Serendipitous Gains of Explaining a Classifier - Artificial versus Human Performance and Annotator Support in an Urgent Instructor-Intervention Model for MOOCs

Laila alrajhi
Durham University, Durham, UK, King Abdulaziz University, Jeddah, SA
laila.m.alrajhi@durham.ac.uk

Filipe Dwan Pereira
Federal University of Roraima, Boa Vista, Brazil
filipe.dwan@ufrr.br

Alexandra I. Cristea
Durham University, Durham, UK
alexandra.i.cristea@durham.ac.uk

Ahmed Alamri
University of Jeddah, Jeddah, SA
asalamri4@uj.edu.sa

ABSTRACT

Determining when instructor intervention is needed, based on learners' comments and their urgency in massive open online course (MOOC) environments, is a known challenge. To solve this challenge, prior art used autonomous machine learning (ML) models. These models are described as having a "black-box" nature, and their output is incomprehensible to humans. This paper shows *how to apply eXplainable Artificial Intelligence (XAI) techniques to interpret a MOOC intervention model for urgent comments detection.* As comments were selected from the MOOC course and annotated using human experts, we additionally study the confidence between annotators (*annotator agreement confidence*), versus an estimate of the class score of making a decision via ML, to support intervention decision. Serendipitously, we show, for the first time, that *XAI can be further used to support annotators creating high-quality, gold standard datasets for urgent intervention.*

CCS CONCEPTS

• **Natural language processing (NLP)**; • **eXplainable Artificial Intelligence (XAI)**; • **Instructor intervention model**;

KEYWORDS

MOOCs, Comments, Urgent intervention, Text classification, NLP, XAI, Annotator Support

ACM Reference Format:
Laila alrajhi, Filipe Dwan Pereira, Alexandra I. Cristea, and Ahmed Alamri. 2023. Serendipitous Gains of Explaining a Classifier - Artificial versus Human Performance and Annotator Support in an Urgent Instructor-Intervention Model for MOOCs. In *6th Workshop on Human Factors in Hypertext (HUMAN '23), September 04–08, 2023, Rome, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3603607.3613480

HUMAN '23, September 04–08, 2023, Rome, Italy

ACM ISBN 979-8-4007-0239-6/23/09.
https://doi.org/10.1145/3603607.3613480

1 INTRODUCTION

Instructor intervention in MOOC environments is a very challenging task. The proposed solution is to identify urgent comments that need intervention automatically, via ML. This is however a hard ML task, as urgency decisions are difficult, even for a human [1]. Also, the large number of observations may increase the cognitive overhead of annotators [2] (including physical consequences, such as making their vision blurry), where they struggle to make the appropriate decision.

Correct labelling, in natural language processing (NLP), annotating text data correctly, is a critical issue, and plays an important role in supervised ML model prediction. Therefore, and considering that urgency decisions had recently been confirmed to be hard for humans [1], we additionally evaluate the annotator agreement confidence in their decisions, and compare these with the model decisions on every instance of the 'learner comments'. We compared model decision with human decision making, using Captum [3] as interpretation tool, which is state-of-the-art for interpreting transformer models [4]. Thus, this work aims to explore and measure word attribution to predicted urgency cases, versus the confidence level of annotators. Therefore, we formalise our research question as:

- RQ: How can XAI be employed to improve human annotators' decisions about the urgency of comments (i.e., deciding on which comments need intervention)?

In terms of the contributions of this paper, to the best of our knowledge:

- This is the first time that the Artificial Intelligence (AI) prediction error has been shown to be connected to human (lack of) confidence (i.e. appearing for the same instances, here, comments).
- This is the first time where explainable models have been shown how they can be used for annotator support, for creating high standard corpora.

