An Approach to the Frugal Use of Human Annotators to Scale up Auto-coding for Text Classification Tasks

Li'An Chen[1], Hanna Suominen[1,3,4]
firstname.lastname@anu.edu.au
[1] ANU, Centre for the Public Awareness of Science / Canberra, Australia
[2] The Australian National University (ANU), School of Computing (SoCo) / Canberra, Australia
[3] Data61, Commonwealth Scientific and Industrial Research Organisation / Canberra, Australia
[4] University of Turku / Turku, Finland
[*] Address for Correspondence: ANU CPAS, 42A Linnaeus Way, Canberra, ACT 2601, Australia

Abstract

Human annotation for establishing the training data is often a very costly process in natural language processing (NLP) tasks, which has led to frugal NLP approaches becoming an important research topic. Many research teams struggle to complete projects with limited funding, labor, and computational resources. Driven by the Move-Step analytic framework theorized in the applied linguistics field, our study offers a rigorous approach to the frugal use of two human annotators to scale up auto-coding for text classification tasks. We applied the Linear Support Vector Machine algorithm to text classification of a job ad corpus. Our Cohen's Kappa for inter-rater agreement and Area Under the Curve (AUC) values reached averages of 0.76 and 0.80, respectively. The calculated time consumption for our human training process was 36 days. The results indicated that even the strategic and frugal use of only two human annotators could enable the efficient training of classifiers with reasonably good performance. This study does not aim to provide generalizability of the results. Rather, it is proposed that the annotation strategies arising from this study be considered by our readers only if they are fit for one's specific research purposes.

1 Introduction

In natural language processing (NLP), human annotation is an indispensable and decisive step. The human annotation process directly influences the quality of the training data in NLP tasks, and consequently, it influences the quality of machine-generated results. In this regard, Song et al. (2020) have revealed how significant the risk of reaching an incorrect conclusion could be if the quality of human annotation used for validation cannot be guaranteed. Unfortunately, the science of annotation is progressing very slowly (Hovy and Lavid, 2010; Song et al., 2020). In many NLP studies, methodological details concerning the human annotation process have not been fully disclosed (Song et al., 2020). Such a lack of disclosure may hinder readers' judgment of the soundness of human annotation procedures (Hovy and Lavid, 2010; Song et al., 2020). It is time for NLP researchers to attach greater importance to the methodological rigor of human annotation in NLP tasks.

Where the funding and labor are limited, institutions or researchers might have to turn to the 'frugal' use of human annotators for text labelling tasks. For instance, Andreotta et al. (2019) acknowledged the limitation of not being able to afford high computational and labor costs in their machine learning (ML)-assisted analysis of Tweeter commentary. Johnson et al. (2018) also point out cost control that many engineering teams may need to deal with and emphasize the importance of minimizing labor cost and required training data to meet target results in NLP projects. Therefore, well-planned investment of labor and training resources for NLP and ML tasks is a topic worth considerable scholarly attention. We need to investigate how to make the best use of limited labor and monetary resource to achieve the optimal machine-generated outcomes, while preserving methodological rigor.

Crowdsourcing is often put forward as a solution to the human coder resource problem. Aside from the fact that crowds are often not experts, this kind of human annotation is allowed only in some national contexts, such as in the US (e.g., Munro et al., 2010; Pavlick et al., 2014). This solution is not broadly applicable and has ethical implications with respect to researchers exploiting free or cheap labor. For instance, such option does not conform to the requirement for minimum hourly salaries under employment laws in national contexts such as Australia (Australian Government, 2020). Under circumstances of regulatory limitations and within ethical constraints, it becomes necessary to resort to the frugal use of human annotators to scale up

data analyses.

