component, confirmed this, and sorting the table by the columns "ground-truth" and "prediction" provided a detailed breakdown of the specific confusions. Interestingly, when sorted by confusion frequency (clicking on the frequency column's header), we found the most frequent confusion was mistaking *getting_spare_card* for *top_up_by_card_charge*, which happened three times. Our goal became understanding this confusion and why the model most frequently predicted *top_up_by_card_charge* false positively.

To understand the top confusion, we first looked at the local words (ignoring stop words) with the default frequency threshold of 20 (selected interactively to balance visual clutter) and found that "card" was in the intersected area (shown by enclosing hulls corresponding to each label) of the two labels, and that "top" and "up" were only in the area of top_up_by_card_charge. The most common word across the two labels was "card", which appeared in 66.3% of samples (Fig. 4a), which is expected based on the label names. Furthermore, we looked at the three errors using the Sample level component. In all three errors, the word "card" was found to be a confounding factor (Fig. 4b shows one of these errors), that related to both the incorrect *top_up_by_card_charge label* and the correct *getting_spare_card* label, which suggested that the model associated the word "card" to the former more strongly.

To confirm this, we activated compare mode and looked at the two labels side by side and found that "card" indeed occurred in more samples predicted to be *top_up_by_card_charge* despite there being an equal number of samples for each label in the dataset (Fig. 4c). This was also true against the label that was second most frequently mistaken for *top_up_by_card_charge* (*supported_cards_and_currencies*).

Furthermore, when we compared the samples that were *predicted* to be *top_up_by_card_charge* with those that actually belonged to this label, the word "card" was more frequent in the former group than in the latter. On the other hand, the words "top" and "up" were less frequent in the former than in the latter. These suggested the model was giving the word "card" more importance than it should when predicting *top_up_by_card_charge*.

In summary, our analysis showed that the model associated the word "card", which was common among many other labels, too strongly with *top_up_by_card_charge*, which explains why the model most frequently predicted this label false positively. Based on this insight, we experimented with further fine-tuning the model on a small training set comprising 1) samples of *top_up_by_card_charge* that do not contain the word "card" and 2) samples of those (target) labels confused with *top_up_by_card_charge* (e.g., *getting_spare_card*) that do contain the word "card". Doing so removes all confusions of *top_up_by_card_charge* with the target labels and increases accuracy from 88.67% to 88.90%. When we do not filter the samples based on the word "card", confusion with some of the target labels still remains and the accuracy increase (to 88.77%) is less despite more training data. These results concretely show that SemLa can provide practical insights that can lead to model improvements.

8.2. Hidden Conceptual Relations in Multi-Domain Datasets

HWU64 [LESR21] and CLINC150 [LMP*19] are *multi-domain* intent recognition datasets with 64 labels in 18 domains and 150 labels in 10 domains respectively. We analyzed BERT-base models (each fine-tuned on the training split of the respective dataset) on the test splits containing 1076 and 4500 samples respectively. On these datasets, we often found unexpected cross-domain conceptual relationships between seemingly unrelated labels.

The most interesting case was on CLINC when looking at the most frequent confusion.

The most frequent confusion was between the seemingly unrelated labels *vaccines* and *cancel_reservation*. To investigate why, we clicked on the confusion to see the two labels on the Map view. Upon not finding an apparent connection between the labels when looking at the local words, we switched to visualizing the local concepts. Then, we found there are many countries mentioned in *cancel_reservation* label ("spain", "mexico", "america", "china" and "zimbabwe"), and that the error cases all contained the word "cuba", which is also a country. Even though "cuba" was not among any sample that actually has the label *cancel_reservation*, the model likely made these errors after recognizing that "cuba" was similar to other country names (see image above). The system automatically extracted and instantly showed us this hidden conceptual relation, which otherwise would not have been apparent without manually looking through the data in detail.