2 RELATED WORK

A literature review on intervention in MOOCs shows the area to have gained great momentum, proposing a variety of text classification models, to classify urgent comments. These models range from shallow ML [5] to deep learning [6] and transformers as embedding [7] with different level of inputs [8] but all those works did not

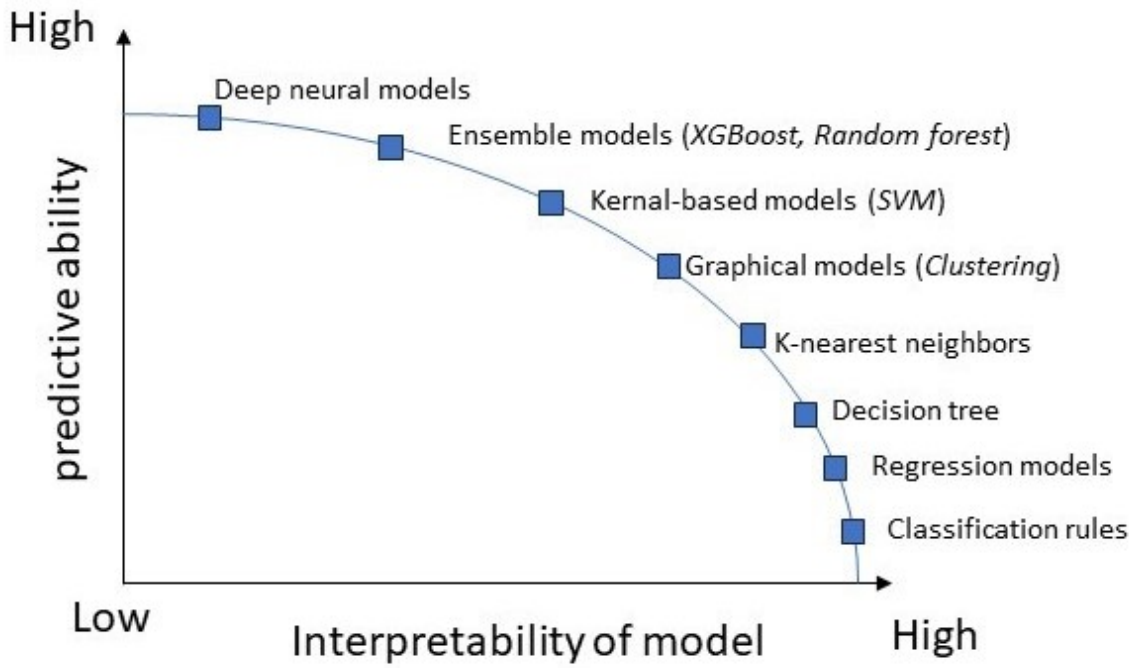

Figure 1: Predictive ability vs. interpretability trade-off [13].

interpret and explain the model's decision (excluding a brief work in [9]). Despite the fact that some shallow ML, such as classification trees and naïve Bayes algorithms, are simple to understand and interpret, they are less accurate than other approaches [10], especially for large data sets. Therefore, complex models, which achieve better performance have been proposed [6] [11] [12]. These models are considered as a 'black- box' and are consequently difficult to understand for the end-user, as depicted in Figure 1 [13], although some large models are interpretable [14].

Recently, outside of our specific area, a research direction has become very active and trendy: that aiming to explain and interpret 'black-box' predictions and ML models in general, for different sectors. Model interpretability is a field of explainable AI that attempts to explain model internals and results in human-understandable terms [15] [16]. A wide range of powerful tools have been proposed, such as the Local Interpretable Model-agnostic Explanations (LIME) [17], the SHapley Additive exPlanations (SHAP) [18], InterpretML [19] and Captum [3]. Explainability in AI is essential to developers, to understand and improve models, and also to end-users, to increase model-decision trust [20]. Please note that interpretability and explainability are often used as having the same meaning, but there are some papers that distinguish between them [21]. Here, however, for simplicity, we do not make that distinction.