Unlike human annotation tasks for ordinary image annotation (e.g., dog vs. cat recognition), many text annotations require expert knowledge because they are simply more demanding. For instance, the labelling of research skills in job ads involved human annotators who worked as researchers and educators at universities in Mewburn et al. (2020). These researchers point out that it can be extremely time and money-consuming to hire multiple expert human annotators. In many cases, if annotation procedures were well-devised, the frugal option generated results that were as good as the more costly option (Chang et al., 2017; Cocos et al., 2015). From the perspective of cost control, a better option would be to also involve non-expert annotators with well-designed annotation schemes to reach the optimal annotation outcomes (Chang et al., 2017). Therefore, it is in the interest of textual-data scientists to investigate if there is a way to guarantee the quality of manual annotation with the frugal use of human coders for automatic textual data analyses at scale. As many social science disciplines (e.g., applied linguistics or sociology) have a record of excellent human annotation frameworks, it is worth considering if annotation frameworks in any of these fields could help us enhance the methodological soundness for human annotation process in NLP tasks.

The research questions of this study are posed as follows:

1. For automatic text classification tasks, how could we design human annotators' workshop frugally and at the same time maintain good performance of the machine?

2. How could we design the human annotators' workshop to enable easy identification and fixation of problems in the human annotation schema?

3. If multiple human annotators were involved, which annotator's labelled data should be adopted for training?

The primary outcomes of this study were as follows:

1. The frugal use of an expert annotator and a non-expert annotator generated an averaged Cohen's Kappa of 0.76.

2. The total time investment of our frugal approach to human annotation was 376 hours (the time consumed by two human annotators).

3. The frugal use of only two human annotators plus a limited amount of labelled data resulted in an averaged area under the receiver operating characteristic (ROC) curve (AUC) score of 0.80.

4. Differentiation of coarse-grained and fine-grained labels allowed for enhanced interpretability of the ML performance. It also allowed for strategically hybrid use of multiple human annotators' labels to optimize the ML performance.

2 Methods

2.1 Data

Our human coders annotated job ad data from a corpus of high research skill intensive job ads of computing and healthcare positions[1]. In total, 1,800 job ads were chosen randomly from a large corpus consisted of health-domain and computing-domain job postings. The word counts of the 1,800 job ads reached 680,367. The randomly chosen job ads contained 900 health-domain job ads and 900 computing-domain job ads. As we aimed to minimize the labor and time cost, as well as the amount of data used for training and validation, the selection of only 1,800 job ads was based on a balanced consideration of the machine's performance and the time investment on manual annotation.

The job ad corpus was purchased from Burning Glass Technology Inc. Due to legal constraints, the data used for this study cannot be shared. However, it is assumed that our audience would be those who do not necessarily need to conduct analyses of job ads, but potentially other text classification tasks. Alternatively, readers interested in obtaining the same data for a verification of the results could contact Burning Glass Technology Inc. directly.

2.2 Ethics

We went through necessary ethics procedures to avoid potential conflict of interests. We obtained the approval for the data to be used for our research purpose. The manuscript of the paper was read by a legal consultant in our team and a representative from Burning Glass Technology Inc. to ensure our publication met contractual agreements. We also signed an agreement with our human annotators for clarification of responsibilities and task specifications. The agreement with the human annotators was approved by our ethics delegate. Thus, we believe that ethical issues were mitigated to the best of our abilities.

[1] We only analyzed computing-domain and health-domain job postings because the current paper is part of a large project to contextualize high-RSI job requirements for pedagogical purposes.

2.3 Human Annotators' Workshop

Our study involved two human annotators for the labelling of requirements in job ads. The first human annotator N1was one of the authors of the paper. N1 was an expert annotator and a PhD candidate who held a master's degree in applied linguistics with extensive experience in identifying job requirements from textual data. The second annotator N2 was hired as a volunteer for our task. N2 held a master's degree in finance with experience in classifying news information, her experience was less relevant compared to N1. Hence, N2 played the role of a novice human annotator in the annotators' team.

Before assigning the job ads to N1 and N2, the job ad texts were segmented into sentences to be labelled by the annotators. The purpose of segmenting the job ad data into sentences was to reduce cognitive burdens for both annotators.

