

An Interdisciplinary Perspective on Evaluation and Experimental Design for Visual Text Analytics: Position Paper

Kostiantyn Kucher*
Linköping University
Linnaeus University

Nicole Sultanum†
University of Toronto

Angel Daza‡
Vrije Universiteit Amsterdam

Vasiliki Simaki§
Lund University

Maria Skeppstedt¶
CDHU, Department of ALM,
Uppsala University

Barbara Plank||
LMU Munich
IT University of Copenhagen

Jean-Daniel Fekete**
Inria
Université Paris-Saclay

Narges Mahyar††
UMass Amherst

ABSTRACT

Appropriate evaluation and experimental design are fundamental for empirical sciences, particularly in data-driven fields. Due to the successes in computational modeling of languages, for instance, research outcomes are having an increasingly immediate impact on end users. As the gap in adoption by end users decreases, the need increases to ensure that tools and models developed by the research communities and practitioners are reliable, trustworthy, and supportive of the users in their goals. In this position paper, we focus on the issues of evaluating visual text analytics approaches. We take an interdisciplinary perspective from the visualization and natural language processing communities, as we argue that the design and validation of visual text analytics include concerns beyond computational or visual/interactive methods on their own. We identify four key groups of challenges for evaluating visual text analytics approaches (data ambiguity, experimental design, user trust, and “big picture” concerns) and provide suggestions for research opportunities from an interdisciplinary perspective.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

The research on computational methods for language and specifically text data, in particular, natural language processing (NLP), has achieved impressive advances in the past several decades. The more traditional approaches have been complemented and in some cases replaced by data-driven approaches relying on machine learning (ML) and deep learning (DL), resulting in very large and capable models [13, 19, 64], which can eventually be used for particular downstream applications [67, 76]. However, a number of concerns have been voiced regarding the true capabilities [12], availability, biases, and risks [11], miscalibration [32, 54], and performance testing [77] of such modern models. In order to improve users’ understanding and eventually trust [46, 60] for computational models while making use of the advantages of human perception and cognition, researchers and practitioners in NLP might benefit from the knowledge and methods developed in the information visualization [22] and visual analytics [51] communities, which have been

traditionally open towards interdisciplinary collaborations and domain applications [36, 84, 96].

The purpose of visual text analytics (VTA) approaches in most cases is to support the users in accomplishing their tasks and achieving their goals that involve text data, computational methods including NLP/ML, and visual/interactive techniques (i.e., text visualization) [55, 61]. Examples of such approaches include systems for retrieval and exploration of patients’ medical histories for physicians [92] or investigation of news archives for historians [43]. In order to ensure that the proposed VTA tool indeed supports the users in achieving their goals and behaves in a reliable, predictable manner, it is imperative for designers of VTA solutions to consider the respective evaluation/validation concerns [23, 33, 47, 58, 75]. The choice of appropriate evaluation techniques and experimental designs is not a trivial issue, though, and it presents a number of open challenges. Some of such challenges include the trade-off between the study precision, realism, and generalizability [23], experimental data collection and analysis [75], scope and focus of the evaluation [58], expected reporting rigor [47], and applicability of typical evaluation patterns [33], among others.

The need for appropriate evaluation and experimental design strategies was in fact an important dimension of the discussions at a recent Dagstuhl seminar on VTA (<https://www.dagstuhl.de/22191>), which included 28 researchers and practitioners from both the NLP and visualization communities (10 and 18 participants, respectively; the overall group was also relatively well balanced with 11 females and 17 males, including 11 junior and 17 senior participants in terms of expertise). This diverse group has identified and discussed a number of VTA aspects and challenges in smaller subgroups. These topics included, among others, the diversity of data sources and applications of VTA, uncertainty and bias in VTA, NLP model interpretation, and visual representations for text data.

The diversity of the methods and traditions existing across these two disciplines has implications for the overall design and validation process of VTA solutions. In this regard, the need to establish the common ground between the NLP and visualization community representatives in order to make progress in VTA research was evident, as even the term “evaluation” leads to different associations and expectations in these disciplines, while the efforts from both sides are actually required. For example, the concerns regarding the performance of the underlying computational models (e.g., a review text classifier) must be complemented with human-centered concerns driven by the expectations and reactions of various actors involved with such VTA solutions, including the target end users (e.g., marketing managers), but also data annotators and experts in charge of the NLP/ML models, for instance. The existing bodies of knowledge on data annotation [6], performance evaluation of computational models for NLP [90], design and evaluation in information visualization and visual analytics [47, 63, 66, 84], and specifically design and evaluation of text visualization and VTA techniques [2, 55, 61, 85] provide us with initial pointers and particular recipes. However, we would argue that truly interdisciplinary

*e-mail: kostiantyn.kucher@{liu.se,lnu.se}

†e-mail: nicolebs@cs.toronto.edu

‡e-mail: j.a.dazaarevalo@vu.nl

§e-mail: vasiliki.simaki@englund.lu.se

¶e-mail: maria.skeppstedt@abm.uu.se

||e-mail: bplank@{cis.lmu.de,itu.dk}

**e-mail: jean-daniel.fekete@inria.fr

††e-mail: nmahyar@cs.umass.edu

Table 1: The main groups of challenges and further topics related to evaluation and experimental design for VTA identified within our discussion group consisting of visualization and NLP researchers.

High-level challenge group	Examples of specific concerns
Data ambiguity	Language/text data ambiguity; lack of single ground truth; making use of interactive visualization to reveal issues with raw and annotated data; concept drift implications for data and models
Experimental design	Taxonomies/categorizations of users, data, NLP tasks & scales, and respective interactive visualization evaluation methods; representation and evaluation issues specific to text data; evaluation for tools with multiple views and interactions
User trust	Ensuring user trust for NLP models and complete VTA solutions; going beyond time and error measurements on VTA evaluation; considering interpretability and explainability of models and decisions for trust in VTA; measuring the effects of model errors/noise on trust
VTA evaluation “big picture”	Systematic analysis of previous VTA evaluation efforts; guidelines for conducting and reporting VTA evaluations; repository of standard experimental designs for VTA; reproducibility concerns for case and user studies; long-term VTA adoption studies in real-world scenarios; retrofitting human-centered evaluation approaches for NLP models and tools

efforts are necessary to advance the understanding and practice of evaluation of VTA approaches, and furthermore, the results of such efforts could be beneficial for the respective communities beyond VTA-specific findings.

