Why Do We Need Domain-Experts for End-to-End Text Classification? An Overview

Jakob Smedegaard Andersen

Department of Computer Science, Hamburg University of Applied Science, Germany

Keywords: Text Classification, Human-in-the-Loop, Hybrid Intelligent Systems.

Abstract: The aim of this study is to provide an overview of human-in-the-loop text classification. Automated text classification faces several challenges that negatively affect its applicability in real-world domains. General obstacles are a lack of labelled examples, limited held-out accuracy, missing user trust, run-time constraints, low data quality and natural fuzziness. Human-in-the-loop is an emerging paradigm to continuously support machine processing, i.e. text classification, with prior human knowledge, aiming to overcome the limitations of purely artificial processing. In this survey, we review current challenges of pure automated text classifiers and outline how a human-in-the-loop can overcome these obstacles. We focus on end-to-end text classification and feedback of domain-experts, which do not process technical knowledge about the algorithms used. Further, we discuss common techniques to guide human attention and efforts within the text classification process.

1 INTRODUCTION

Involving domain experts in text classification can bridge the gap between machine learning (ML) research and real-world applications. While recent automated text classifiers have achieved great success on many benchmarks (Devlin et al., 2019; Yang et al., 2019), the applicability of automated text classifiers in real-world environments is still limited and faces many challenges. Even state-of-the-art text classifiers, such as BERT (Devlin et al., 2019), are not able to consistently reach desirable classifications results on arbitrary datasets. Automated text classifiers generally lack reliability, explainability, interpretability and human trust.

Semi-automated approaches have lately gained in prominence (Keim et al., 2008; Amershi et al., 2014; Holzinger, 2016), in which human background knowledge, abilities, and expertise are tightly coupled with ML models. Allowing humans to interact with ML models, aims to overcome the obstacles of purely artificial approaches and ultimately increases the applicability of artificial assisted decision-making. The term human-in-the-loop (HiL) (Holzinger, 2016) emerged to describe a semi-automated process characterized by the continuous support of machine processing by human feedback. HiL problem-solving aims to achieve what neither a human nor a machine can achieve on their own. In this work, we survey HiL for text classifiers with a focus on non-ML expert systems, where domain experts support the machine processing with their domain-specific prior knowledge and expertise. We analyse the main types of HiL implementations for text classification and highlight how these solve various challenges related to pure automated approaches.

Text classification is a widespread research challenge with high practical demands. It describes the process of assigning predefined class labels to natural language texts. Since large-scale manual labelling of text documents is a tedious, time-consuming and expensive task, there is a high demand for automation. In order to increase the applicability of automated classifiers, the question arises how domain-experts and artificial classifiers can work together efficiently. In particular, explicit uncertainty information (Der Kiureghian and Ditlevsen, 2009) and explanations (Adadi and Berrada, 2018) have shown to provide valuable insights into automated decision-making that help to effectively spend human efforts. To our best knowledge, there is no survey, focusing on overcoming the limitations of automated text classification via domain-experts. Wu et al. (Wu et al., 2022) provide a general survey of human-in-the-loop in conjunction with a range of ML tasks. Wang et al. (Wang et al., 2021) survey how several *natural language processing* (NLP) tasks can benefit from human feedback.

Andersen, J.
Why Do We Need Domain-Experts for End-to-End Text Classification? An Overview.
DOI: 10.5220/0011605900003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 17-24
ISBN: 978-989-758-623-1; ISSN: 2184-433X

The remainder of the paper is structured as follows: Section 2 introduces the task of automated text classification and outlines current challenges in its application. Then, Section 3 defines the human-in-the-loop approaches and introduces common types of human feedback. Section 4 looks at approaches to efficiently bring humans in the classification loop, and Section 5 reviews generic applications of the HiL paradigm to support text classification. Finally, Section 6 discusses open challenges and Section 7 concludes the paper.

2 AUTOMATED TEXT CLASSIFICATION

Text classification is about assigning text documents to predefined classes (Sebastiani, 2002). It has become a major challenge for research, as large amounts of textual data are produced daily in many real-world applications. However, fully automated classification algorithms remain imperfect and have several limitations that negatively impact their applicability in real-world domains. In the following, we highlight general limitations of purely automated text classifiers.