It was decided that there should be both coarse-grained labels and fine-grained labels. The decision was theoretically driven and inspired by an inductive analytic framework called 'Move-Step analysis' pioneered by the renowned applied linguist John Swales (1990). Move-Step analysis is a widely adopted linguistic approach to the systematic examination of different genres (or text types). Genre theorists (Miller, 1984; Bhatia, 2014; Moreno and Swales, 2018) advocate that writing is a social action, and so a specific genre serves as a tool to achieve a social purpose that is shared among a community of practice. In our case, the purpose of the job ad genre is the communication of various skills, qualifications and capabilities required of a particular job vacancy, by the employer to potential hirees. To achieve an overarching purpose of a genre, writers need to involve conventionally acknowledged components in their writing (Swales, 1990). Swalesian genre theorists differentiated the conventional textual components of a genre into coarse-grained moves and fine-grained steps. The intention of differentiating granularity levels derives from the pedagogical orientation shared among the Swalesian genre theorists (Bhatia, 2014; Maswana et al., 2015; Moreno and Swales, 2018) for clarifying concepts more clearly in class. Move-step analysis has previously been applied by NLP researchers such as Chen et al. (2020) for projects with a strong pedagogical orientation. As argued by Chen et al. (2020), the provision of coarse-grained and fine-grained conventions embedded in the writing of a genre would allow students to learn more efficiently. The pedagogical orientation of move-step analysis aligns well with our intention to identify job requirements to enrich employability training[2].

To give the readers a clearer sense of what we meant by a coarse-grained/move-level job requirement label and its associated fine-grained/step-level labels, we give the example of the job requirement 'Continuous education' below:

Coarse-grained/Move-step label:

• Continuous education.

Its associated fine-grained labels:

• Passion & Self-motivation,
• Participation in training,
• Sharing of knowledge,
• Seeking advice, and
• Self-reflection.

Moreover, we assumed that the differentiation between coarse-grained and fine-grained labels might have other potential benefits. Having coarse- and fine-grained labels may speed up the annotation process. In this regard, Tange et al. (1998) showed that the combination of coarse and fine-grained labels helped the readers of informatics process information faster and more accurately.

After introducing move-step analysis and assigning the task to the two annotators, N1 conducted the first round of annotation of 200 job ads, as she had the expert skills and knowledge relevant to the task. It was then decided that the unit to be annotated could contain multiple labels, as N1 found that the employers sometimes put multiple requirements in one sentence. Hence, our task was multi-label text classification. After N1 finished the first round of annotation, she came up with a coding schema that listed all the coarse-grained and fine-grained job requirement categories, and she gave the schema to N2. From the second to the last round of annotation, both N1 and N2 were involved in the task. N1 and N2 conducted their annotation tasks individually. The two annotators used the annotation tool Dataturks to label the texts.

Overall, there were nine rounds of annotation. In between every two rounds of annotation, the

[2]How to use the identified job requirements to enrich employability training is not covered in the current paper. Our main focus in this study is still the demonstration of the frugal use of human annotators. The point of mentioning the alignment between our pedagogical aim and the use of move-step analysis is to advocate a well-justified selection of analytic framework to be used in human annotators' workshop to fit one's specific research aim.

two annotators met once to discuss their compared results. If a high level of inconsistency measured by Cohen's Kappa was found regarding a particular fine-grained label (e.g., Continuous education - Passion & Self-motivation), N1 and N2 would randomly scan through several inconsistent instances and give their justifications about why they labeled in their ways. If the agreement was reached concerning how to label similar instances in the future, both of them would write the agreed approach in their notepads. However, if an agreement was not reached after their justifications were given, they would note down the dubious items and leave them for the next meeting when they labeled more data and had further justifications to convince each other.

The inter-rater reliability between the two human annotators was measured by Cohen's Kappa. For assessing coders' agreement on the annotation of categorical variables, Hallgren (2013) recommends Cohen's Kappa as the measurement. The Cohen's Kappa equation was given in (1) as follows:

$$K = \frac{P(a) - P(e)}{1 - P(e)} \tag{1}$$

where P(a) denotes the observed percentage of the human annotators' agreement and P(e) refers to the probability that the agreement is met by chance.