The Bidirectional Encoder Representations from Transformers (in short, BERT) has been released at the end of 2018, and has been extremely popular, being applied in text classification models with high performance, such as in [22] [23]. Importantly for our current research, recent studies have proposed techniques for using XAI combined with BERT. Kokalj et al. [24] proposed TransSHAP (Transformer-SHAP) adapting and extending SHAP [18], to operate on BERT, by building custom functions and visualising the results in a sequential way. They demonstrated that the visualisation approach used on TransSHAP was simpler than that of other tools (LIME and SHAP). However, this approach is considered limited in terms of only supporting random word sampling, which may result in unintelligible and grammatically incorrect sentences, as well as wholly uninformative texts. Another study, Szczepański et al. [25], proposed a new approach for explainable BERT-based fake news detectors, using two XAI techniques (LIME and Anchors). They used the Kaggle dataset, and their findings support the use of multiple methods to construct explanations. However, there is a problem with Anchor, as it is not always being able to find an explanation.

On the other hand, Captum is an open-source multi-modal (image, text, audio or video) library for transformer model interpretability [3] [4]. It is an open-source library developed by Facebook AI and offers cutting-edge techniques, such as Integrated Gradients, that make it simple for researchers and developers to identify which features contribute to a model's decision and output [26]. This package has been drawing great attention and some researchers used this package in their applications. For instance, Levy et al. [27] utilised it to interpret a BERT model that was used as a one of different ML models, to predict current procedural terminology (CPT) codes from pathology reports.

Hence, we build in this research an explainable instructor intervention classifier model as a text classification task, deploying the Captum package, due to it being one of the most commonly used for transformer models [4].

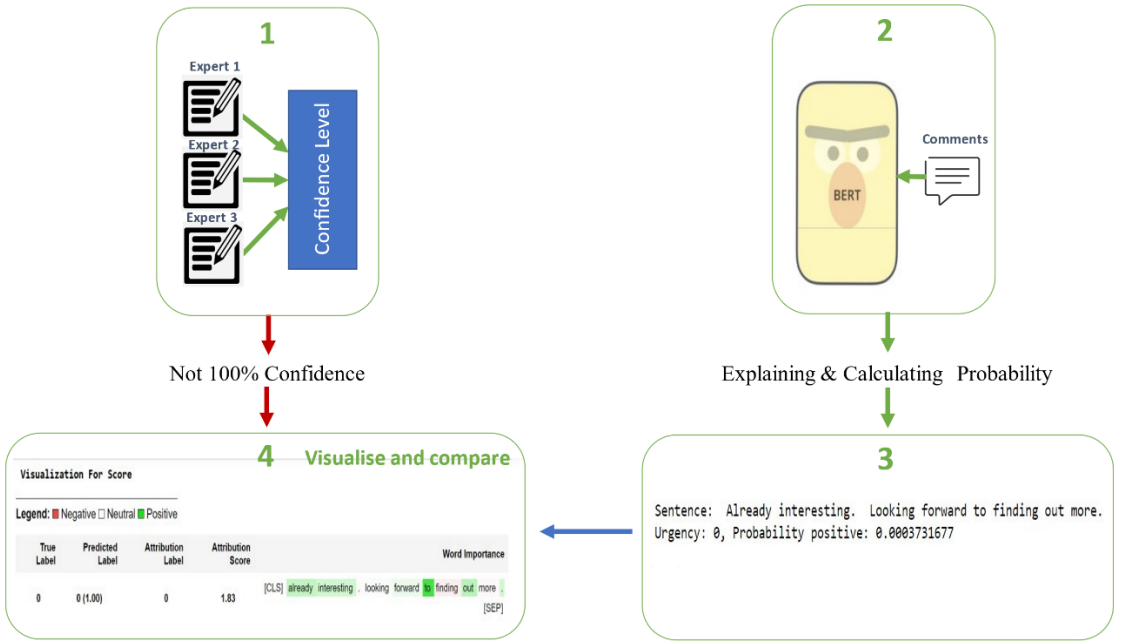

Figure 2: Human annotator vs. machine pipeline: basic stages.