In this position paper, we report on the outcomes of the discussion that culminates in a vision of four groups of open challenges in evaluation and experimental design for VTA from the interdisciplinary perspective across the NLP and visualization communities. The authors of this manuscript formed a discussion group focusing on the issues of evaluation and experimental design at the seminar. Our group featured an equal number of participants with a background in visualization and NLP (4+4), which helped us with the identification of concerns and concepts related to all parts of the (visual) text analytics workflow that affect the choice of evaluation methods. More specifically, our prior experiences included the work on NLP problems such as disagreement and errors in data annotation [72, 95], semantic role labeling [30], and stance analysis [86], among others. With respect to text visualization and VTA, our experiences included works on clinical text visualization [91, 92], VTA support for data annotation and active learning for stance analysis [57], investigations of news archives [43], online reviews [49], historical [16] and scientific [20] document collections, as well as meta-analyses of the text visualization field [9, 55, 56]. Furthermore, several of the group participants shared prior interdisciplinary collaboration experiences on topic [88] and stance [87] analysis problems.

The discussion of our group at the seminar started with a short round of introductions and interests with respect to the group topics (some of which were shaped by our prior experiences mentioned above), and then we proceeded with a brainstorming session using sticky notes from individual participants. We analyzed the contents of the collected notes and organized them into several categories listed in Table 1. The following major topics were then discussed in smaller subgroups and eventually with other group participants:

- How should we account for the *lack of single ground truth in the complex and ambiguous real-world text data*? How can interactive visualization complement the traditional methodologies of data labeling and metric-based performance evaluation used in computational linguistics?
- What are the concerns that affect the *experimental design for text visualization and interaction techniques* in VTA approaches, taking the specifics of the respective NLP techniques into account?
- Moving beyond the design and evaluation of performance and usability, how can we ensure that the validation process accounts for the *users’ trust in the models* used by VTA solutions?
- What is the existing body of knowledge of evaluation in VTA approaches? Which concerns should be taken into account

when conducting *systematic analyses* of the existing VTA evaluation studies? How can *guidelines for future studies* be shaped?

The rest of this paper is organized according to these four topics. In Section 2, we highlight the issues of the traditional performance-based evaluation of NLP models (according to the usual use of the term “evaluation” in the NLP community) and highlight the opportunities for interactive/visual approaches to contribute to the respective activities carried out by NLP researchers and practitioners. Afterwards, in Section 3, we propose the dimensions of experimental design for human-centered modules of VTA solutions that follow from a discussion among NLP and visualization researchers in our discussion group. We argue for the importance of ensuring user trust for VTA systems as part of evaluation efforts in Section 4 and highlight the specific issues of trust assessment that should be considered. These and further challenges lead us to Section 5, where we argue for the need to review the “big picture” of VTA evaluation from an interdisciplinary perspective (including NLP, visualization, and further concerns) and propose first steps towards such a review. Finally, the paper is concluded with Section 6.

2 CONCERNS OF TEXT DATA AMBIGUITY AND LACK OF SINGLE GROUND TRUTH

While the visualization community primarily associates the term “evaluation” with human-centered validation approaches for particular techniques or tools, it is important to establish that the NLP community has different associations in this regard. Evaluation of NLP models is typically carried out by comparing the automatic classification performed by the model to a gold standard based on manual annotations [8]. That entails that the evaluation of any system based on NLP models is, in part, dependent on a gold standard. The creation of a gold standard for text data is not always a straightforward task: we typically assume that there is a single ground truth, and that an objective gold standard expressing this single ground truth can be created. In some cases, this assumption is, however, an oversimplification. Text data is rich and ambiguous, and there is no single canonical data representation or universally accepted interpretation [5, 8, 9, 68, 72, 73]. Ambiguity is a pervasive language phenomenon that can be found at all levels of linguistic analysis, from the phonological (*too vs two*) and the grammatical level (*run* as a noun vs *run* as a verb) to the semantic and conceptual level (as in the sentence “*I cannot find the mouse anywhere in the room*” that can have different interpretations in context depending on whether *mouse* refers to the animal or the computer device).

Ambiguity in language may lead to different interpretations of the text by the annotators, which as a result affect their final decisions. This is a frequent phenomenon especially in discourse annotation, as some relations such as *cause and effect* or *contrariety* are explicitly signalled in discourse by a connector/cue phrase (e.g., *because, but*) while other may not contain any linguistic items to signpost a more

implicit relation [45]. This may affect the annotator’s understanding of the text and thus their decision. But even in cases where annotators have a similar (or even identical) understanding of the text and they rely on a detailed, restrictive and well-crafted annotation protocol, their decisions may still differ at the end due to factors related to their perceptual system, pragmatics, and their background.

The following example illustrates plausible disagreement between two annotators (A and B) in a stance detection task:

Instance: “*even when our passions run high, our compassion must run deep.*” — Kevin Deyoung

Labels: *necessity* (Annotator A), *source of knowledge* (Annotator B)

In the example above, taken from a stance annotation task on Twitter as discussed in our previous work [86], Annotator A labeled the tweet with the *necessity* tag, probably due to the presence of the auxiliary verb *must* that denotes necessity, while Annotator B labeled the tweet as *source of knowledge* relying on the person’s name (Kevin Deyoung) as the source of the quoted text. Here, the annotators have a different perception of the overall stance expressed in this tweet, which one is the dominant one, and they probably do not share the same idea regarding the purpose of the tweet’s author. Such disagreements are observed when refined concepts, such as stance, are to be identified, grouped and annotated into different categories [87].

In some contexts, such as communication or language variation studies, the different perceptions of readers over a text due to language ambiguity may be an interesting phenomenon to explore and derive insights [69]. From the NLP perspective though, in the cases where the divergence between the annotators’ decisions is high, the task needs to be more closely specified (and/or simplified) and the annotation guidelines further refined, in order to reach an acceptable level of inter-annotator agreement (IAA) [6]. Traditionally, high IAA implies higher data quality (in terms of reproducibility of data labeling). But due to the nature of language itself and the various meanings that exist, assuming high agreement (and by consequence, a single ground truth) is at best an idealization [95]. This assumption of a ground truth neglects genuine disagreement, which can provide valuable insights about language use and interpretation. For some text interpretation tasks, too restrictive guidelines might even hide the fact there is no real ground truth for this particular task. While the body of work on learning from disagreement is growing in NLP, including our own prior efforts [95], there is a lack of interdisciplinary research between the NLP and visualization communities in this regard.

In particular, there are many examples of visualizations that create an overview of the results of text annotations, for visualizing the co-occurrence of different annotation categories. Approaches range from using standard visualization techniques (e.g. bar charts, scatter plots, Sankey and chord diagrams, and treemaps [50]) to developing new visualization techniques specific to the text annotation task (e.g., the ALVA tool for supporting annotation and active learning for stance analysis [57], which was developed in the scope of prior collaboration between several authors of this manuscript [87]). However, user interfaces for showing the difference in annotation categories between different annotators typically focus on 1) individual annotation instances (e.g., WebAnno [101]), or 2) IAA statistics for the annotators, such as matrices showing the inter-annotator agreement for pairs of users (e.g., ERAS [41]). That is, the aim of these visualizations is not to provide the user with an overview of the differences in annotation choices. This is natural, as these interfaces typically are created for curators who use them for annotation adjudication [14], i.e., merging each conflicting annotation into the one correct annotation, rather than exploring the annotated data set in order to gain insights into the reasons for disagreement. For

instance, insights into to what extent there is a single ground truth for the annotation task.