After the Kappa was calculated for each coarse-grained and fine-grained category, we also calculated the standard error for the calculation of the 95% confidence intervals for the Kappa. The standard error equation is given in (2) as follows:

$$\alpha_K = \sqrt{\frac{P(a)(1 - P(e))}{N(1 - P(e))^2}} \tag{2}$$

where N refers to the overall numbers of classified tokens.

2.4 Machine Learning Methods

The algorithm chosen for running the auto-coding task was the Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with the linear kernel. Linear SVM is a good choice in a low-resource context (Zhang et al., 2012), such as in ours. Linear SVM had also a low computational cost and at the same time good prediction results (Vijayan et al., 2017). For multi-label text classification tasks, Linear SVM could have good ability to generate prediction results close to those generated by manual efforts (Qin and Wang, 2009; Yang et al., 2009; Wang and Chiang, 2011).

We involved several steps in preprocessing the data. As mentioned in the description of the human coders' workshop, we segmented the job ad texts into sentences as labeling units. The classification task hence was also at sentence level. There were 63,504 sentence units overall. The average number of labels per sentence was 1.8. The segmentation into sentences supported calculation of the job requirements more accurately. Additionally, we removed stop words (e.g., conjunctions, articles) from the texts via the stop-word list given in the Natural Language Toolkit (NLTK) corpus v.3.5. The data were then put in a machine-readable format with the word representation tool TfidfVectorizer (term frequency times inverse document frequency) from the Scikit-learn v0.24.1.

We separated the processed data into 70%, 15%, and 15% chunks for the training, testing, and validation purposes. The ratio of the training/test/validation sets was based on the conventional practice suggested in Muller and Guido (2016) and Ng (2020). We were aware of other validation approaches such as K-fold cross-validation (CV). Considering that the tuning of the hyper-parameters (e.g., K value and ratio) in other CV approaches could be time-consuming and computationally expensive whilst their gain limited (as in Anguita et al., 2012 and Racz et al., 2021), we chose to proceed with the frugal option of 70%, 15% ,15% split of the data for train/test/validation.

For the parameter-tuning function of the Linear SVM classifier, we adopted the GridSearchCV tool from the Scikit-Learn v0.24.1. More specifically, the parameters tuned were 1) Loss, 2) Max-iteration, 3) Tolerance, 4) Fit intercept, and 5) Intercept scaling.

The performance of the Linear SVM classifier was measured by the AUC. The reason that we chose the AUC is that it, compared to the accuracy, F1 or other such measurements, was less prone to biased results from class imbalance (Suominen et al., 2008; Narkhede, 2018).

After the AUC values were calculated, we also computed the 95% confidence intervals for our automatic classifier.

3 Results

The inter-rater agreement measured by Cohen's K reached an average of 0.76 (see Section 3.1),

meaning that most of our manually labeled categories can be used for making at least tentative conclusions. The results related to the total time investment in the human annotation process (see Section 3.2) suggested that two human annotators, each working 5 hours a day, would need approximately 36 days to complete the task. Section 3.3 is concerned with the performance of the two automatic classifiers trained with data labeled by our two annotators. Although the two classifiers both reached an averaged AUC of 0.80, a closer examination of fine-grained categories revealed potential room for further improvement to the human annotation schema. These findings posed the question of whether high-inter rater agreement is more important than the ML results' interpretability. Moreover, strategic hybrid use of the two classifiers for optimization was introduced in Section 3.3.

3.1 Inter-rater Agreement

The averaged inter-rater reliability measured by Kappa for all the identified categories reached 0.76 (see Table 1). For the fine-grained categories, the Kappa ranged from the minimum 0.60 to the maximum of 0.94. At the coarse-grained level, the Kappa range from 0.68 to 0.83. Based on the Kappa interpretation guidelines suggested by Krippendorff (2018), Kappa values under 0.67 indicate that any conclusion should not be counted. Values ranging from 0.67 to 0.80 point to tentative conclusions to be made. Values above 0.80 indicate that definite conclusions can be made. Based on Krippendorff's guidelines, it is safe to claim that only 9 out of 72, or 12.5% of the fine-gained categories, did not reach the standards for making a tentative conclusion. The rest 87.5% of fine-grained categories reached the 'Pass' Kappa threshold defined by Krippendorff, which has been deemed among the strictest (Hallgren, 2013). If we use guidelines defined by Landis and Koch (1977), who viewed Kappa under 0.61 as enough for the indication of a moderate agreement between two annotators, most of our fine-grained categories can be used for making at least tentative conclusions.