The closest to our area is the work of [28], using XAI to analyse online discussion. However, no methods of XAI have been applied yet to urgent intervention, with the exception of a very recent and brief work in [9]. Moreover, *none of these works clearly explain how to connect AI prediction error to human (lack of) confidence*, or *use explainable models for annotator support, for creating high standard corpora.*

3 METHODOLOGY

This section summarises the methods used to generate and build our gold-standard corpus, with consideration of the measure of confidence between annotators, together with the tool and technique to explain the ML model. Thus, our research consists of four basic stages (see Figure 2) as follows:

1) First, construct an 'urgent' gold-standard dataset, via human experts annotating comments and computing their label confidence levels (section 3.1).
2) Second, via BERT, build an automatic urgent intervention model (section 3.2).
3) Third, automatically explain the model as a local explanation and calculate the probability and word attribution of urgency (section 3.3).
4) Fourth and last, we compare the two (machine and confidence), visualise and discuss (section 3.4, and section 4).

3.1 Building the Gold-standard Dataset

The data used was collected from the FutureLearn platform, the 'Big data' course, from the first 5 among 9 weeks. This was due to the fact that we wished to be able to catch issues and problems early on in the course. These comments were manually annotated by three experts of the domain, following the instructions of Agrawal and Paepcke [29] in urgency, over the scale (1-7) that represents ordinal values (from not urgent to very urgent) [9]. More details can be found in the online annotators instructions[1]. To briefly explain Agrawal and Paepcke [29] scale, we used the semantics as per their website, with: no reason to read the post → 1, not actionable; read if time → 2, not actionable; maybe interesting → 3, neutral: respond if spare time → 4, somewhat urgent: good idea to reply, teaching assistant might suffice → 5, very urgent: good idea for instructor to reply → 6, extremely urgent: instructor definitely needs to reply → 7. When we determined Krippendorff's [30] agreement value between all annotators for validation, we discovered that there was low agreement between any subgroups. To mitigate this problem and as for the current work we only needed a binary classifier (urgent comments for the instructor to read versus non-urgent: as the instructor can only read a comment or not, so a non-binary scale is not useful for an application scenario), we converted the scale to:

- 1,2 & 3 → 0 (do not read, not urgent enough).
- 4, 5, 6 & 7 → 1 (read, urgent messages).

We used voting between three annotators to determine the final label value, since voting is the most popular method of gathering different opinions for the same activity [31]. From this process we obtained:

- 4903 classified as '0' non-urgent;
- 883 classified as '1' urgent.

To add the annotator agreement confidence, we considered the three annotators' decision, after converting to a binary value. Therefore, we assigned (Figure 2, Step 1):

[1] CoderInstructions.docx(stanford.edu)

- If the three annotators agreed → 100% agreement confidence.
- Otherwise → not 100% agreement confidence (i.e. ~67% = 2/3 agreement).

3.2 The BERT Model (Fine-tuning)

The data as above was split using the stratify method [32] into training (80%), where the distribution of the training set is (0: 3922, 1: 706) and testing (20%), with distribution (0: 981, 1: 177). The training set was then split again, as 90% for training and 10% as validation.

We fine-tuned BERT, by using the 'bert-base-uncased', to train a text classifier to classify comments as *urgent* or *non-urgent*. We used the 'bert-base-uncased' version, then we trained the model, by set: batch size = 8, epochs = 4 and optimiser = AdamW, with learning rate = 2e-5 (Figure 2, Step 2).

3.3 Explaining the BERT Model

The BERT model is next automatically explained using the Captum package. From Captum_BERT colab [33], BertForSequenceClassification was applied. This is done by creating the Layer Integrated Gradients explainer and attribute method, to generate feature importance and identify which words (tokens) have the highest attribution to the model's output. As based on the gradient of the model's output (prediction) with respect to the input, integrated gradients [34] are a way to calculate the attribution score of each input feature of a deep learning model (here, BERT). This attribution score can be used to determine which words are important to the outcome that our model predicts. The final attribution score is calculated by the average value for each word (Figure 2, step 3).

In this experiment we also inspect 3 comments without 100% confidence, which means the final label was set by majority voting (2/3, with one annotator disagreeing). The selected comments are selected to showcase 3 scenarios reflecting the differences on agreement between human annotators: 1: large difference; 2: slight difference; 3: in-between. This will be further clarified in section 4.