Lack of Knowledge. State-of-the-art classifiers are usually deep neural networks (Minaee et al., 2021) consisting of millions of parameters. Such classifiers require a lot of training data to efficiently model the prediction function. However, labelled data are typically scarce and represent a significant bottleneck in classification. Human labelling is time-consuming, labour-intensive, and costly, especially when domain-experts are needed. Therefore, the limited availability of training data is a serious obstacle to the application of automated text classifiers.

Further, it is generally assumed that the samples a model faces during deployment come from the same distribution as the training data set. However, the distribution of new in-domain data is generally unknown, and the underlying distribution is prone to shift over time (Ovadia et al., 2019). This makes it difficult to maintain a representative and meaningful training dataset, which is essential for training a well performing classifier.

Lack of Performance and Reliability. A primary goal of text classification is to achieve the highest possible accuracy. However, classification algorithms are inherently uncertain and misclassifications must be expected (Der Kiureghian and Ditlevsen, 2009). The exact relationship between class labels and text inputs remains unknown and can only be approximated by classifiers. Misclassifications usually occur because of missing training data, inappropriate selection of the classification algorithm, input noise, limited number of processable features (Gao et al., 2021), overfitting (Roelofs et al., 2019) or because unknown classes in the training data are mistakenly perceived as different labels (Zhao et al., 2021). Even when a large amount of training data is available, the most advanced text classifiers rarely achieve 100% accuracy on the test dataset and even more rarely on unseen data. Typically, measured held-out accuracies overestimate the real performance of classifiers on real-world data (Ribeiro et al., 2020). Given a labelled text corpus, the performance of a classifier usually converges to a maximum achievable accuracy that the model itself cannot exceed. Benchmarks indicate that an accuracy of around 90% on well scoped tasks can be expected (Devlin et al., 2019; Yang et al., 2019).

Lack of Transparency. Classification algorithms, especially deep learning approaches, are considered "black-boxes" for humans because they do not provide comprehensible insights into their decision-making. Practitioners are confronted with classification results without being told why and how these predictions were made. Without any human-readable explanations, practitioners can hardly be convinced of the classification result. Practitioners do generally not trust artificial classification results if they cannot understand why and how these decisions were made. Mechanisms for transparency aim to increase the trustworthiness, conformity and ultimately the applicability of text classifiers in practice (Adadi and Berrada, 2018). However, it is not enough to just explain the classifiers internal behaviour, humans must also understand them. Explainability can only be achieved through the interaction between humans and ML models (Adadi and Berrada, 2018).

Computational Complexity. The increasing accessibility of powerful computational instructions is paving the way for more complex classification algorithms that continue to push the boundaries of the state-of-the-art. Recent classifiers consist of more than 100 million trainable parameters (Devlin et al., 2019; Yang et al., 2019). Such complex models require a lot of computing time and resources, which precludes their use in product environments and on large datasets. Excluding many practitioners from their application. Long training or inference times also negatively impact the user experience when humans are involved in the classification loop, as long waiting times occur between interactions. A common workaround is to use simple linear classification models instead, such as *FastText* (Joulin et al., 2017) or less complex neural networks (Kim, 2014), which reduce training, testing, and inference time but compromise accuracy.

Data Quality. Data quality can be viewed as the degree to which a data set is fit for a particular purpose (Gudivada et al., 2017), i.e. analytical task. Missing, biassed or insufficient data are a common source of misclassifications, affecting the overall accuracy. Raw text is usually noisy, inconsistent, heterogeneous and has to be initially cleaned to reduce quality issues. Furthermore, classifiers are easily disrupted by an imbalanced data distributions, i.e. some classes occur much more often than others, causing a bias towards selecting the majority classes (Sun et al., 2009). In general, it is assumed that the more labelled data available, the better a model can be trained. However, it has been shown that equally high accuracies can be achieved with few but high quality data instances (Lewis and Gale, 1994). Overall, learning powerful classifiers requires good quality data.

Fuzzy Classification Objective. Text is naturally fuzzy, and its interpretation is highly subjective. Boundaries between different classification objectives are commonly fluid and cannot be sufficiently decided by black and white thinking. Ambiguous borderline cases will arise that cannot be decided objectively. Studies show that even experts disagree on the detection of hate speech (Waseem, 2016). It may happen that some texts appear as valid examples for several or all classes. In general, it is assumed that if domain experts cannot agree on a certain class membership, an algorithm will not be able to do better (Boguslav and Cohen, 2017).

3 FROM AUTOMATED TO HiL TEXT CLASSIFICATION

This section defines and scopes the HiL approach, and outlines common types of human feedback to support the collaboration between humans and text classifiers.