We believe that VTA provides opportunities for mutual benefits of research within NLP and text visualization to provide insights into the nature of disagreement. In addition to visualizing disagreement between annotators in terms of labeling divergences, there are additional factors that contribute to disagreement and could be visualized. For instance, the annotators could be given the opportunity to not only annotate, but also specify the level of uncertainty for the annotation, or to opt out of annotating an instance due to it being too difficult. Also the time it takes for the annotators to make an annotation decision could be gathered as an information source to indicate the difficulty of the task.

In order to make sense of all this information on the annotators and the annotation process, and decide in what direction to steer the annotators (to make the task more restrictive, or to accept that there is no single ground truth), the following information provided from the annotation process could be visualized:

- For which annotation cases do the annotators agree?
- For which annotation cases do the annotators disagree?
- How can you...
 - First provide an overview to what extent they agree/disagree;
 - Provide an overview of the cases with agreement/disagreement;
 - Make it possible to zoom in to different cases of agreement and disagreement;
 - Let the user zoom in on groups of cases of disagreements, and on the actual instance of disagreements;
 - Let the user make their own classifications of the agreement and disagreement and also make it possible to visualize these classifications; and
 - Visualize automatically derived classifications of annotator agreement and disagreement.

It might also be interesting to compare different annotation tasks, or the same annotation task on different corpora for their level of agreement/disagreement. A visualization might provide insights and a quick overview into the difference between different corpora, and thereby into the level of difficulty automating a classification task.

Most importantly, giving the possibility to let the user understand agreement/disagreement between annotators on different levels might help one get insight into whether the annotation task should be more clearly specified, or if there is no single ground truth for the text annotation task that has been given to the annotators.

3 CONCERNS FOR EXPERIMENTAL DESIGN: AUDIENCE, TASKS, SCALE, MODELS

Moving from the data-driven concerns towards the issues of validation of text visualization approaches, we consider that functionality of an NLP system and the respective (interactive) visualizations should be evaluated in light of at least the following four dimensions (see Figure 1): target *audience* (who?), list of *tasks* to support (why?), *scales* to consider in the tasks (linguistic objects of interest), and *models* supporting the tasks at the given scales (how?).

Audience: we start with the users who define what they want to perform in the system (goals). We can define the following User Groups (from larger to smaller in terms of the system knowledge and expected functionality): 1) *Debugger* or *Developer* of the system, 2) *Linguist* or *NLP Expert* in charge of the modeling, 3) *Domain Specialist* using the system with no particular knowledge of NLP

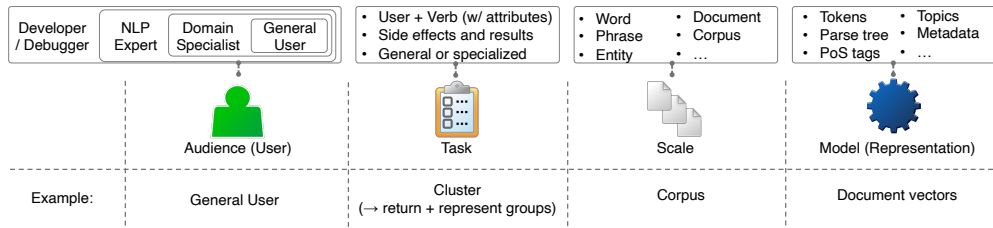


Figure 1: Dimensions of experimental design for text visualization techniques that take into account the specifics of the underlying NLP tasks and representations.

or CS (e.g., a physician or a historian), and 4) *General User*, where each member of a group has the ability to perform the tasks of their own group and the smaller groups contained within (cf. Figure 1). Several examples of our own published studies demonstrated the use of VTA approaches with public healthcare [88], clinical [91, 92], and civic [9, 10] text data sets, for instance. Considering the real-world implications of the use of such approaches, the importance of carefully considering the intended audience and their goals for design and evaluation of visual text analytics cannot be overstated.

Tasks: we then define the tasks that will achieve each of the users’ goals. A task is the combination of a User + Verb (and attributes); it can lead to a side effect (change in the visualization) or to a result (list of items). For example: *overview* (depends on the visual representation, no parameters), *query*, *search* (return a list of items), *cluster* (visualize groups and return a list of groups), etc. Some prototypical tasks are:

- Search documents containing given text snippets;
- Search sentences containing two adjectives;
- List documents similar to document X; or
- Group sentences containing keywords.

Specialized tasks can be domain-specific, but also specific to the linguistic model (aimed at the NLP specialist) or to the programmer, allowing the respective user to understand the internal data structure of the NLP and visualizations. Such tasks are not useful for the domain expert or to a lay user. Further potential tasks can be identified in the previous works on the more general task taxonomies in information visualization and visual analytics [3, 18, 52, 59, 82] as well as the sources focusing on text visualization and visual text analytics [2, 21, 25, 48, 61, 85, 93, 99], including the studies contributed by the authors of this manuscript [55, 56], among others.

Scale: each (or almost each) task can be applied or affect the text at a different level of granularity: *word*, *phrase*, *entities*, *tree structure* (syntax), *proposition* (semantics), *sentence*, *section*, *paragraph*, *document*, *corpus*, or *stream* [99], for instance.

Models: layers of linguistic information are encoded and structured differently: *parse tree*, *part-of-speech tags*, XML-encoded text (*metadata*), *stand-off annotation* (e.g., relevant spans in text), *topics*, etc. The model influences the tasks in the sense that it allows some tasks to be done more or less accurately, and with diverse levels of uncertainty, quality, errors, and artifacts. The models also constrain the scale of possible tasks and affect the choice of suitable visual encodings and interaction techniques.

4 USER TRUST IN AUTOMATION IN THE CONTEXT OF VISUAL TEXT ANALYTICS EVALUATION

Given the complex, multifaceted, and potentially ambiguous nature of text, there is a wide error range in NLP outcomes, as we cannot expect NLP systems to work perfectly in every circumstance, e.g., for each task and for each data set. This is something users

must take into account when leveraging text visual analytics systems for a particular problem, and is reflected in the extent users trust automated outcomes and take them into account. User trust is an evolving process [46], and it reflects perceived system performance over use and the risk associated with decision-making. For example, spelling and grammar correction in documents is a relatively mature problem with low impact for errors, and something users are generally comfortable delegating to automation. On the other hand, automatic summarization of patient records for clinical decision making is an open research problem with significant consequences for wrong medical choices [71].