3.4 Visualising and Comparing

The final step is to visualise the explainability results with the attribution score and highlight the word importance, as input for human consumption and potential future decision support. The visualisation is done by using VisualizationDataRecord method. Green highlights are used to indicate the tokens that contribute positively to the model's prediction. While red highlights are placed on the tokens that have a negative impact on the model's prediction (see Figure 2, step 4).

4 RESULTS AND DISCUSSIONS

At the beginning, we calculate the agreement between the three annotators, as previously explained, and we found that: the number of comments that have 100% confidence between annotators is 4190; and, on the other hand, the number of comments without 100% confidence is 1596 from the total data. From the test data, the total number of comments among annotators with 100% confidence is 833. However, there were 325 comments without 100% confidence.

We use different metrics to evaluate the BERT classifier, average accuracy and (precision, recall and F1-score) for every class, for a comprehensive understanding of the outcomes, as in Table 1. Please note that BERT has been selected here as it is one of the state-of-the-art classifiers; however, *the method of explainable decision about urgent comments for instructors, and comparing machine prediction to the human classification, is generalisable, and can be used with other deep learning models.*

Figure 3: Confusion matrix of the BERT classifier.

These measurements are based on the confusion matrix, which is depicted as a table, with 4 different combinations of predicted and true values: 'TN', 'FP', 'FN' and 'TP' standing for true negative, false positive, false negative and true positive, respectively, as the results from the BERT classifier reported - see Figure 3.

We analyse, as said, the relation between machine results and the confidence agreement level between human annotators. From the confusion matrix we can study different cases, as shown in Table 2.

The aim of this research is to help annotators to find urgent cases that BERT can classify as urgent. Thus, we focus on true positive (TP), especially case 2, where confidence between human annotators is not 100%. The reason for focusing on TP is that these are the ones we would like both the algorithm and the annotators to find. There are 126 TP cases, as we reported in Figure 3. The experiment and investigation results, interestingly, show that, for 79 out of the 126, the classifier and the annotators agree that the comments need urgent intervention with confidence level = 100%. For the rest 47 cases (case 2), we found that the confidence level between annotators is not 100%. Therefore, (47 cases) need explanation and visualisation to the annotator who disagrees with the other two annotators, to potentially change their mind. In addition, FP, where a comment is considered by the algorithm (BERT) as urgent, but not by the annotators, may be a potential issue, if the label should be 'True' but is not. The number of FP is 47 with 17 cases with confidence level = 100% and 30 cases with confidence is not 100% (case 4). That means that at least one of annotators from 30 cases believes it is urgent, similar to BERT. Thus, these are the cases that should be explained and shown to annotators, especially with the highlights illustrating the reason of BERT's decision, to help them refine theirs. In general, however, any of the comments where annotators disagree could potentially be reinspected by annotators, to ensure that they increase their confidence.

Next, we inspected some of these comments from TP, to interpret the probability of predicting the urgency by the classifier and to understand if and how this may be related to the disagreement between annotators.

To better understand in-depth our findings, we consider three scenarios, based on the agreement between human annotators, as shown in Table 3; we selected these three cases, according to level

Table 1: The results of the BERT classifier

Class	AverageAccuracy	Precision	Recall	F1-score
0	.92	.95	.95	.95
1		.73	.71	.72

Table 2: Machine prediction correctness (from the BERT confusion matrix), vs. human annotator classification correctness, with (binary) confidence between (human) annotators and number of comments for each case, Bold/Italics: cases that should/could be explained to annotators

Cases	True class (human annotators)	BERT confusion	BERT prediction	Confidence between human annotators	Number of comments
1	1	TP	1	100%	79
2	**1**		**1**	**Not 100%**	**47**
3	0	FP	1	100%	17
4	**0**		**1**	**Not 100%**	**30**
5	1	FN	0	100%	17
6	*1*		*0*	*Not 100%*	*34*
7	0	TN	0	100%	741
8	0		0	Not 100%	193

Table 3: Three scenarios based on TP with agreement between human annotators: 1: large difference; 2: slight difference; 3: in-between