3.1 Scope and Definition

The phrase *human-in-the-loop* (HiL) (Holzinger, 2016) describes a computational ML paradigm and field of research characterized by the adaptation of machine processing by human skills, background knowledge and expertise. HiL systems aim to facilitate problem-solving with the cost of human involvement. At the time of the survey, there is no generally accepted definition for HiL. Many attempts have been made covering different aspects and use-cases of human-machine collaboration. Fails and Olsen Jr (Fails and Olsen Jr, 2003) were the first to use the term *interactive Machine Learning* (iML) to describe a continuous *train-feedback-correct* loop for interactively training a model. In their framework, humans continuously provide additional training data to a model until an acceptable level of accuracy is reached. This approach is also called *active learning* (Settles, 1995). Amershi et al. (Amershi et al., 2014) argue for the importance of an extended user-centric perspective in iML, focusing on human factors and the rapid and incremental nature of interaction cycles. They see iML as an opportunity for domain-experts to incorporate their knowledge directly into ML models. Another definition of HiL is provided by Holzinger (Holzinger, 2016). They use the term HiL to describe a concept which looks for *algorithms which interact with agents and can optimize their learning behaviour through this interaction*. This definition focuses mainly on the machine centred aspects of HiL, where models actively ask for feedback to support their behaviour during the learning phase. In general, human-machine cooperation can be both user- and machine-centred. In a machine-centred approach, a model asks the human directly for information. In a human-centred approach, humans select information themselves and make them available to classifiers. The later viewpoint is outlined by Dudley and Kristensson (Dudley and Kristensson, 2018) which define iML as "*a co-adaptive process, driven by the user, but inherently dynamic in nature as the model and user evolve together during training*". This definition focuses on the user and illuminates the process of knowledge generation that occurs during the progressive interaction between humans and the model. Humans gain insight and knowledge about their data by observing its structure and model results, while machines learn from human feedback (Sacha et al., 2014). The knowledge acquired can help to further improve the quality of subsequent feedback. HiL does not just focus on training. It has also been shown that involving domain-experts during inference can increase the accuracy of text classifiers (Kivlichan et al., 2021; Andersen and Maalej, 2022). Endert et al. (Endert et al., 2014) take a step further and advocate a "Human-*is*-the-Loop" methodology in exploratory settings to highlight the importance of seamlessly integrating human capabilities in the process of knowledge discovery. The application of HiL is not limited to model development. HiL can also be applied in deployment to further refine a model in the field. For the purposes of this survey, we define HiL as "a generic semi-automated process in which models and humans interact and learn from each other to improve the outcomes or applicability of ML algorithms."

3.2 Human Feedback

Users of HiL systems should not require a deep understanding of the model they are interacting with (Fails and Olsen Jr, 2003). Therefore, interactions have to concentrate on the exchange of domain-specific knowledge. In the following, we discuss different types of human feedback to support end-to-end text classification, where no manual text-processing and features-engineering is performed.

Feedback from domain-experts within text classification is mostly limited to label existing text documents to be added to the training data. The classifier is then re-trained on the extended training dataset and possibly improved (Lewis and Gale, 1994). Bernard et al. (Bernard et al., 2018) distinguish between two labelling scenarios to support the training process of a classifier. In **pre-labelling**, training data is collected to build the first batch of training data, which is used to initially train a classifier. In **incremental learning**, a model is re-trained when additional training data is available. In this case, humans continuously provide new labelled data to refine an already trained model. Re-training a model is important to strengthen its robustness and prevent it from deteriorating over time (Ovadia et al., 2019). Human labels can also be obtained by letting humans agree or disagree with artificially derived labels (Andersen et al., 2021). Domain-experts can also provide new text instances to support a classification model. Textual feedback can be used to reduce blind spots or misconceptions, such as providing missing evidence that is not included in the current training data (Attenberg et al., 2011). Here, users are asked not only to provide labels, but also to provide new or modified text examples.

4 ENABLERS

HiL aims for fast, efficient, continuous and beneficial interactions between ML models and humans. While artificial decision-making is considerably cheap and fast, human involvement is usually the bottleneck of HiL systems. To keep human involvement efficient and sparse, it is desirable to obtain high-quality feedback while avoiding redundant and unnecessary interactions. In the following, we discuss three general techniques that enable and support the exchange of high-quality feedback. These are **predictive uncertainties**, **explanations** and **visualizations**.

