More than a label: machine-assisted data interpretation

Maja Trębacz [1] **Luke Church** [1]

Abstract

We motivate and describe a prototype that presents an alternative view on how labelling of data can be done, with the goal of not only efficiently attaching labels to the data, but also supporting a researcher gaining an understanding of the data in the process of labelling.

1. Introduction

The automatic assignment of categorical labels to data has been one of the key capabilities introduced by advances in machine learning systems. These labels can then be used to support a range of activities, for example, spam detection, credit scoring or skin cancer identification (Guzella & Caminhas, 2009; West, 2000; Esteva et al., 2017).

Before the use of machine learning, labelling had to be done as a manual activity, a form of 'informational work' often done to 'sort out' (Bowker & Star) complex contexts so that bureaucratic and corporate infrastructures could work with them. Whilst machine learning has vastly reduced the labour of such activities, it has come with a number of costs. Contemporary discussions of the ethics of machine learning include the bias introduced into the training sets, and the respect (or lack of) for intellectual property and privacy norms associated with the data.

However, apart from significant concerns around the labour conditions of the workforce performing the labelling activity (Fort et al., 2011), comparatively little attention is given to the process of labelling itself. This is a missed opportunity: If the process of labelling is a form of interpretative abstraction, then the people doing it will have learnt things about the data during the process. The paper asks the question: How can we go about support this learning process, such that it takes advantage of, but is not replaced by, an automated labelling system?

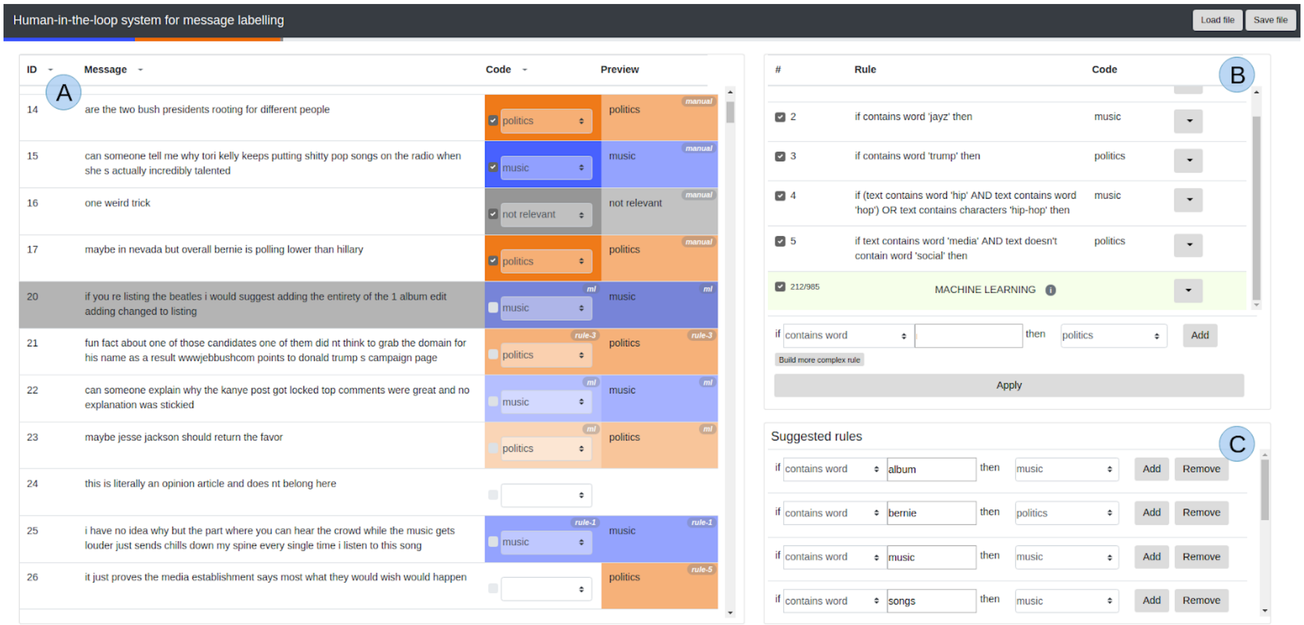
This question is motivated by work undertaken at Africa's Voices Foundation (AVF). AVF is a UK and Nairobi-based, non-profit that engages citizens in Kenya and Somalia to understand their perspectives and represent them to authorities. AVF, working with partner organisations produces radio shows. As part of these shows, the audience is asked open-ended questions and encouraged to send free-text replies via SMS. Some of these replies are subsequently read out on air, to maintain the discussion about a topic. People who text in are sent follow on questions asking about demographics, and for further opinions. In contrast with traditional surveys, the method does not pre-frame the answers allowing for surprise and nuance in the data, but shift the burden of interpretation - making sense of what is being said - to the researchers at AVF. For example in a recent show the question "What is your community doing to help the most vulnerable during coronavirus?" was asked, resulting in 19,177 messages[1].