#	Text	Firstannotator	Second annotator	Thirdannotator
1	What's the shortcomings of the crowdsourced data? It's hard for me to understand.	1	6	7
2	I wonder if there is also a correlation between future orientation index and GDP per capita when the search terms are two years ahead and two years before (e.g. search terms "2009" and "2013" in the search year 2011).	3	4	4
3	I am seeking to build the following: 1-Multiple big data DB on VPS over the internet so they will be MySQL on CentOS. 2-Multiple DB manipulation engines that will Read or write on them from the BigData source. 3-Multi-agent simulations on a given basemap 4-The data model running on the simulation and DB manipulation engines will be an XML based model. Can anyone help me to build this environment?	2	6	4

of classification (large difference, slight difference or in between). Please note that in Table 3 we show the rating from annotators before converting to binary as urgent or non-urgent, to understand their real decision.

4.1 Scenario 1 (large difference)

In this scenario, we observe that some comments lead to large differences between annotators. Therefore, we interpret the model and highlight the important words, as shown in Figure 4. We can see that the attribution score = 2.13, which is high; the words 'hard for me' are the important words that affect the decision. Thus, it may draw the attention of the annotator, to lead them to the correct decision.

4.2 Scenario 2 (slight difference)

In this scenario, the agreement is strong on being a threshold case (between urgent and not non-urgent). When visualising (Figure 5), we find that the words 'I wonder if' are important. The meaning of 'wonder' involves asking for help, but also just thinking. Thus, this scenario can be used as a confirmatory analysis for annotators.

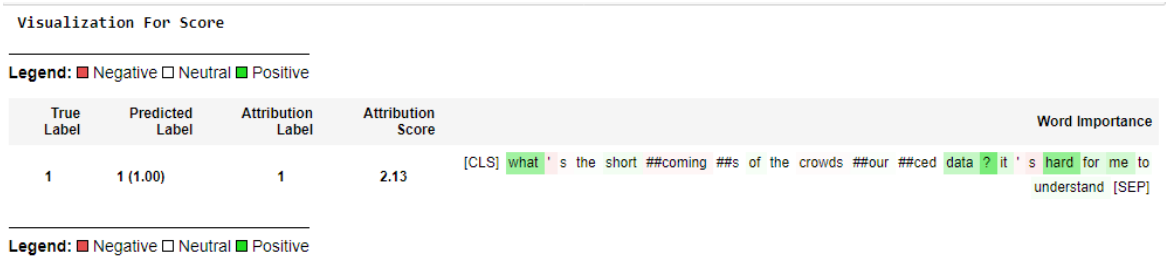

Figure 4: Screenshots of Captum explanations for scenario 1 (large difference and 'not 100% confidence' between annotators).

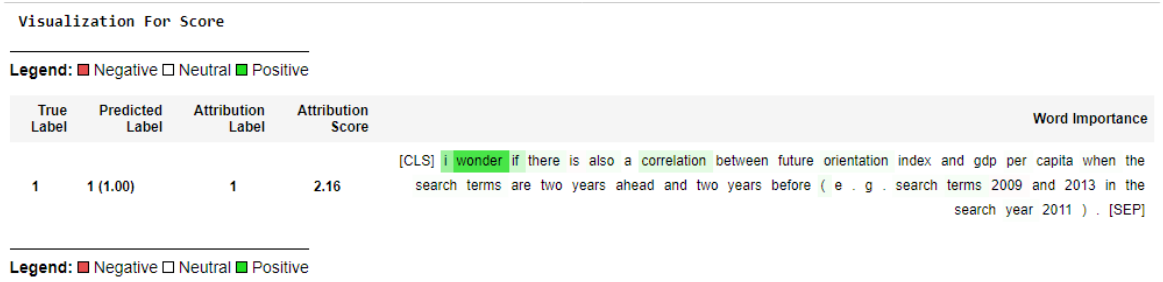

Figure 5: Screenshots of Captum explanations for scenario 2 (slight difference and 'not 100% confidence' between annotators).

Figure 6: Screenshots of Captum explanations for scenario 3 (in between and 'not 100% confidence' between annotators).