3.2 Time Investment in Human Annotation

The annotators reported that averagely they spent ten seconds annotating each sentence token in the task when they were fully concentrating on the task. The two annotators both labelled 63,504 sentence tokens. Therefore, the total time investment in the completion of a single-person annotation task was

Coarse-grained	Fine-grained	*Kappa* (Fine)	*95% CI* (Fine)	*Kappa* (Coarse)
People skills	Peer practitioners	0.76	(0.72 - 0.80)	0.70
	Interpersonal skills	0.77	(0.71 - 0.82)	
	Multidisciplinary collaboration	0.76	(0.72 - 0.81)	
	Decision makers	0.64	(0.59 - 0.69)	
	Public sectors	0.71	(0.68 - 0.74)	
	Private sectors	0.60	(0.55 - 0.66)	
	Business partners	0.71	(0.68 - 0.74)	
	General public	0.60	(0.54 - 0.66)	
	Research participants	0.70	(0.67 - 0.74)	
	Research institutions	0.78	(0.75 - 0.81)	
Empathy with	Clients	0.87	(0.81 - 0.93)	0.83
	Less experienced	0.73	(0.69 - 0.77)	
	Ethnic minorities	0.83	(0.80 - 0.87)	
	Children	0.89	(0.86 - 0.92)	
	Public welfare	0.65	(0.61 - 0.69)	
	Clients' families	0.85	(0.81 - 0.90)	
	Aging population	0.86	(0.82 - 0.91)	
	Disabled	0.88	(0.85 - 0.92)	
	Women	0.77	(0.73 - 0.91)	
	LGBTQI+	0.94	(0.91 - 0.99)	
Personal attributes	Leadership skills	0.82	(0.78 - 0.86)	0.77
	Time efficiency	0.79	(0.74 - 0.83)	
	Commercial orientation	0.69	(0.65 - 0.74)	
	Safety awareness	0.81	(0.77 - 0.85)	
	High-pressure management	0.72	(0.67 - 0.77)	
	Personal impact	0.81	(0.77 - 0.84)	
	Result orientation	0.82	(0.77 - 0.86)	
	Attention to details	0.77	(0.74 - 0.81)	
	Independence	0.80	(0.76 - 0.85)	
	Agility	0.71	(0.68 - 0.75)	
Continuous education	Passion & Motivation	0.78	(0.73 - 0.82)	0.78
	Knowledge sharing	0.83	(0.80 - 0.87)	
	Participation in training	0.80	(0.77 - 0.84)	
	Seeking advice	0.77	(0.72 - 0.81)	
	Self-reflection	0.70	(0.64 - 0.75)	
Cognitive abilities	Analytic skills	0.80	(0.77 - 0.84)	0.81
	Problem understanding & solving	0.83	(0.77 - 0.88)	
	Needs interpretation	0.76	(0.72 - 0.80)	
	Innovation	0.85	(0.81 - 0.89)	
Pursuit of job quality	Organisational environment	0.82	(0.78 - 0.86)	0.77
	Long-term contract	0.76	(0.72 - 0.79)	
	Specified payment	0.74	(0.69 - 0.79)	
	Unspecified payment	0.73	(0.70 - 0.77)	
	Surrounding environment	0.80	(0.75 - 0.84)	
Proof of qualification	Registration in institutions	0.79	(0.76 - 0.82)	0.77
	Writing skills	0.72	(0.68 - 0.75)	
	Oral & Presentation skills	0.79	(0.73 - 0.84)	
	Residency	0.80	(0.77 - 0.83)	
	Tertiary degree	0.81	(0.75 - 0.87)	
	Years of industry experience	0.80	(0.76 - 0.83)	
	General IT skills	0.70	(0.64 - 0.77)	
Professional standards	Policy & Regulation familiarisation	0.82	(0.78 - 0.85)	0.78
	Background check	0.88	(0.84 - 0.92)	
	Ethical conduct	0.61	(0.57 - 0.65)	
	Awareness of confidentiality	0.79	(0.75 - 0.84)	
Aesthetics	Detecting defects & debugging	0.77	(0.72 - 0.83)	0.81
	Refined design	0.85	(0.81 - 0.89)	
	Maintenance of environment	0.80	(0.72 - 0.87)	
Courage	Change management	0.79	(0.76 - 0.83)	0.72
	Risk management	0.77	(0.71 - 0.83)	
	Execution from concept	0.60	(0.56 - 0.64)	
	Travelling & Driving	0.81	(0.75 - 0.87)	
	On-call availability	0.69	(0.61 - 0.77)	
	Conflict management	0.68	(0.62 - 0.73)	
	Working in harsh environment	0.70	(0.62 - 0.77)	
Resource management	Capital & Budget management	0.70	(0.66 - 0.75)	0.71
	Configuration management	0.81	(0.77 - 0.84)	
	Resource allocation	0.63	(0.60 - 0.67)	
Hiring procedure	ASAP orientation	0.65	(0.61 - 0.69)	0.70
	Quality selection process	0.75	(0.70 - 0.81)	
Subject knowledge	Computer science subject knowledge	0.75	(0.70 - 0.80)	0.68
	Medical science subject knowledge	0.60	(0.56 - 0.65)	
None category	* Step tokens with no information about job requirements	0.92	(0.90 - 0.93)	N/A