In order to see macro-trends in the data, the datasets need to be labelled by AVF's researcher who is fluent in the local languages. Historically this has been done manually, first in spreadsheets and then in a tool called Coda (Church et al., 2018). Attempts to use fully automated, machine learning based approaches, have not proven successful due to the nature of the labelling, the languages in use and the need for confidence in the analysis. However, as well as not being practical, the use of machine learning also misses the broader point that it is through the reading and interpretation of the messages that the researchers gain insight into the dataset that allows it to be interpreted in a way that is contextually relevant. Developing this insight is crucial, both to shape future shows, and to represent the audiences' opinions to policymakers.

This project attempts to get the best of both worlds using a hybrid approach. It augments user actions with the abilities of a machine learning (ML) classifier and end-user programming to allow for large datasets to be labelled, whilst supporting the researchers developing an understanding of the dataset. In doing so, it builds on the paradigm of Human-In-The-Loop Systems (Amershi et al., 2014; Fails & Olsen Jr, 2003; Zanzotto, 2019)

[1] Department of Computer Science and Technology, University of Cambridge, UK. Correspondence to: Maja Trębacz <trebacz-maja@gmail.com>, Luke Church <luke@church.name>.

[1] The analysis of the results is listed at https://www.africasvoices.org/case-studies/covid19-kenya-trusted-two-way-mass-and-individual-health-communications-and-rapid-socio-epidemiological-insights/



ID	Message	Code	Preview
14	are the two bush presidents rooting for different people	☑ politics	politics manual
15	can someone tell me why tori kelly keeps putting shitty pop songs on the radio when she s actually incredibly talented	☑ music	music manual
16	one weird trick	☑ not relevant	not relevant manual
17	maybe in nevada but overall bernie is polling lower than hillary	☑ politics	politics manual
20	if you re listing the beatles i would suggest adding the entirety of the 1 album edit adding changed to listing	☐ music ml	music ml
21	fun fact about one of those candidates one of them did nt think to grab the domain for his name as a result wwwjebbushcom points to donald trump s campaign page	☐ politics rule-3	politics rule-3
22	can someone explain why the kanye post got locked top comments were great and no explanation was stickied	☐ music ml	music ml
23	maybe jesse jackson should return the favor	☐ politics ml	politics ml
24	this is literally an opinion article and does nt belong here	☐	
25	i have no idea why but the part where you can hear the crowd while the music gets louder just sends chills down my spine every single time i listen to this song	☐ music rule-1	music rule-1
26	it just proves the media establishment says most what they would wish would happen	☐	politics rule-5

#	Rule	Code
☑ 2	if contains word 'jayz' then	music
☑ 3	if contains word 'trump' then	politics
☑ 4	if (text contains word 'hip' AND text contains word 'hop') OR text contains characters 'hip-hop' then	music
☑ 5	if text contains word 'media' AND text doesn't contain word 'social' then	politics
☑ 212/985	MACHINE LEARNING	

if contains word then politics Add

Build more complex rule

Apply

Suggested rules

if contains word album then music Add Remove

if contains word bernie then politics Add Remove

if contains word music then music Add Remove

if contains word songs then music Add Remove

Figure 1. Screenshot of the user interface. The Message Table (A) shows the individual messages in rows and the applied labels in the Code column. The Rule Table (B) provides interface for defining and managing an ordered set of custom rules and the ML classifier. The Suggested Rules Panel (C) proposes rules produced by the statistical inference.