9. Reflection and Discussion

We identified several takeaways from our design study after analyzing our evaluation results and reflecting on our collaboration with the domain experts. These apply to developing a generalizable system that addresses the needs of different user profiles.

Cast a wide net to generalize The diverse backgrounds (dialogue system, medicine, abuse detection) of the participants in our design study entailed a wide range of requirements and individual differences in how they prioritized the requirements and model aspects (performance, explainability, robustness). This was reflected in the expert feedback, as the most novel feature according to each expert was often related to their background. Often, a system feature that one expert paid little attention to was the most novel to another. For example, the idea of our VIFI view, which was merely acknowledged by most experts, was highlighted as one of the most novel features with strong significance by expert E5 who has high level of experience and in-depth understanding of XAI methods. Therefore, our reflection is that what may seem like unnecessary complexity to one expert can be a necessity for another. When addressing a task with wide applicability like text classification, to prevent overly tailoring our system design to only a subset of potential users, it was worthwhile to ensure that our requirements analysis encompassed not only the practical challenges experienced by the experts, but also the common problems addressed by previous works and the background domain knowledge in XAI.

Resist the temptation to simplify Simplicity is a key factor in usability and an important design principle behind our novel visualizations. However, as previously discussed, neglecting individual differences in user requirements for simplicity would lead to poor generalizability. Furthermore, based on the questions we received from some of the experts, omitting low-level details behind the visualizations or key information that needs to be clear (e.g., how are the link strengths calculated in our relation graphs, or what are the labels that correspond to each column) can hurt transparency and reduce the simplicity *experienced* by the user.

Use XAI techniques responsibly As each individual explanation method simplifies model reasoning, each can only offer a limited perspective. Furthermore, as multiple competing methods disagree with each other (Section 3), acknowledging these limitations to the users and offering them multiple perspectives is essential for preventing misconceptions and using these methods responsibly. This applies to users at both ends of the spectrum when it comes to how much they know about the explanation methods and how likely they are to trust them. Based on our discussions with the experts, for users who lack knowledge of explanation methods, providing multiple perspectives and acknowledging the limitations are critical in reducing vulnerability to misconceptions, whereas for users who are generally familiar with explanation methods, the same multiple perspectives are required to address the current issue of "trusting the explanation methods themselves" [SSSEA20]. This pervasive need for multiple perspectives motivates adopting visual analytics systems when applying XAI methods in practice and new ways of integrating multiple explanations together into the same system.

Accompany freedom with guidance In our last evaluation round, two experts suggested a common direction for improving SemLa for production, which was guiding users to intuitively follow a series of steps to complete common tasks. They suggested providing documentations and tutorials within and outside the system, or tailoring the default settings and the UI design for the important and common usage scenarios. They explained that these comments are actually based on the system's strength– the wide variety of features and full freedom to explore the model and the data. This resonated with the initial challenge we faced in Iteration 1, which was deciding how to best exploit the freedom-to-explore based on user requirements. We reflect that while offering the users a high-degree of freedom is useful for generalizability, ensuring usability by providing guidance tailored to user requirements is essential.

10. Conclusion and Future Work

In this paper, we detailed the intensive iterative design study involving a total of six NLP experts (with different backgrounds and different roles) that resulted in our visual analytics system SemLa for analyzing fine-grained text classification models and datasets. Our evaluation of the final design based on expert feedback and case studies shows that SemLa effectively addresses the special challenges posed by the task and that it can overall be a useful tool for assisting experts in their workflow of analyzing models and datasets with a diverse range of use cases. In future iterations, we would like to refine SemLa to be used in production scenarios by adding more ways of extracting insight from data using our LWC algorithm and assisting in communication between experts and non-experts. We are also excited by the prospect of extending our techniques to other application domains of deep learning, such as image processing and multi-modal input processing.