Table 1: Cohen's K and the respective 95% confidence interval (CI) for the inter-rater agreement.

approximately 177 hours. Suppose a research team hires two annotators to do the coding task concurrently, and both the annotators work five hours a day. A project of a size comparable to ours might need about 36 days for the manual labeling to be completed. We considered such a time span as reasonably moderate. In addition, if the hired annotators could work for over five hours each day, the completion of the manual labeling process could be even faster. The exact hours allocated to a human annotator per day might vary based on different research teams' consideration.

The total labeling hours of the two annotators were 354 hours. Our corpus contained 826,891 words. Therefore, the approximate time investment per word for our labelling task was 1.6s. There were nine rounds of meetings (one hour for every meeting) plus the two-hour orientation time. Hence, two-person efforts for orientation and meetings cost 22 hours. In total, our two-annotator labeling task incurred a 376-hour time investment. Any team who also wants to use a similar frugal approach to their human-labeling task would find our results of interest.

3.3 Performance of the Automatic Classifier

The two automatic classifiers trained and tested with the data labeled by our two human annotators both reached an averaged gold-standard AUC value of 0.80. Table 2 suggest that 58% of the coarse-grained categories reached AUC values above 0.80 with Machine N1 on data labeled by N1. Around 57% of step-level categories reached AUC values above 0.8 with Machine N2 on data labeled by N2. The scores of AUC given by the machine trained and tested from data labeled by annotator N1 ranged from 0.52 to 1.00. The scores of AUC given by the machine trained and tested from data labeled by annotator N2 ranged from 0.58 to 0.99. Interestingly, when we calculated the average of the AUC results given by Machine N1 trained and tested on Data N1 for all the fine-grained categories, the value reached 0.80. Similarly, the averaged AUC results given by Machine N2 trained and tested on Data N2 reached 0.80, too. This reminded us of the likelihood that even when a machine's performance seems outstanding at a coarse-grained level, potential problems at a fine-grained level might be invisible.

Certain coarse-grained categories such as 'Decision makers' and 'Public welfare' were low in AUC scores. We would pay particular attention to these categories in our future attempt for continuous improvement. Our approach of identifying both the fine- and coarse-grained categories proved to be one that could increase the interpretability of the results. More specifically, if we had not differentiated between the fine- and coarse-grained categories, we would not have been able to know where the problem lay in the human annotation schema. With the information about which fine-grained categories did well and which did not, we could allow more efficient future attempts to drive continuous improvement on the human coding schema.