4.3 Scenario 3 (in between)

In this scenario, the score is incremental (2, 4, 6). To understand it, visualisation (see Figure 6) shows that ('can', 'anyone', help', '?') are important words for the algorithmic decision. The annotator difference may be due to some annotators considering that, by using the word 'anyone', the learner asks for help from their peers, not the instructor.

4.4 Discussion

Finding urgent messages is vital for instructors in online courses. However, it is a daunting prospect for instructors in MOOCs, due to the sheer volume of the comments. Thus, classification models for automatically analysing comments and predicting their urgency are stringently needed, and some accurate models have been proposed in the past [6] [7] [8]. However, the story does not end here. Whilst some of these models have cutting-edge performance, however, just increasing performance may not be enough as with any black-box system, explainability is key and related to trust in the system. Moreover, correct labels are key. However, the type of comments appearing in a learning system are hard even for experts to reliably classify, as our experiments with annotators show. This further supports the addition of (automatic) explanations to the recommendation, to better contextualise the information presented for annotators. Indeed, highlighting the most important words could facilitate the annotators work on deciding and finding whether a comment is urgent. That is, with our method, we can facilitate the annotators work, leading to further improving the dataset annotation process.

Here, thus, we use explainable AI in a different, novel way: we turn it around, to explain to us not the errors in the algorithm, but the errors in human annotation (which may well lead to or explain

errors in the algorithm[2]). Thus, we analyse three extreme scenarios, based on true positives (TP). As explained, these are the cases we really want to deal with and help the disagreeing annotator to check the decision, to increase the quality of the dataset.

We show how colour-based highlighting functionality of explainable AI can give us an in-depth understanding of where the different answers of the annotators, as well as the algorithm, may stem from. Thus, we found serendipitous gains, in automatic explanations of annotators. Such systems could support thus annotators, in facilitating/fast-tracking their work in detecting interventions points and also, bringing them to a common denominator, and helping them make informed decisions on a sample of already-labelled data, to then be able to confidently label new, unseen data, in a rigorous and systematic way.

5 CONCLUSION

We aimed on this paper is to evaluate the annotators' agreement confidence, obtained from labelling for the urgent instructor intervention task in a MOOC environment. In particular, we would like to highlight the contribution of explaining individual predictions in the urgent intervention task and assessing annotators decisions, when they label the comment corpus. We have presented a BERT model to classify urgent comment cases. To better understand what causes the errors we have made the interesting discovery of the relation between the ability of the classifier to find urgent cases, and the confidence between human annotators on making a decision in labelling data. Moreover, we are offering a new method for supporting annotators.

REFERENCES

[1] Chandrasekaran, M.K., *et al.*, *Learning instructor intervention from mooc forums: Early results and issues.* arXiv preprint arXiv:1504.07206, 2015.

[2] Dong, H., *et al.*, *Automated social text annotation with joint multilabel attention networks.* IEEE Transactions on Neural Networks and Learning Systems, 2020. 32(5): p. 2224-2238.

[3] Kokhlikyan, N., *et al.*, Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896, 2020.

[4] Bennetot, A., *et al.*, *A Practical Tutorial on Explainable AI Techniques.* arXiv preprint arXiv:2111.14260, 2021.

[5] Almatrafi, O., A. Johri, and H. Rangwala, *Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums.* Computers & Education, 2018. **118**: p. 1-9.

[6] Guo, S.X., *et al.*, Attention-Based Character-Word Hybrid Neural Networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums. IEEE Access, 2019. **7**: p. 120522-120532.

[7] Khodeir, N.A., *Bi-GRU urgent classification for MOOC discussion forums based on BERT.* IEEE Access, 2021. **9**: p. 58243-58255.

[8] Alrajhi, L. and A.I. Cristea. Plug & Play with Deep Neural Networks: Classifying Posts that Need Urgent Intervention in MOOCs. in International Conference on Intelligent Tutoring Systems. 2023. Springer.