Trust is a crucial factor when measuring overall task performance and user experience. Misplaced trust leads to misguided decision-making, when relying too much on wrong assessments (i.e., overtrust), generating user overhead in verifying largely accurate outcomes (i.e., undertrust), and in some cases leads users to completely ignoring the results due to lack of results and preferring the manual sensemaking process [65], as observed in our several prior studies [43, 49, 91, 92], for instance. As such, there will always be a need for a system that uses NLP to communicate to what extent the user can trust the automatic output of the system [79]. In visual text analytics, this is often enabled by providing access to original text sources, allowing users to calibrate their trust. Allowing users to fix errors is another important consideration [31], but propagating corrections across a dataset is not always a straightforward task. In the end, while guidelines have been proposed to communicate uncertainty and foster trust in visual (text) analytics systems [24, 26, 60, 79], designing for trust is not always seen as a first class concern.

Apart from design considerations to foster adequate trust, from an evaluation perspective, it is also important to consider how trust can be measured, and whether it matches NLP capabilities [53]. For example, the study by Dasgupta et al. [29] operationalizes user trust as a degree of perceived confidence in the analysis results; the self-reported confidence can then be compared across various user tasks, for instance. While the methodology of that study might be adapted to at least some VTA evaluation efforts, in general there are many challenging components to measuring adequate trust within a visual text analytics context. It entails 1) an understanding of the limitation areas of NLP algorithms within a particular application, 2) assessing user perception of the limits of automated outcomes, 3) assessing how well user perception matches real capabilities of automation, and 4) how trust levels evolve over time. While there is research in understanding and designing for trust, strategies for measuring trust are still not widely adopted in the evaluation of visual text analytics systems. Moving forward, we posit that the community should consider developing and fostering use of standardized and easily applicable metrics and instruments (e.g., questionnaires [53]) for measuring trust in this specific context, which should be applicable at different stages of use (i.e., measurable over a period of time).

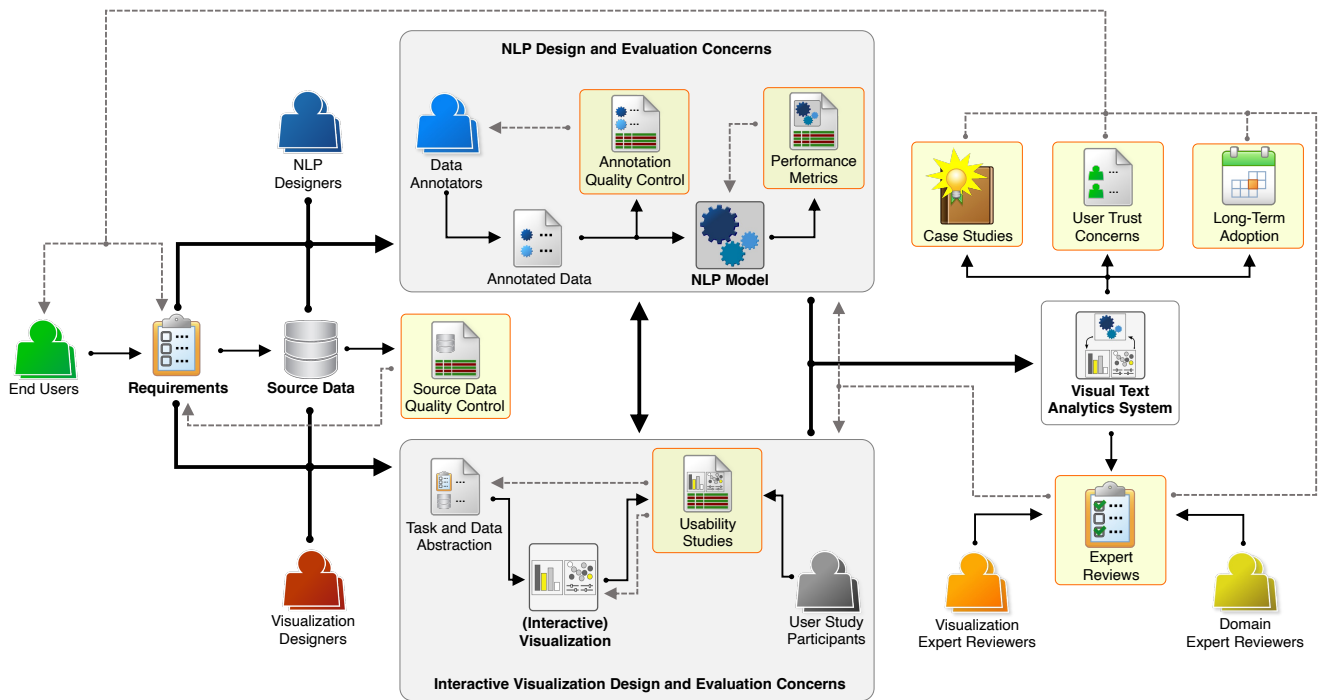


Figure 2: Sketch of a visual text analytics design and evaluation workflow with explicit identification of validation concerns which can guide our future efforts for systematic analysis and preparation of guidelines. The nodes with bold font labels indicate the main components and artifacts leading to the implementation of VTA solutions; the nodes highlighted with yellow background indicate quality control and validation activities; the grey shaded boxes represent the areas focusing on the particular NLP and visualization concerns; black edges indicate interactions and effects; thick black edges indicate the main interactions and effects leading to the implementation of VTA solutions; and grey dashed edges indicate feedback effects resulting from quality control and validation activities.

5 BIG PICTURE: SYSTEMATIC ANALYSIS AND GUIDELINES FOR EVALUATION OF VISUAL TEXT ANALYTICS

One of the challenges of evaluating visual text analytics approaches is related to the breadth and variety of concerns regarding computational and interactive, human-centered aspects of such approaches and tools. The efforts focusing on systematic analysis, categorization, and formulation of guidelines for evaluation (cf. the work by Isenberg et al. [47]) would thus be beneficial for researchers, practitioners, and end users of such visual text analytics tools. Here, the topics and dimensions would include an interdisciplinary analysis/view and include both NLP/ML-focused aspects (e.g., the methods and measurements of intra- and inter-annotator agreement, or the measurements of NLP model performance for a given task beyond standard ML metrics such as accuracy or F_1 -score) as well as human-centered aspects (e.g., the methods of usability measurement and long-term adoption observation). While the complexity and variety of visual text analytics problems would most likely not allow us to recommend a single prescribed method for validation, the outcomes of the work on systematic analysis would result in a collection of guidelines that would benefit the research efforts and applications.