2. User interface

The tool supports multiple styles of interaction. The researcher can label the individual messages manually, review the machine learning predictions, define custom rules for labelling, and review the rules generated by the system. Figure 1 presents the interface that consists of three tables supporting these interactions: Message Table (A), Rule Table (with a special machine learning row) (B) and Suggested Rules Table (C).

2.1. Message Table (A)

The Message Table displays the messages that are already labelled or need to be coded. To label a message the user can click on the dropdown arrow and select the option from the list, or using keyboard shortcuts. The label will apply to the currently selected row, colouring the cell according to the label, and the selection will advance to the next row.

To support staged decisions, the predictions suggested by the rule and ML components are displayed in the preview column. The confidence score of the prediction is indicated by the shading and the source is displayed in a badge in the top-right corner (e.g. showing rule id). In order to accept the preview, the researcher can click the ‘Apply’ button in the rule table. Then the predictions will be transferred to the ‘Code’ column. Then the researcher can verify the predictions by ticking the checkbox on the label or using the Enter key.

2.2. Rule Table (B)

The researcher can define simple rules based on word presence (“if contains *word* then *label*"), the occurrence of a sequence of characters, message length, or regular expressions. The rules can be combined with binary operators. The rules are presented in a table (B). The ordering of the rules in the table reflects the application order, the first matching rule is applied.

2.3. Machine Learning

Machine Learning classifier is expressed as a special row in the rule table. It becomes available once the number of manually labelled messages reaches a boundary. The predictions are applied only to the messages for which the confidence of the model is above a predefined threshold. Including the ML component component within the rule table reduces the complexity of the system and provides the user with freedom of changing the application order. That is, the ML predictions can be applied after, before or between the user-defined rules.

The system uses Naive Bayes classifier (Maron, 1961). Such a simple model was used to explore interaction modality rather than the performance of Machine Learning system. The classifier can be replaced with another model of choice.

2.4. Suggested Rules Panel (C)

The system automatically generates appropriate classification rules in the Suggested Rules Panel (C). The rule induction is based on the information gain calculation phase in the ID3 decision tree induction algorithm (Quinlan, 1986; Johnson et al., 2002). The researcher may accept, reject or edit the proposed rules.

3. User study

To determine whether the interaction supported by the tool represented a viable alternative to manual or fully automated labelling we performed a small pilot study with 10 student participants 5 of whom had an existing experience in programming. They ranged in age from 19 to 23. We used quantitative performance measures in a controlled experiment to assess whether they were significantly more productive using the system compared to a spreadsheet, and qualitative and experience reports to investigate whether they gained an interpretative insight into the data they were labelling.

3.1. Experimental tasks

The participants were asked to perform classification labelling using different versions of the system. They were given four pairs of tasks.

In Task 1, the participants were classifying Reddit comments (Qiu, 2016) from categories ‘music’ versus ‘movies’. Participants performed the categorisation once in LibreOffice Calc, and once using the whole version of the proposed system. Task 2 involved manual labelling of an SMS spam detection dataset (Almeida et al., 2011), once in Calc and once in the tool with the interface only the manual labelling facilities. Task 3 involved sentiment classification of Amazon reviews (He & McAuley, 2016). Participants performed the task once in a full version of the system, and once with the functionalities limited to manual labelling and defining rules. Task 4 involved the classification of short messages based on their source of Reddit vs Twitter. Participants performed the task in a full version of the system, and once with the functionalities limited to manual labelling and verifying ML predictions (without defining custom rules). The time limit was 3 minutes for each subtask in tasks 1 and 2, and 5 minutes in tasks 3 and 4. The subtasks were presented to the participants in a different order among the control groups. Each subtask had a random stratified sample of 1000 messages, out of which 200 were already labelled and used to train the initial ML model that was updated as an effect of labelling.