[9] Alrajhi, L., *et al.* A Good Classifier is Not Enough: A XAI Approach for Urgent Instructor-Intervention Models in MOOCs. in International Conference on Artificial Intelligence in Education. 2022. Springer.

[10] Kowsari, K., *et al.* Hdltex: Hierarchical deep learning for text classification. in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017. IEEE.

[11] Sun, X., *et al.* Identification of urgent posts in MOOC discussion forums using an improved RCNN. in 2019 IEEE World Conference on Engineering Education (EDUNINE). 2019. IEEE.

[12] Alrajhi, L., K. Alharbi, and A.I. Cristea. A Multidimensional Deep Learner Model of Urgent Instructor Intervention Need in MOOC Forum Posts. in International Conference on Intelligent Tutoring Systems. 2020. Springer.

[2]Please note that we do not consider here the algorithm beyond errors. However, being able to compare their own decision against the algorithm may give the human additional insight to revise (some of) their opinions.

[13] Kumar, A., S. Dikshit, and V.H.C. Albuquerque, *Explainable artificial intelligence for sarcasm detection in dialogues.* Wireless Communications and Mobile Computing, 2021. **2021**: p. 1-13.

[14] Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019. **1**(5): p. 206-215.

[15] Gilpin, L.H., *et al.* Explaining explanations: An overview of interpretability of machine learning. in 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). 2018. IEEE.

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

[17] Ribeiro, M.T., S. Singh, and C. Guestrin. " Why should i trust you?" Explaining the predictions of any classifier. in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[18] Lundberg, S. and S.-I. Lee, *A unified approach to interpreting model predictions.* arXiv preprint arXiv:1705.07874, 2017.

[19] Nori, H., *et al.*, *Interpretml: A unified framework for machine learning interpretability.* arXiv preprint arXiv:1909.09223, 2019.

[20] Confalonieri, R., *et al.*, *A historical perspective of explainable Artificial Intelligence.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021. **11**(1): p. e1391.

[21] Došilović, F.K., M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. in 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). 2018. IEEE.

[22] Fonseca, S.C., *et al.*, *Automatic subject-based contextualisation of programming assignment lists.* International Educational Data Mining Society, 2020.

[23] Pereira, F.D., *et al.* Towards a Human-AI Hybrid System for Categorising Programming Problems. in Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. 2021.

[24] Kokalj, E., *et al.* BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. in Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021.

[25] Szczepański, M., *et al.*, *New explainability method for BERT-based model in fake news detection.* Scientific reports, 2021. **11**(1): p. 1-13.

[26] Captum. *Captum - Model Interpretability for PyTorch.* 2021; Available from: https://captum.ai/docs/introduction.html.

[27] Levy, J., *et al.*, *Comparison of machine-learning algorithms for the prediction of current procedural terminology (CPT) codes from pathology reports.* Journal of Pathology Informatics, 2022. **13**: p. 100165.

[28] Hu, Y., R.F. Mello, and D. Gašević, Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence. Computers and Education: Artificial Intelligence, 2021. **2**: p. 100037.

[29] Agrawal, A. and A. Paepcke. *The Stanford MOOCPosts Data Set.* 2020 27/1/2020]; Available from: https://datastage.stanford.edu/StanfordMoocPosts/.

[30] Hayes, A.F. and K. Krippendorff, *Answering the call for a standard reliability measure for coding data.* Communication methods and measures, 2007. **1**(1): p. 77-89.

[31] Troyano, J.A., *et al.* Named entity recognition through corpus transformation and system combination. in International Conference on Natural Language Processing (in Spain). 2004. Springer.

[32] Farias, F., T. Ludermir, and C. Bastos-Filho, *Similarity Based Stratified Splitting: an approach to train better classifiers.* arXiv preprint arXiv:2010.06099, 2020.

[33] Captum. *Captum_BERT.* 2022; Available from: https://colab.research.google.com/drive/1pgAbzUF2SzF0BdFtGpJbZPWUOhFxT2NZ.

[34] Sundararajan, M., A. Taly, and Q. Yan. *Axiomatic attribution for deep networks.* in *International conference on machine learning.* 2017. PMLR.