One possible way to look at the issues of VTA evaluation from an interdisciplinary perspective (and taking the issues and opportunities for cross-/interdisciplinary effects and benefits, as discussed in the previous sections) would be to consider a sketch of a possible complete (visual) text analytics design and evaluation workflow. Inspired by the well-known visual analytics knowledge generation model by Sacha et al. [80], we consider the relations between the data, NLP model, visualization, users, researchers/designers from both NLP and visualization communities, and the respective design and evaluation concerns. In the rest of this section, we focus on

the sketch depicted in Figure 2 and the respective actors, activities, and artifacts, starting with the first steps towards the design of VTA solutions from scratch (the left part of the figure):

End users and requirements: while listed as “end” users, it is typically the intended goals and requirements of users that guide the design of visual text analytics approaches, for instance, an expert in bioinformatics might be interested in an interactive solution for exploring a large collection of research literature in their field and finding related work based on their interests, or a user of some social media platform might be interested in checking their own personal data over time with respect to sentiment and topics [25], or a product manager might be interested in investigating the opinions expressed in customer reviews. From these examples as well as the considerations given above in Section 3, we can establish that the level of expertise with respect to NLP and visualization can be very different from scenario to scenario. The previous work on this part of the VTA design and evaluation workflow includes, for instance, the study focusing on domain experts as the target users by Wong et al. [100], previously presented at BELIV, and the discussions of pre-design and discovery activities for visualization design and evaluation [58, 84].

Data: this includes all the different types of language data, either from written and/or oral discourse, in one or several languages [30, 74]. Depending on the nature of the data and its sources, different methods and techniques are used to extract and process it in order to bring it to the desired shape for the different NLP and visualization tasks. In many cases, complementary material is included in the data set to address ambiguity at different levels, as for instance in data from sign language (videos with gestures) and texts transcribed from oral discourse (audio files). The different genres, content, topics, and communicative situations/purposes are also

important characteristics to be identified, as discussed in Section 2. Furthermore, for complete VTA solutions, the raw data may include not only language/text information, but also further data types not directly addressed by NLP methods, for example, user metadata from social media platforms [25].

Data quality control: it is worth noting that before the specific design decisions for NLP and visualization are made, the raw data associated with the intended VTA application should be assessed, as the issues of missing, duplicate, or ambiguous records may have to be identified and addressed in some way. The previous work on data quality assessment can be considered here [70], including the implications for NLP/ML [40] and potential data augmentation techniques for NLP [38], among others.

We are now proceeding from the left to the central part of the workflow figure.

NLP designers and visualization designers: we have added the actors responsible for the overall NLP and visualization components in the VTA design and evaluation workflow sketch in Figure 2, but please notice that the same individuals might be “wearing a different hat” and play further roles within the design and evaluation process, e.g., as data annotators or eventual end users. We discuss the NLP components next (the bordered block in the upper central part of the workflow sketch), and then address the visualization components (the lower central part).

Annotators: an annotator is the person who is asked to read, assess, and label a given text according to specific instructions. Depending on the purpose of the study, it can be a domain expert (e.g., a linguist observing dialectal characteristics in language data collected from a specific region, or a historian assessing the relevance/trustworthiness of materials and sources regarding the French revolution) who has a background and knowledge relevant to the study. In many cases, the domain expert is also the one who designs the study and the annotation task, and assesses the results. In other cases, the annotator can be randomly selected and they can either belong to a specific target group (e.g., native speakers of English annotating the quality of texts written by learners of English), or just people interested in participating in a coding task (e.g., via crowdsourcing [89]).

Annotation process and annotation quality control: as discussed in Section 2, the process of attributing labels to text may differ significantly depending on the nature of the data and the purpose of the study. However, some general rules apply for the implementation of the study. A protocol that guides the annotator is drawn, in which the research expert define the concepts involved and describe the various steps that need to be followed. Based on the annotation task, annotators are selected and a pilot round is conducted. The results of this round are discussed with the research expert, and arisen issues are discussed in order to be addressed in the upcoming steps of the study. This can lead to refinements and/or changes in the annotation protocol. After this, the actual annotation round takes place. The annotation outcomes can be statistically evaluated, usually by using metrics such as Cohen’s kappa [27] that calculate the agreement between the annotation decision of two or more annotators (IAA). Opinions regarding the most suitable metrics for the IAA calculation and the acceptable level of this agreement vary, and researchers have not reached a consensus towards a common path to evaluate the quality the annotation results [6, 42, 62], which provides opportunities for further contributions.

Computational (NLP) models for text analysis: without going into excessive detail on the choice of particular algorithms and activities such as hyperparameter search, which are well-covered by the related work in NLP and ML, we should mention the issues of over-relying on pre-trained models [11, 12] and their trustworthiness, which might require additional calibration [32, 54]. Such issues have implications not only for the NLP model itself, but also other components of the VTA design and evaluation workflow.

Model evaluation methods and results: the traditional approach for assessing performance of NLP models relies on quantitative estimation of the model error and performance metrics calculated in relation to the ground truth data, such as accuracy, precision, recall, F_1 -score, etc. [90]. However, not only numbers of hits and misses should be taken into account, but also the respective task, audience, etc., as discussed in Section 3. The excessive reliance on benchmark performance evaluation of NLP models has also drawn criticism due to the eventual detriment of such properties as model size, energy efficiency, and practical utility for real-world language data [17, 34, 81]. In this regard, we believe that there is an opportunity for the NLP community to benefit from collaboration with the visualization community and make use of some of the human-centered evaluation considerations. Furthermore, the need for interactive approaches to allow NLP experts to prepare and investigate additional tests for their models in an efficient and flexible way has been recognized by the community [77].

We now switch to the visualization design and evaluation concerns (the lower central part in Figure 2).

Task and data abstractions: this is a more familiar territory for researchers in visualization, and the existing work on the respective methodologies [3, 18, 52, 59, 63, 66, 84] will definitely provide the basis for systematic analyses of VTA evaluation efforts. At the same time, we should take the concerns specific to language/text data and NLP models into account, too [93], as discussed in Section 3.

Interactive visual representations: to continue the previous point, we could comment on one representation that is widely used for text visualization tasks (and is often explicitly requested by end users), while raising issues with visualization researchers—word clouds [97] and their varieties. The various design issues and recommendations for application of this technique have been discussed in the literature [1, 37, 44], and it is an example of VTA design and evaluation concerns that must be considered in context of the target audience, tasks, and underlying representations (cf. Section 3).

Usability studies: while the traditional human-computer interaction approach for empirically assessing usability of (interactive) visualization techniques typically relies on task completion time and error rate measurement in laboratory settings [39], the visualization community (including the BELIV workshop) has long argued that this is not sufficient. Even considering the text visualization essentially on its own at this point (without the NLP model and synergy concerns), we can make use of the rich body of knowledge existing in the community in this regard [33, 47, 58, 75], including the work focusing on text visualization in particular [10] and even making use of traditional approaches for analyzing text data from the humanities [7], for instance.

User study participants: a particular role included in our workflow sketch in Figure 2 is of participants invited to the user studies of (interactive) text visualization approaches. It is important to remember of the audience and task considerations here (cf. Section 3), as undergraduate students or members of the general public may or may not be the suitable audience for particular studies, posing a risk for their validity. Furthermore, while we mentioned the existing efforts on making use of crowdsourcing for data annotation on the computational side of VTA [89] above, there are also discussions of feasibility and concerns of such activities for user studies [4], which should also be taken into account for VTA evaluation.