3.2. Quantitative Results

Task 1 - Spreadsheet vs whole tool: The number of manually labelled messages was not significantly different between the proposed tool and spreadsheet. During this time, in the proposed tool, the participants were also writing rules and providing data for training an ML model, both result in a statistically significant increase in the number of labelled messages, but these labels haven’t been manually confirmed.

Task 2 - Spreadsheet vs manual labelling: On average, the participants managed to label 109.5±31.3 messages when using the manual labelling in the tool compared to 68.5±22.4 in a spreadsheet ($p<0.05$). This result demonstrates the value of designing a custom application that optimises the user interface for the specific task, improving the efficiency and ease of the manual mode of interaction.

Task 3 - Rules and manual labelling vs whole tool: In the limited version of the system, the participants categorised and confirmed an average of 47.0±12.6 messages, compared to 62.8±17.9 messages in the whole tool with ML classifier ($p<0.05$). This results suggest that the inclu-

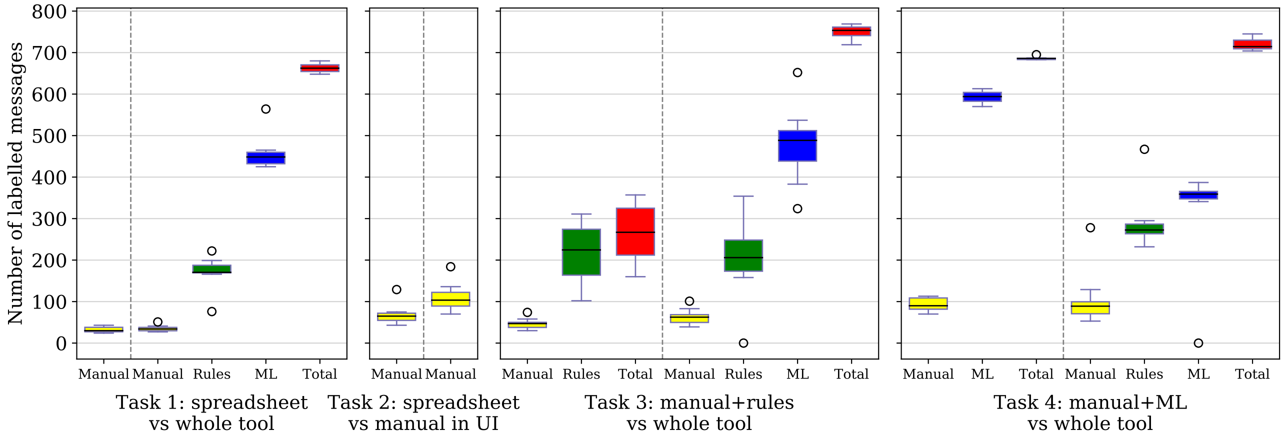


Figure 2. Box and whiskers plots showing the number of messages labelled using different styles of interaction. The yellow boxes show data from manual labelling (and manually confirming predictions). The green boxes show data from automatic labelling with user-defined rules. The blue boxes show data from automatic labelling with an ML classifier. The red boxes show the total of all types of labelling.

sion of ML predictions improved the users' productivity and allowed to categorise more messages by performing the verification task rather than traditional labelling combined with defining rules. The addition of the automated classification performed by the ML algorithm did not reduce the coverage of the rules inserted to the rule table.

Task 4 - Inference and manual labelling vs whole tool: In the version of the system limited to manual labelling and ML, participants categorised on average 93.6±14.6 messages by manual labelling and reviewing predictions. The ML module classified further 592.3±13.43 messages with the accuracy of 94%. In the full version of the system (with components for defining rules) participants manually labelled on average 104.6±61.52. They also defined on average 3.6 rules that applied to 289.1±61.72 messages labelling them with 99% accuracy. The ML module classified further 325.4±109.3 messages.

3.3. Discussion of quantitative results

The results demonstrate that the application of the custom tool, the rules and a simple machine learning system, lead to an increased productivity over manual labelling in a spreadsheet. As well as offering these benefits, the creation of both rules and a trained model represent investments allowing labelling of future unseen data.

These results are not surprising in themselves, what they demonstrate is that the whole system has a level of productivity that supports labelling a substantial amount of data. This opens the question of 'does the tool support the researchers in building an interpretative understanding to the data' which we investigated using a qualitative approach.