Acknowledgement

We thank all experts who participated in this study and the anonymous reviewers who provided valuable feedback. This work was supported by UK Research and Innovation [EP/S023356/1] and the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

References

[ACT*24] ATZBERGER D., CECH T., TRAPP M., RICHTER R., SCHEIBEL W., DÖLLNER J., SCHRECK T.: Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *IEEE Transactions on Visualization and Computer Graphics 30*, 1 (2024), 902–912. doi:10.1109/TVCG.2023.3326569. 6

[BAL*15] BROOKS M., AMERSHI S., LEE B., DRUCKER S. M., KAPOOR A., SIMARD P.: FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), pp. 105–112. doi:10.1109/VAST.2015.7347637. 2

[BLDB24] BATTOGTOKH M., LUCK M., DAVIDESCU C., BORGO R.: Simple framework for interpretable fine-grained text classification. In *Artificial Intelligence. ECAI 2023 International Workshops* (2024), Springer Nature Switzerland, pp. 398–425. doi:10.1007/978-3-031-50396-2_23. 2, 3, 4, 5

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 30 (Mar. 2003), 993–1022. URL: https://dl.acm.org/doi/10.5555/944919.944937. 2

[CMQ21] CHENG F., MING Y., QU H.: DECE: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization & Computer Graphics 27*, 02 (Feb. 2021), 1438–1447. doi:10.1109/TVCG.2020.3030342. 3

[CTG*20] CASANUEVA I., TEMČINAS T., GERZ D., HENDERSON M., VULIĆ I.: Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for Conversational AI* (July 2020), pp. 38–45. doi:10.18653/V1/2020.NLP4CONVAI-1.5. 1, 2, 7, 8

[CZMX22] CHEN J., ZHANG R., MAO Y., XU J.: ContrastNet: A contrastive learning framework for few-shot text classification. *Proceedings of the AAAI Conference on Artificial Intelligence 36*, 10 (Jun. 2022), 10492–10500. doi:10.1609/aaai.v36i10.21292. 7

[DCLT19] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: 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* (Minneapolis, Minnesota, June 2019), pp. 4171–4186. doi:10.18653/v1/N19-1423. 7

[DJR*20] DEYOUNG J., JAIN S., RAJANI N. F., LEHMAN E., XIONG C., SOCHER R., WALLACE B. C.: ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), pp. 4443–4458. doi:10.18653/v1/2020.acl-main.408. 3

[DMAK*20] DEMSZKY D., MOVSHOVITZ-ATTIAS D., KO J., COWEN A., NEMADE G., RAVI S.: GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), pp. 4040–4054. doi:10.18653/v1/2020.acl-main.372. 8

[EAKC*20] EL-ASSADY M., KEHLBECK R., COLLINS C., KEIM D., DEUSSEN O.: Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 1001–1011. doi:10.1109/TVCG.2019.2934654. 2

[ECCB23] EBERLE O., CHALKIDIS I., CABELLO L., BRANDL S.: Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore, Dec. 2023), Bouamor H., Pino J., Bali K., (Eds.), pp. 6907–6920. doi:10.18653/v1/2023.emnlp-main.427. 1, 8

[EMK*21] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics 27*, 3 (2021), 2153–2173. doi:10.1109/TVCG.2019.2944182. 5

[FWC23] FENG J., WU K., CHEN S.: TopicBubbler: An interactive visual analytics system for cross-level fine-grained exploration of social media data. *Visual Informatics 7*, 4 (2023), 41–56. doi:10.1016/j.visinf.2023.08.002. 2

[GHYB20] GOMEZ O., HOLTER S., YUAN J., BERTINI E.: ViCE: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2020), IUI '20, p. 531–535. doi:10.1145/3377325.3377536. 3

[GHYB21] GOMEZ O., HOLTER S., YUAN J., BERTINI E.: AdViCE: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)* (2021), pp. 31–35. doi:10.1109/VIS49827.2021.9623271. 3

[Gro22] GROOTENDORST M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* (2022). doi:10.48550/arXiv.2203.05794. 6