4.1 Predictive Uncertainties

Classification algorithms are inherently imperfect due to their probabilistic nature. Artificial predictions are corrupted by uncertainties which emerge during the classification process (Der Kiureghian and Ditlevsen, 2009). Misconceptions, corruptions, ambiguities, noise, a lack of evidence, limited representativeness, conflicting evidence within the training data, or out-of-distribution inputs might cause highly unreliable predictions which are probably wrong. Unfortunately, automated classifiers are incapable of recognizing when they fail to provide reliable outcomes. Thus, it is difficult for humans to reason about the reliability and trustfulness of predictions. In the worst case, a prediction is considered as correct even though it is not. Quantifying predictive uncertainties could help to handle these difficulties and is a first step towards more accountable and transparent predictions. Therefore, classifiers are required to additionally report uncertainty scores beside the usual class outcome, when a certain level of safety is needed. While uncertainty can arise and be passed on in any part of the ML pipeline and in human interaction with them (Sacha et al., 2015), recent research focuses on the automatic estimation of classification uncertainty in individual classification results (Blundell et al., 2015; Gal and Ghahramani, 2016).

Uncertainty is generally considered as a lack of confidence in a prediction (Li et al., 2012). It can also be seen as an indicator of unpredictability, indicating that instances contain much information the model might need (Lewis and Gale, 1994). Being aware of uncertainties helps to take special care of unreliable predictions and reduces the risk of trusting incorrect model behaviour, i.e., misclassifications (Hendrycks and Gimpel, 2017; Andersen and Maalej, 2022; Andersen and Zukunft, 2022). Uncertainty also facilitates the detection of out-of-distribution examples to which classifiers typically do not generalize well (Hu and Khan, 2021; Hendrycks and Gimpel, 2017). Interactions guided by uncertainty seek to spend human efforts most efficiently by focusing feedback on machine misconceptions and information needs.

4.2 Explanations

State-of-the-art classifiers are more complex and less interpretable than ever. Especially deep learning approaches are considered black-boxes since humans can not reason about their internal decision-making. The lack of transparency makes it challenging to understand how and why decisions were made. It stays unknown what the model has learned. Understanding the rationale for decisions would increase the confidence, trust, and applicability of artificial decisions. Especially since, users are usually not willing to apply artificial models when they do not trust them.

Explainable ML (Adadi and Berrada, 2018) aims to open the black-box of classifiers and ensure that humans can understand and justify why certain class results were delivered. Explanations enable us to discover what a model has learned and how to further improve it (Adadi and Berrada, 2018). Making artificial reasoning explicit helps humans to make their decision-making more efficient. Generally, explanation can enhance a classifiers robustness and user trust, knowledge transferability, as well as prevent faulty behaviour, weak points, undesired biases, unfairness, and discrimination (Arrieta et al., 2020; Confalonieri et al., 2021). Explanations are also used to provide better cognitive support and enhance the collaboration between humans and models.

In order to explain text classifiers, human comprehensible interpretations of classification results are needed. Explanations of classifiers are usually modelled as an additional output alongside the final class predictions. For deep learning based classification, typically *local* or *introspective explanations* (Confalonieri et al., 2021) are provided to explain input-output pairs based on a subset of input features that justify the classification result. For text classification, these are the words of the input text that contribute most to a particular class outcome.

4.3 Visualizations

Visual perception is one of the most important skills that enables humans to discover and understand local patterns and relationships in visual representations of problems statements (Tropmann-Frick and Andersen, 2019). Information visualization is about mapping data in a visual context so that it is easier for humans to understand and draw insights from it. Visual interfaces are essential for HiL since these enable domain-experts to cooperate with ML model without requiring any additional programming (Fails and Olsen Jr, 2003). Just displaying the labels, predicted by artificial classifiers, greatly improves human accuracy and speed of manually labelling (Desmond et al., 2021).

A common visualization technique for exploring large textual data is to embed their high-dimensional feature vectors, i.e. semantic meaningful text representations, into a typically two-dimensional vector-space using dimensionality reduction techniques (Benato et al., 2020). The reduced vectors can then easily be visualized via a scatter-plot.

5 GENERIC HiL FRAMEWORKS

In this section, we present three common HiL implementations that aim to overcome some major limitations in the application of automatic text classifiers.