After focusing on the NLP and visualization components on their own, we finally reach the stage of complete VTA solutions (see the right part of Figure 2), with the expectation of helping users achieving better results, knowledge, and particular insights through the visual analytics process [51].

Visual text analytics solutions: the extra capacity and flexibility of such approaches might come at the cost of overcomplicated solutions relying on dozens of NLP models and dozens of interactive views; the issue is, once again, how to ensure that such solutions

work well for the intended audience, tasks, and data. The problem of evaluating visual analytics systems in comparison to individual views or interactions has been raised for quite some time in the visualization community, and we would argue that it is still an open challenge. However, the relevant considerations, e.g., for holistic vs reductionist views on evaluation [28] or evaluation of interactive visual machine learning systems [15], should definitely be included in the future systematic analysis efforts for VTA evaluation.

Expert reviews: some of the existing evaluation methods that could be applied to the complete VTA solutions rather than individual underlying techniques include reviews by the experts. Here, we could specifically refer to the work focusing on the feedback from experts in visualization and human-computer interaction [94], including not only interviews, but also structured questionnaires such as ICE-T with the focus on value added by visualization [98]; and the work focusing on the involvement of domain experts [33]. One comment that we should add in this regard is that the recommendations (or even requirements) for reporting the results of such reviews should cover the explicit statement on the overlap or other conflicts of interest between the experts and the authors of respective research studies (which might not always be evident, e.g., during the double-blind manuscript review process).

Case studies: we are now reaching the final part of the workflow sketch proposed in Figure 2, where the complete VTA solutions are applied by the intended end users. The outcomes of such applications with real-world data, reported as case studies, provide another opportunity for evaluation [33, 47, 58], and the particular studies should be included in the systematic analysis efforts for VTA evaluation. One additional note to mention here is related to the reporting of case study data, steps, and results, which should contribute to the reproducibility of the results by reviewers and eventual readers of the respective studies [35].

User trust concerns: as discussed in Section 4, the issues of user trust [26, 60, 79] permeate the VTA design and evaluation process (and the positioning of the respective node in Figure 2 can be seen as an oversimplification, for instance). The existing work on trust in automation, visual analytics for enhancing trust [24], evaluation of the respective concerns in visualization and visual analytics [29], and specific work on text visualization and VTA should be considered for the future systematic analysis efforts; furthermore, we envision a number of opportunities for future work in this direction.

Long-term adoption: finally, we should consider longitudinal studies or other potential approaches to gauge adoption of VTA solutions in the long term by the respective audiences. Such evidence, including observations and studies conducted outside of the academic environment, for instance, in commercial [78, 83] and public [9] settings, will be extremely valuable for both visualization and NLP communities.

6 CONCLUSION

In this position paper, we have taken an interdisciplinary perspective on evaluation of visual text analytics approaches based on the existing work and our own prior knowledge in NLP and visualization. Our main message is that there is definitely interest and opportunity to benefit from collaborations in the scope of VTA evaluation research. We have discussed the issues of data ambiguity and lack of single ground truth which are usual for real-world language and text data; the important experimental design dimensions to consider for evaluation of particular (interactive) text visualization techniques; the importance of user trust issues for VTA design and evaluation; and finally, the need to conduct further systematic analyses of VTA evaluation efforts. Our proposed sketch of a VTA design and evaluation workflow relates to a number of prior works and it has already led us to further points in our interdisciplinary group discussions, including the existing knowledge gaps within and across the disciplines. For instance, if we compare the typical NLP and visualization

evaluation methodologies, we would notice that the visualization community often relies on time and error rates in usability studies, while the NLP community puts a lot more focus on error rates rather than time/efficiency concerns—can we learn from each other in this and further regards? Can we align the design and evaluation methodologies from our fields in order to leap forward together?

ACKNOWLEDGMENTS

This paper is a result of the discussion group work at Dagstuhl Seminar 22191 “Visual Text Analytics”. The authors would like to thank the organizers and the rest of seminar participants for this opportunity and the feedback provided for our group work.

REFERENCES

- [1] E. Alexander, C.-C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. Perceptual biases in font size as a data encoding. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2397–2410, Aug. 2018. doi: 10.1109/TVCG.2017.2723397
- [2] M. Alharbi and R. S. Laramée. SoS TextVis: An extended survey of surveys on text visualization. *Computers*, 8(1), 2019. doi: 10.3390/computers8010017
- [3] R. A. Amar and J. T. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, July–Aug. 2005. doi: 10.1109/TVCG.2005.63
- [4] D. Archambault, H. Purchase, and T. Höbflöd, eds. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions*. Springer, 2017. doi: 10.1007/978-3-319-66435-4
- [5] L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564
- [6] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, Dec. 2008. doi: 10.1162/coli.07-034-R2
- [7] A. Bares, S. Zeller, C. D. Jackson, D. F. Keefe, and F. Samsel. Using close reading as a method for evaluating visualizations. In *Proceedings of the IEEE Workshop on Evaluation and Beyond — Methodological Approaches to Visualization, BELIV '20*, pp. 29–37. IEEE, 2020. doi: 10.1109/BELIV51497.2020.00011
- [8] V. Basile. It’s the end of the gold standard as we know it. In M. Baldoni and S. Bandini, eds., *AIXIA 2020 – Advances in Artificial Intelligence*, pp. 441–453. Springer International Publishing, 2021.
- [9] E. P. S. Baumer, M. Jasim, A. Sarvghad, and N. Mahyar. Of course it’s political! A critical inquiry into underemphasized dimensions in civic text visualization. *Computer Graphics Forum*, 41(3), June 2022. doi: 10.1111/cgf.14518
- [10] E. P. S. Baumer, J. Snyder, and G. K. Gay. Interpretive impacts of text visualization: Mitigating political framing effects. *ACM Transactions on Computer-Human Interaction*, 25(4), Aug. 2018. doi: 10.1145/3214353
- [11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pp. 610–623. ACM, 2021. doi: 10.1145/3442188.3445922
- [12] E. M. Bender and A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198. ACL, July 2020. doi: 10.18653/v1/2020.acl-main.463
- [13] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug. 2013. doi: 10.1109/TPAMI.2013.50
- [14] C. Biemann, K. Bontcheva, R. Eckart de Castilho, I. Gurevych, and S. M. Yimam. Collaborative web-based tools for multi-layer text annotation. In N. Ide and J. Pustejovsky, eds., *Handbook of Linguistic*