3.4. Qualitative findings

After finishing the tasks, participants took part in a semi-structured interview. The transcripts of these interviews, together with the recording of their use of the tool were reviewed by the authors for evidence that indicated or counter-indicated the formulation of an understanding of the data by the participants.

The small pilot study, with participants who were not ideally representative of the eventual user population imposed a number of limitations to the external validity of the task, we'll outline the tentative results and then discuss them in the context of the limitations.

Observations of the participants in the study suggested that they where building an understanding of the data whilst performing the tasks. In the fourth task (Twitter vs Reddit), 8 out of the 10 participants defined a rule that "if text *contains characters '#'* then *twitter*" and 10 out of 10 defined rule "if text *contains characters '@'* then *twitter*". These two rules alone result in a 41% coverage of the dataset with 100% accuracy.

As well as the discovery of this rule, P8 recognised that the dataset used in the second task was sourced from a Singaporean context, observing the prevalence of 'leh' and 'lor', common words in colloquial Singlish. This was an observation about the provenance of the data that the authors were not aware of, and had to verify, but was indeed the case. This suggests that the interaction modality encourages substantive engagement with the content of the data.

Participants also expressed a preference for their choice of interaction mode. P4 completed task 4 by spending their time examining the data to build a list of 8 rules and then confirming their responses. This strategy resulted in an accuracy of 92.6%, comparable to the accuracy of the an ML system trained on 200 examples (giving 92.1% accuracy).

P6 reported that they found the task of verifying a label to be easier than the performing the full interpretation: "I found that the machine learning module is really useful because I'm already primed for an answer. Such that if I see the answer already being 'negative' I'll more easily find negative parts in it such as 'hate' or 'didn't like'. And I can make the decision much faster than when I'm looking at a blank space.". This self reported behaviour leads to further questions about the risks of the participants exhibiting confirmation bias, and to what extent that inhibits the interpretation they are performing.

These results are encouraging for demonstrating engagement with the data. Due to the context and short duration of the study, the tasks only needed shallow, primarily syntactic interpretation. The study was also conducted with students, several of whom are completing degrees in STEM subjects, rather than the intended users experienced social science researchers. Taken together, these limitations prevent us from making a strong claim about the benefit of the tool in context but we suggest that the results indicate the possibility that such a tool can efficiently support labelling whilst also enabling the researchers using it to build an interpretative understanding of the data.

4. Conclusion

This investigation has explored the potential of a different perspective on the configuration of systems, that machine learning tools should be considered as systems of augmentation of human capabilities as opposed to automated replacement. The prototype shows a design strategy of enabling multiple styles of interaction with data. This, together with preliminary results, demonstrates that such an approach is plausible and may result in substantial benefits over either fully manually or fully automated labelling in contexts where understanding the data, not merely categorising it, matters.

References

Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, pp. 259–262. ACM, 2011.

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

Bowker, G. C. and Star, S. L. *Sorting things out: classification and its consequences*. MIT Press, Cambridge, Mass. ISBN 0262024616 9780262024617.

Church, L., Zágoni, R., Simpson, A., Srinivasan, S., and Blackwell, A. Building socio-technical systems for representing citizens voices in humanitarian interventions. In *Designing Technologies to Support Human Problem Solving-A Workshop in Conjunction with VL/HCC 2018*, 2018.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

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

Fort, K., Adda, G., and Cohen, K. B. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, 2011.

Guzella, T. S. and Caminhas, W. M. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222, 2009.

He, R. and McAuley, J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pp. 507–517. International World Wide Web Conferences Steering Committee, 2016.

Johnson, D. E., Oles, F. J., Zhang, T., and Goetz, T. A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal*, 41(3):428–437, 2002.

Maron, M. E. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417, 1961.

Qiu, L. Reddit comment and thread dataset. https://github.com/linanqiu/reddit-dataset, 2016.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

West, D. Neural network credit scoring models. *Computers & Operations Research*, 27(11-12):1131–1152, 2000.

Zanzotto, F. M. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.