5.1 Training Data Acquisition

While obtaining raw text instances usually is not a problem nowadays, the lack of labelled examples is the bottleneck of text classification. Generally, labelling must be done manually, which requires a lot of human labour. To reduce the effort required to manually label a sufficient training dataset is an important task in text classification. Approaches are needed to efficiently provide knowledge to classifiers.
Active Learning (AL). (Lewis and Gale, 1994; Settles, 1995) describes an incremental process in which a classifier accumulates knowledge by soliciting feedback from human annotators for the purpose of training. An actively trained model continuously improves its learning behaviour by querying human knowledge until the model reaches the desired accuracy. In the simplest case, a human is prompted over several iterations to specify the correct class labels for selected data instances. In several iterations, the potential training instances are ranked according to their uncertainty. Human annotators are asked to manually label instances that are believed to have the greatest impact on the model's learning behaviour. Then, the labelled examples are added to the training dataset and the model is re-trained. AL has proven to be very successful in various text classification tasks (Lewis and Gale, 1994). A general survey of AL is provided by Settles (Settles, 1995).

5.2 Moderation of Classification Outcomes

Highly accurate classifiers are required to adequately automate information retrieval. However, during training, the accuracy of text classifiers may converge to a level that does not meet the requirements of an application domain. Faulty and unreliable predictions are likely to occur, which might stay unnoticed. To improve the accuracy or prevent classification mistakes of an already trained classification model, humans can also be involved during inference post model training.

Classifier Moderation (CM). (Kivlichan et al., 2021; Andersen and Maalej, 2022) aims at increasing the applicability and reliability of an already trained model. Trained and deployed classifiers generally perform well, with only a small fraction of the data responsible for incorrect and unreliable model behaviour. CM seeks to maintain a superior level of accuracy by involving humans to prevent unreliable predictions as the model is used in practice (Karmakharm et al., 2019; Kivlichan et al., 2021; Andersen and Maalej, 2022). Humans are responsible for manually checking highly unreliable, i.e. uncertain, and fuzzy instances and correcting their labelling accordingly. If the model is certain of its prediction, no human involvement is required. Although not all misclassifications can be prevented (Attenberg et al., 2011), CM has the potential to lead to much better decision outcomes at the expense of human labour (Zhang et al., 2019; Andersen and Zukunft, 2022).

Since it might be impractical to let humans check all unreliable outcomes, Andersen and Maalej (Andersen and Maalej, 2022) suggest a saturation-based stop-criterion for CM. They aim to maximize the overall classification accuracy while spending human labour highly efficient. Pavlopoulos et al. (Pavlopoulos et al., 2017) suggest an approach which aims to maximize the performance of a classifier with respect to a given moderation effort, e.g. 10% of the data.

Geifman and El-Yaniv (Geifman and El-Yaniv, 2017) propose an approach to guarantee a certain risk level. Their approach is based on selective classification, where classifiers reject predictions which then have to be made by a human.

5.3 Interactive Labelling and Data Exploration

Interactive Labelling (IL). (Knaeble et al., 2020) is a user-centred variation of the AL process. Like AL, IL aims to reduce the number of labelled examples to adequately train a model. In contrast, humans are tasked with selecting the instances to be labelled. IL is based on the assumption that humans can select representative data instances more efficiently than automatic query strategies. For example, uncertainty sampling strategies are prone to sample outliers that contribute little or nothing to the learning behaviour (Lewis and Gale, 1994) of the models and tend to be overconfident (Guo et al., 2017).

To enable human-centred data sampling, the label query problem is reformulated into a *visual analytics* problem (Keim et al., 2008). Users are provided with adaptable visual-interactive interfaces to strategically select and label data instances. IL draws strength from extensive use of human expertise, background knowledge and visual perception. A key advantage over AL is that users generate additional and expanded knowledge and insights about their data through the exploratory nature of IL (Sacha et al., 2014). The accumulated knowledge can then further support the labelling process. As with AL, the model is interactively re-trained until a desired accuracy is achieved. Preliminary research show that IL can come close or even complement AL in terms of achieved accuracy (Bernard et al., 2018).

6 OPEN CHALLENGES

Previous studies demonstrate the usefulness of HiL text classification compared to a pure automatic analysis (Lewis and Gale, 1994; Kivlichan et al., 2021; Andersen and Maalej, 2022). However, the cost of using human labour is usually very high, whether in terms of money or time. Practitioners have to decide whether a pure automated approach is applicable and can solve a task appropriately, or whether a human in the loop is actually required and affordable. Further, human annotations should be taken with a grain of salt, as they can also be wrong. When human annotations are not reliable, biassed or too noisy, it negatively impacts the interaction between humans and machines (Andersen and Zukunft, 2022).