[GZL*21] GOU L., ZOU L., LI N., HOFMANN M., SHEKAR A. K., WENDT A., REN L.: VATLD: A visual analytics system to assess, understand and improve traffic light detection. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2021), 261–271. doi:10.1109/TVCG.2020.3030350. 2

[HKPC19] HOHMAN F., KAHNG M., PIENTA R., CHAU D. H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics 25*, 8 (2019), 2674–2693. doi:10.1109/TVCG.2018.2843369. 2

[JG20] JACOVI A., GOLDBERG Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), pp. 4198–4205. doi:10.18653/v1/2020.acl-main.386. 3

[JKA*23] JEON H., KUO Y.-H., AUPETIT M., MA K.-L., SEO J.: Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics* (2023). doi:10.48550/arXiv.2308.00278. 5

[JSR*21] JACOVI A., SWAYAMDIPTA S., RAVFOGEL S., ELAZAR Y., CHOI Y., GOLDBERG Y.: Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic, Nov. 2021), pp. 1597–1611. doi:10.18653/v1/2021.emnlp-main.120. 2, 3

[JW19] JAIN S., WALLACE B. C.: Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Minneapolis, Minnesota, June 2019), pp. 3543–3556. doi:10.18653/v1/N19-1357. 3

[KDEP21] KIM H., DRAKE B., ENDERT A., PARK H.: ArchiText: Interactive hierarchical topic modeling. *IEEE Transactions on Visualization and Computer Graphics 27*, 9 (2021), 3644–3655. doi:10.1109/TVCG.2020.2981456. 2

[LBS*23] LIEBE. L., BAUM. J., SCHÜTZE. T., CECH. T., SCHEIBEL. W., DÖLLNER. J.: unCover: Identifying AI generated news articles by linguistic analysis and visualization. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR* (2023), pp. 39–50. doi:10.5220/0012163300003598. 2

[LCHJ16] LI J., CHEN X., HOVY E., JURAFSKY D.: Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (San Diego, California, June 2016), pp. 681–691. doi:10.18653/v1/N16-1082. 2

[LESR21] LIU X., ESHGHI A., SWIETOJANSKI P., RIESER V.: Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems* (Singapore, 2021), Springer Singapore, pp. 165–183. doi:10.1007/978-981-15-9323-9_15. 8, 9

[LL17] LUNDBERG S. M., LEE S.-I.: A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 4768–4777. URL: https://dl.acm.org/doi/10.5555/3295222.3295230. 2

[LLLZ21] LUO Q., LIU L., LIN Y., ZHANG W.: Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Online, Aug. 2021), pp. 2773–2782. doi:10.18653/v1/2021.findings-acl.245. 2

[LMP*19] LARSON S., MAHENDRAN A., PEPER J. J., CLARKE C., LEE A., HILL P., KUMMERFELD J. K., LEACH K., LAURENZANO M. A., TANG L., MARS J.: An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), pp. 1311–1316. doi:10.18653/v1/D19-1131. 9

[LPL*22] LI Q., PENG H., LI J., XIA C., YANG R., SUN L., YU P. S., HE L.: A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology 13*, 2 (Apr. 2022). doi:10.1145/3495162. 1

[LRBB*23] LA ROSA B., BLASILLI G., BOURQUI R., AUBER D., SANTUCCI G., CAPOBIANCO R., BERTINI E., GIOT R., ANGELINI M.: State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum 42*, 1 (2023), 319–355. doi:10.1111/cgf.14733. 2

[LS99] LEE D. D., SEUNG H. S.: Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (Oct. 1999), 788–791. doi:10.1038/44565. 2

[LST20] LERTVITTAYAKUMJORN P., SPECIA L., TONI F.: FIND: Human-in-the-loop debugging deep text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, Oct. 2020), pp. 332–348. doi:10.18653/v1/2020.emnlp-main.24. 1, 2, 6