- Annotation*, pp. 229–256. Springer Netherlands, 2017. doi: 10.1007/978-94-024-0881-2_8
- [15] N. Boukhelifa, A. Bezerianos, R. Chang, C. Collins, S. M. Drucker, A. Endert, J. Hullman, C. North, and M. Sedlmair. Challenges in evaluating interactive visual machine learning systems. *IEEE Computer Graphics and Applications*, 40(6):88–96, Nov.–Dec. 2020. doi: 10.1109/MCG.2020.3017064
- [16] N. Boukhelifa, E. Giannakis, E. Dimara, W. Willett, and J.-D. Fekete. Supporting historical research through user-centered visual analytics. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA), EuroVA '15*. The Eurographics Association, 2015. doi: 10.2312/eurova.20151095
- [17] S. R. Bowman and G. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '21*, pp. 4843–4855. ACL, 2021. doi: 10.18653/v1/2021.naacl-main.385
- [18] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, Dec. 2013. doi: 10.1109/TVCG.2013.124
- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [20] P. Caillou, J. Renault, J.-D. Fekete, A.-C. Letournel, and M. Sebag. Cartolabe: A web-based scalable visualization of large document collections. *IEEE Computer Graphics and Applications*, 41(2):76–88, Mar.–Apr. 2021. doi: 10.1109/MCG.2020.3033401
- [21] N. Cao and W. Cui. *Introduction to Text Visualization*, vol. 1 of *Atlantis Briefs in Artificial Intelligence*. Atlantis Press, 2016. doi: 10.2991/978-94-6239-186-4
- [22] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.
- [23] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, eds., *Information Visualization: Human-Centered Issues and Perspectives*, vol. 4950 of *LNCS*, pp. 19–45. Springer, 2008. doi: 10.1007/978-3-540-70956-5_2
- [24] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum*, 39(3):713–756, June 2020. doi: 10.1111/cgf.14034
- [25] S. Chen, L. Lin, and X. Yuan. Social media visual analytics. *Computer Graphics Forum*, 36(3):563–587, June 2017. doi: 10.1111/cgf.13211
- [26] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, pp. 443–452. ACM, 2012. doi: 10.1145/2207676.2207738
- [27] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, Apr. 1960. doi: 10.1177/001316446002000104
- [28] M. Correll, E. Alexander, D. Albers, A. Sarikaya, and M. Gleicher. Navigating reductionism and holism in evaluation. In *Proceedings of the Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV '14*, pp. 23–26. ACM, 2014. doi: 10.1145/2669557.2669577
- [29] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. LaFrance, N. Cramer, K. Cook, and S. Payne. Familiarity vs trust: A comparative study of domain scientists’ trust in visual analytics and conventional analysis methods. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):271–280, Jan. 2017. doi: 10.1109/TVCG.2016.2598544
- [30] A. Daza and A. Frank. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '20*, pp. 3904–3914. ACL, 2020. doi: 10.18653/v1/2020.emnlp-main.321
- [31] M. De-Arteaga, R. Fogliato, and A. Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '20*. ACM, 2020. doi: 10.1145/3313831.3376638
- [32] S. Desai and G. Durrett. Calibration of pre-trained transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '20*, pp. 295–302. ACL, 2020. doi: 10.18653/v1/2020.emnlp-main.21
- [33] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269, 2015. doi: 10.1177/1473871613513228
- [34] K. Ethayarajh and D. Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '20*, pp. 4846–4853. ACL, 2020. doi: 10.18653/v1/2020.emnlp-main.393
- [35] J.-D. Fekete and J. Freire. Exploring reproducibility in visualization. *IEEE Computer Graphics and Applications*, 40(5):108–119, Sept.–Oct. 2020. doi: 10.1109/MCG.2020.3006412
- [36] J.-D. Fekete, J. J. van Wijk, J. T. Stasko, and C. North. The value of information visualization. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, eds., *Information Visualization: Human-Centered Issues and Perspectives*, vol. 4950 of *LNCS*, pp. 1–18. Springer, 2008. doi: 10.1007/978-3-540-70956-5_1
- [37] C. Felix, S. Franconeri, and E. Bertini. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):657–666, Jan. 2018. doi: 10.1109/TVCG.2017.2746018
- [38] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pp. 968–988. ACL, Aug. 2021. doi: 10.18653/v1/2021.findings-acl.84
- [39] E. Frøkjær, M. Hertzum, and K. Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00*, pp. 345–352. ACM, 2000. doi: 10.1145/332040.332455
- [40] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, Mar. 2010. doi: 10.1007/s00521-009-0295-6
- [41] J. S. Grosman, P. H. Furtado, A. M. Rodrigues, G. G. Schardong, S. D. Barbosa, and H. C. Lopes. ERAS: Improving the quality control in the annotation process for natural language processing tasks. *Information Systems*, 93:101553, Nov. 2020. doi: 10.1016/j.is.2020.101553
- [42] K. Gwet. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical methods for inter-rater reliability assessment series, STATAxis Consulting, Apr. 2002.
- [43] A. Handler, N. Mahyar, and B. O’Connor. ClioQuery: Interactive query-oriented text analytics for comprehensive investigation of historical news archives. *ACM Transactions on Interactive Intelligent Systems*, mar 2022. doi: 10.1145/3524025
- [44] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri. An evaluation of semantically grouped word cloud designs. *IEEE Transactions on Visualization and Computer Graphics*, 26(9):2748–2761, Sept. 2020. doi: 10.1109/TVCG.2019.2904683
- [45] J. Hoek and M. Scholman. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the Joint ISO-ACL Workshop on Interoperable Semantic Annotation, ISA-13*, 2017.
- [46] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink. Trust in automation. *IEEE Intelligent Systems*, 28(1):84–88, Jan.–Feb. 2013. doi: 10.1109/MIS.2013.24
- [47] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, Dec. 2013. doi: 10.1109/TVCG.2013.126
- [48] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. Vi-