HiL approaches place special time requirements on the underlying classification model. Very short waiting times and few interruptions between human interactions are required, to maintain user experience. The need for fast interactions and model updates often makes it necessary to trade-off speed with accuracy (Amershi et al., 2014). Reducing iterative feedback latency is critical for HiL systems.

Also, uncertainties and explanations are not perfect and remain an active field of research. In particular, uncertainty estimation is challenging, especially using deep neural networks, since they do not provide an inherent indicator of uncertainty (Gal and Ghahramani, 2016). Inadequate measurements can inadvertently mislead humans into making a false assumption or blindly trusting artificial decision-making, e.g. unknown-unknowns (Attenberg et al., 2011).

7 CONCLUSION

Human-in-the-loop describes a collaborative process for improving the results and applicability of ML procedures through human feedback. We emphasize the need to involve domain experts in the text classification process in order to increase or even enable the applicability of machine-assisted text classification. We survey the current challenges in pure automated text classification and outline techniques to efficiently involve humans in the classification process. This includes the importance of uncertainty-based interactions to effectively guide humans providing feedback, building trust and focusing attention through explanations, and incorporating models into visual analytics environments. Additionally, we light on current human-in-the-loop implementations covering active learning, classifier moderation and interactive labelling.

REFERENCES

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Andersen, J. S. and Maalej, W. (2022). Efficient, uncertainty-based moderation of neural networks text classifiers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1536–1546.

Andersen, J. S. and Zukunft, O. (2022). Towards more reliable text classification on edge devices via a human-in-the-loop. In *International Conference on Agents and Artificial Intelligence 2022*, pages 636–646. SciTePress.

Andersen, J. S., Zukunft, O., and Maalej, W. (2021). Rem: Efficient semi-automated real-time moderation of online forums. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 142–149.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Attenberg, J. M., Ipeirotis, P. G., and Provost, F. (2011). Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Benato, B. C., Gomes, J. F., Telea, A. C., and Falcão, A. X. (2020). Semi-automatic data annotation guided by feature space projection. *Pattern Recognition*, 109(10761):2.

Bernard, J., Zeppelzauer, M., Sedlmair, M., and Aigner, W. (2018). Vial: A unified process for visual interactive labeling. *Vis. Comput.*, 34(9):1189–1207.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Boguslav, M. and Cohen, K. B. (2017). Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in health technology and informatics*, 245:298–302.

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., Finegan-Dollak, C., Brachman, M., Sharma, A., Joshi, N. N., et al. (2021). Increasing the speed and accuracy of data labeling through an ai assisted interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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: Human Language Technologies*, volume 1, pages 4171–4186.

Dudley, J. J. and Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37.

Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., and Andrews, C. (2014). The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435.

Fails, J. A. and Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., Wu, X.-C., Durbin, E. B., Doherty, J., Stroup, A., et al. (2021). Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Gudivada, V., Apon, A., and Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1):1–20.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Hu, Y. and Khan, L. (2021). Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 628–636.

Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Karmakharm, T., Aletras, N., and Bontcheva, K. (2019). Journalist-in-the-loop: Continuous learning as a service for rumour analysis. 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): System Demonstrations*, pages 115–120.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kivlichan, I., Lin, Z., Liu, J., and Vasserman, L. (2021). Measuring and improving model-moderator collaboration using uncertainty estimation. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 36–53.

Knaeble, M., Nadj, M., and Maedche, A. (2020). Oracle or teacher? a systematic overview of research on interactive labeling for machine learning. In *Wirtschaftsinformatik (Zentrale Tracks)*, pages 2–16.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Li, Y., Chen, J., and Feng, L. (2012). Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2463–2482.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment moderation. *ACL 2017*, page 25.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, 32.

Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., and Keim, D. A. (2015). The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249.

Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. A. (2014). Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Settles, B. (1995). Active learning literature survey. *Science*, 10(3):237–304.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719.

Tropmann-Frick, M. and Andersen, J. S. (2019). Towards visual data science-an exploration. In *International Conference on Human Interaction and Emerging Technologies*, pages 371–377. Springer.

Wang, Z. J., Choi, D., Xu, S., and Yang, D. (2021). Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, X., Chen, F., Lu, C.-T., and Ramakrishnan, N. (2019). Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3126–3136.

Zhao, P., Zhang, Y.-J., and Zhou, Z.-H. (2021). Exploratory machine learning with unknown unknowns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10999–11006.