[LWY*22] LI Z., WANG X., YANG W., WU J., ZHANG Z., LIU Z., SUN M., ZHANG H., LIU S.: A unified understanding of deep NLP models for text classification. *IEEE Transactions on Visualization and Computer Graphics 28*, 12 (2022), 4980–4994. doi:10.1109/TVCG.2022.3184186. 1, 2

[LYW19] LIU H., YIN Q., WANG W. Y.: Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), pp. 5570–5581. doi:10.18653/v1/P19-1560. 3

[MGS21] MEKALA D., GANGAL V., SHANG J.: Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic, Nov. 2021), pp. 583–594. doi:10.18653/v1/2021.emnlp-main.46. 2

[MH08] MAATEN L. V. D., HINTON G.: Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 86 (2008), 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html. 5

[MHSG18] MCINNES L., HEALY J., SAUL N., GROSSBERGER L.: UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software 3*, 29 (2018), 861. doi:10.21105/joss.00861. 5

[Mil19] MILLER T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267* (2019), 1–38. doi:10.1016/j.artint.2018.07.007. 2

[Ngu18] NGUYEN D.: Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics* (New Orleans, Louisiana, June 2018), pp. 1069–1078. doi:10.18653/v1/N18-1097. 3

[OUR*23] OKEY O. D., UDO E. U., ROSA R. L., RODRÍGUEZ D. Z., KLEINSCHMIDT J. H.: Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis. *Computers and Security 135* (2023), 103476. doi:10.1016/j.cose.2023.103476. 2

[RMP21] ROSS A., MARASOVIĆ A., PETERS M.: Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Online, Aug. 2021), pp. 3840–3852. doi:10.18653/v1/2021.findings-acl.336. 3

[RSG16] RIBEIRO M. T., SINGH S., GUESTRIN C.: "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* (New York, NY, USA, 2016), KDD '16, p. 1135–1144. doi:10.1145/2939672.2939778. 1, 2

[SACPF21] STEPIN I., ALONSO J. M., CATALA A., PEREIRA-FARIÑA M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access 9* (2021), 11974–12001. doi:10.1109/ACCESS.2021.3051315. 3

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization & Computer Graphics 18*, 12 (2012), 2431–2440. doi:10.1109/TVCG.2012.213. 4

[SO21] SURESH V., ONG D.: 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* (Online and Punta Cana, Dominican Republic, Nov. 2021), pp. 4381–4394. doi:10.18653/v1/2021.emnlp-main.359. 1, 2

[SRL*22] SAHU G., RODRIGUEZ P., LARADJI I., ATIGHEHCHIAN P., VAZQUEZ D., BAHDANAU D.: Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI* (Dublin, Ireland, May 2022), pp. 47–57. doi:10.18653/v1/2022.nlp4convai-1.5. 2

[SSSEA20] SPINNER T., SCHLEGEL U., SCHÄFER H., EL-ASSADY M.: explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 1064–1074. doi:10.1109/TVCG.2019.2934629. 2, 3, 10

[STY17] SUNDARARAJAN M., TALY A., YAN Q.: Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (2017), ICML'17, JMLR.org, p. 3319–3328. URL: https://dl.acm.org/doi/10.5555/3305890.3306024. 2

[Vig19] VIG J.: A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Florence, Italy, July 2019), pp. 37–42. doi:10.18653/v1/P19-3007. 1, 2

[YTJ*19] YAN Y., TAO Y., JIN S., XU J., LIN H.: An interactive visual analytics system for incremental classification based on semi-supervised topic modeling. In *2019 IEEE Pacific Visualization Symposium (PacificVis)* (2019), pp. 148–157. doi:10.1109/PacificVis.2019.00025. 2

[ZXD*23] ZHANG X., XUAN X., DIMA A., SEXTON T., MA K.-L.: LabelVizier: Interactive validation and relabeling for technical text annotations. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)* (2023), pp. 167–176. doi:10.1109/PacificVis56936.2023.00026. 2