When classifier Ni was tested with data labeled by Nj, most of our fine-grained categories did not show a large decrease in the AUC. When the drop was small, we assumed that the two ML classifiers trained by the two annotators performed almost equally well. We only found 15 fine-grained categories to have a relatively large decrease in the AUC. We used an averaged decrease of 0.05 as the threshold (a threshold used in Hiissa et al., 2006) to denote a large decrease in classifier Ni's performance when tested with data labeled by Nj.

These 15 fine-grained categories, which showed a large decrease in performance were 'Peer practitioners', 'Interpersonal skills', 'Safety awareness', 'Agility', 'Passion Motivation', 'Problem understanding & solving', 'Unspecific payment', 'Residency', 'Refined design', 'Change management', 'Risk management', 'Conflict management', 'Working in harsh environment', 'Resource allocation', and 'Medical science subject knowledge' (Table 2).

These 15 fine-grained categories had good performance with Machine Ni tested on data labeled by Ni, but Machine Ni on data labeled by Nj gave a worse performance. This could indicate that the two human annotators' inner-rater reliability was high, but their inter-rater reliability was not as high. When human annotators face categories like these 15 ones in our study, we recommend a check regarding which features the human annotator Ni deemed as relevant to a category, but the human annotator Nj deemed as not. For the rest categories that did not show a large decrease, we recommend that researchers put Machine Ni into the formal use if Machine Ni on Data Nj results in less decrease in the AUC whilst Machine Ni's performance on Data Ni is also good. Instead of relying on the use of a single classifier for classifying all the fine-

grained categories, the hybrid usage of Machine N1 and Machine N2 could optimize the classifier's performance even if the annotators' workshop was frugally designed.

4 Discussion

Our study showed that even the frugal use of only two human annotators plus a limited amount of labeled data resulted in an averaged AUC score of 0.80. Nonetheless, the differentiation between the fine-grained and coarse-grained categories in our coding schema revealed even the averaged AUC of 0.80 did not necessarily mean the quality of human annotation was as good [3]. The differentiation of fine and coarse granularities could enhance the interpretability of the results. In particular, such a differentiation provided a straightforward indication as to where the machine performed well or not and also where the problems lay in the human annotators' coding schema.

Our study had limitations. Although we provided justifications for all the choices we made in our methods, there is room to refine our project's design (e.g., involving classification of other genres) when we have more resources. Compared to most previous coding schemas where no differentiation of granularity levels was made, our approach could allow more to-the-point and efficient fixation of the human annotation for continuous improvement. Our findings regarding the benefits of having two granularity levels echo the results in Chen et al. (2020). Our choice of making the differentiation between granularity levels counters the suggestion given by Hovy and Lavid (2010). They argue that coarser granularity would improve the accuracy of human annotation results. Nonetheless, Hovy and Lavid (2010) have mostly used examples of semantic recognition tasks such as verb-sense annotation to support their argument. Our task of text classification is different from semantic recognition. Therefore, it is worth further investigating whether it is reasonable to always opt for 'neutering' for all NLP tasks only for the sake of reaching a high inter-rater agreement regardless of the research purpose.

Our frugal use of one expert annotator and one non-expert annotators proved to cost moderate annotation time whilst generating reasonably good results. Compared to the recruitment of multiple expert-annotators, our approach certainly was much less costly. The strategically hybrid use of automatic classifiers trained by our two annotators is perhaps comparable to a classifier trained by only expert annotators. However, such an assumption is subject to future investigations where appropriate measures are involved.

Future scholarly attempts could explore this topic of frugal hybrids of machines and human experts further to verify our assumption. In this regard, Fort (2016) and Chen et al. (2020) echo our thoughts by arguing that a well-devised non-expert annotator workshop could allow the labeling quality to be as good as when only expert annotators generate the labeling. Chang et al. (2017) expressed the concern that writing guidelines for even simple concepts for non-expert coders can be very prohibitive, but our approach of mixing both expert and non-expert coders is less likely to incur uncertainties and unexpected costs. To drive the progress of the science of annotation, scholars in the future might find it interesting to compare labeling results generated by pure experts, a mixture of experts non-experts, and crowdsourced workers for the same NLP project.