- sual text analysis in digital humanities. *Computer Graphics Forum*, 36(6):226–250, 2017. doi: 10.1111/cgf.12873
- [49] M. Jasim, C. Collins, A. Sarvghad, and N. Mahyar. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, 2022. doi: 10.1145/3491102.3517649
- [50] E. Kavaz, A. Puig, I. Rodriguez, M. Taule, and M. Nofre. Data visualization for supporting linguists in the analysis of toxic messages. *Computer Science Research Notes*, 3101:59–70, May 2021. doi: 10.24132/CSRN.2021.3101.7
- [51] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds. *Masterying the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010.
- [52] N. Kerracher and J. Kennedy. Constructing and evaluating visualisation task classifications: Process and considerations. *Computer Graphics Forum*, 36(3):47–59, June 2017. doi: 10.1111/cgf.13167
- [53] S. C. Kohn, E. J. De Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, 12, Oct. 2021. doi: 10.3389/fpsyg.2021.604977
- [54] L. Kong, H. Jiang, Y. Zhuang, J. Lyu, T. Zhao, and C. Zhang. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pp. 1326–1340. ACL, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.102
- [55] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific Visualization Symposium*, PacificVis '15, pp. 117–121. IEEE, 2015. doi: 10.1109/PACIFICVIS.2015.7156366
- [56] K. Kucher, C. Paradis, and A. Kerren. The state of the art in sentiment visualization. *Computer Graphics Forum*, 37(1):71–96, Feb. 2018. doi: 10.1111/cgf.13217
- [57] K. Kucher, C. Paradis, M. Sahlgren, and A. Kerren. Active learning and visual analytics for stance classification with ALVA. *ACM Transactions on Interactive Intelligent Systems*, 7(3):14:1–14:31, Oct. 2017. doi: 10.1145/3132169
- [58] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Sept. 2012. doi: 10.1109/TVCG.2011.279
- [59] H. Lam, M. Tory, and T. Munzner. Bridging from goals to tasks with design study analysis reports. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):435–445, Jan. 2018. doi: 10.1109/TVCG.2017.2744319
- [60] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, Mar. 2004. doi: 10.1518/hfes.46.1.50.30392
- [61] S. Liu, X. Wang, C. Collins, W. Dou, F. Ouyang, M. El-Assady, L. Jiang, and D. A. Keim. Bridging text visualization and mining: A task-driven survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2482–2504, July 2019. doi: 10.1109/TVCG.2018.2834341
- [62] M. Lombard, J. Snyder-Duch, and C. C. Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604, Oct. 2002. doi: 10.1111/j.1468-2958.2002.tb00826.x
- [63] M. Meyer, M. Sedlmair, and T. Munzner. The four-level nested model revisited: Blocks and guidelines. In *Proceedings of the BELIV Workshop: Beyond Time and Errors — Novel Evaluation Methods for Visualization*, BELIV '12. ACM, 2012. doi: 10.1145/2442576.2442587
- [64] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký. Strategies for training large scale neural network language models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding*, ASRU '11, pp. 196–201. IEEE, 2011. doi: 10.1109/ASRU.2011.6163930
- [65] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27:527–539, Nov. 1987. doi: 10.1016/S0020-7373(87)80013-5
- [66] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009. doi: 10.1109/TVCG.2009.111
- [67] D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, Feb. 2021. doi: 10.1109/TNNLS.2020.2979670
- [68] E. Pavlick and T. Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, Nov. 2019. doi: 10.1162/tacl.a.00293
- [69] S. T. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, Mar. 2012. doi: 10.1016/j.cognition.2011.10.004
- [70] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, Apr. 2002. doi: 10.1145/505248.506010
- [71] R. Pivovarov and N. Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, Sept. 2015. doi: 10.1093/jamia/ocv032
- [72] B. Plank, D. Hovy, and A. Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '14, pp. 507–511, 2014.
- [73] M. Poesio and R. Artstein. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 76–83. ACL, 2005.
- [74] E. M. Ponti, H. O’Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601, Sept. 2019. doi: 10.1162/coli.a.00357
- [75] H. C. Purchase. *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press, 2012. doi: 10.1017/CBO9780511844522
- [76] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, Oct. 2020. doi: 10.1007/s11431-020-1647-3
- [77] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. ACL, 2020. doi: 10.18653/v1/2020.acl-main.442
- [78] D. M. Russell. Simple is good: Observations of visualization use amongst the Big Data digerati. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, pp. 7–12. ACM, 2016. doi: 10.1145/2909132.2933287
- [79] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, Jan. 2016. doi: 10.1109/TVCG.2015.2467591
- [80] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, Dec. 2014. doi: 10.1109/TVCG.2014.2346481
- [81] D. Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, ACL-IJCNLP '21, pp. 670–674. ACL, 2021. doi: 10.18653/v1/2021.acl-short.85
- [82] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, Dec. 2013. doi: 10.1109/TVCG.2013.120
- [83] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz. Evaluating information visualization in large companies: Challenges, experiences and recommendations. In *Proceedings of the Beyond Time and Errors: Novel Evaluation Methods for Information Visualization Workshop*,

- BELIV '10, pp. 79–86. ACM, 2010. doi: 10.1145/2110192.2110204
- [84] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, Dec. 2012. doi: 10.1109/TVCG.2012.213
- [85] A. Shamim, V. Balakrishnan, and M. Tahir. Evaluation of opinion visualization techniques. *Information Visualization*, 14(4):339–358, Oct. 2015. doi: 10.1177/1473871614550537
- [86] V. Simaki. Getting into the influencer’s head: Stance-taking in social media discourse. In *Proceedings of the International Conference on Stance, (Inter)Subjectivity & Identity in Discourse, STANCEDISC '20*, pp. 9–10, 2020.
- [87] V. Simaki, C. Paradis, M. Skeppstedt, M. Sahlgren, K. Kucher, and A. Kerren. Annotating speaker stance in discourse: The Brexit Blog Corpus. *Corpus Linguistics and Linguistic Theory*, 16(2):215–248, Oct. 2020. doi: 10.1515/clt-2016-0060
- [88] M. Skeppstedt, K. Kucher, M. Stede, and A. Kerren. Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. In *Proceedings of the Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources at LREC '18, VisLR III*. ELRA, 2018.
- [89] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, pp. 254–263. ACL, 2008.
- [90] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009. doi: 10.1016/j.ipm.2009.03.002
- [91] N. Sultanum, M. Brudno, D. Wigdor, and F. Chevalier. More text please! Understanding and supporting the use of visualization for clinical text overview. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '18*. ACM, 2018. doi: 10.1145/3173574.3173996
- [92] N. Sultanum, D. Singh, M. Brudno, and F. Chevalier. Doccurate: A curation-based approach for clinical text visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):142–151, Jan. 2019. doi: 10.1109/TVCG.2018.2864905
- [93] M. Tofiloski, L. Odilinye, J. Sabharwal, and F. Popowich. A taxonomy of the target-problem space for visual text analytics. In *Proceedings of the UI Workshop on Visual Text Analytics, TextVis '15*, 2015.
- [94] M. Tory and T. Möller. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, Sept. 2005. doi: 10.1109/MCG.2005.102
- [95] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, Dec. 2021. doi: 10.1613/jair.1.12752
- [96] J. J. van Wijk. Bridging the gaps. *IEEE Computer Graphics and Applications*, 26(6):6–9, Nov. 2006. doi: 10.1109/MCG.2006.120
- [97] F. B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, July 2008. doi: 10.1145/1374489.1374501
- [98] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):491–500, Jan. 2019. doi: 10.1109/TVCG.2018.2865146
- [99] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim. State-of-the-art report of visual analysis for event detection in text data streams. In *Proceedings of the Eurographics Conference on Visualization — STARs, EuroVis '14*, pp. 125–139. The Eurographics Association, 2014. doi: 10.2312/eurovisstar.20141176
- [100] Y. L. Wong, K. Madhavan, and N. Elmqvist. Towards characterizing domain experts as a user group. In *Proceedings of the IEEE Evaluation and Beyond — Methodological Approaches for Visualization Workshop, BELIV '18*, pp. 1–10. IEEE, 2018. doi: 10.1109/BELIV.2018.8634026
- [101] S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL '13*, pp. 1–6. ACL, 2013.