5 Conclusion

In this study, we advocate a methodologically sound approach to the frugal use of two annotators to conduct human annotation tasks for NLP projects. Our approach has multiple benefits. Specifically, the time and resource consumption of our frugal approach were moderate compared to the more expensive choice of hiring multiple expert annotators. Having multiple rounds of annotation activities and ongoing meetings makes it possible to make timely justification and adjustments for the annotation schema. Moderate cost, timely communication of dubious labels, joint development of the annotation schema, and reasonably good ML outcomes are the features of our frugal but theoretically sound approach to human annotation. These features make the frugal use of minimally two hu-

[3]The point of constantly mentioning the coarse-grained categories in this paper is to emphasize how coarse granularity alone was unable to ensure the optimal performance for our specific annotation task. Single granularity level has been pervasively used in many text classication tasks (Chen et al., 2018; Da San Martino et al., 2019; Heinisch and Cimiano, 2021). Nonetheless, recent studies (Chen et al., 2018; Da San Martino et al., 2019; Heinisch and Cimiano, 2021) suggest that single granularity cannot guarantee the optimal performance for certain tasks, which echo our findings here. In addition, we feel it necessary to keep the coarse granularity because the high-level categories are always useful when presenting complex results to the public

Coarse-grained	Fine-grained	AUC (Machine N1 on N1)	AUC (Machine N2 on N2)	AUC (Machine N1 on N2)	AUC (Machine N2 on N1)	Drop (Machine N1)	Drop (Machine N2)	95% CI (Machine N1)	95% CI (Machine N2)
People skills	Peer practitioners	0.82	0.84	0.77	0.77	0.05	0.07	(0.74 - 0.86)	(0.76 - 0.89)
	Interpersonal skills	0.85	0.83	0.76	0.81	0.09	0.02	(0.82 - 0.88)	(0.80 - 0.87)
	Multidisciplinary collaboration	0.86	0.84	0.84	0.78	0.02	0.06	(0.80 - 0.92)	(0.78 - 0.90)
	Decision makers	0.69	0.65	0.62	0.65	0.07	0.00	(0.63 - 0.74)	(0.60 - 0.70)
	Public sectors	0.69	0.65	0.66	0.69	0.03	0.00	(0.59 - 0.78)	(0.55 - 0.74)
	Private sectors	0.68	0.70	0.70	0.72	0.00	0.00	(0.59 - 0.79)	(0.61 - 0.80)
	Business partners	0.79	0.72	0.80	0.71	0.00	0.01	(0.73 - 0.84)	(0.67 - 0.78)
	General public	0.67	0.62	0.61	0.60	0.06	0.02	(0.61 - 0.72)	(0.55 - 0.69)
	Research participants	0.73	0.70	0.78	0.72	0.00	0.00	(0.67 - 0.79)	(0.64 - 0.77)
	Research institutions	0.81	0.83	0.88	0.79	0.00	0.04	(0.75 - 0.87)	(0.77 - 0.89)
Empathy with	Clients	0.88	0.92	0.87	0.88	0.01	0.04	(0.84 - 0.92)	(0.88 - 0.95)
	Less experienced	0.82	0.75	0.84	0.76	0.00	0.00	(0.77 - 0.86)	(0.70 - 0.79)
	Ethnic minorities	0.87	0.82	0.83	0.79	0.04	0.03	(0.82 - 0.91)	(0.77 - 0.86)
	Children	0.92	0.89	0.87	0.86	0.05	0.03	(0.88 - 0.96)	(0.84 - 0.93)
	Public welfare	0.52	0.58	0.60	0.56	0.00	0.02	(0.47 - 0.58)	(0.53 - 0.64)
	Clients' families	0.88	0.85	0.91	0.86	0.00	0.00	(0.84 - 0.92)	(0.80 - 0.89)
	Aging population	0.89	0.87	0.86	0.90	0.03	0.00	(0.83 - 0.94)	(0.81 - 0.93)
	Disabled	0.97	0.92	0.91	0.96	0.06	0.00	(0.94 - 0.99)	(0.87 - 0.97)
	Women	0.82	0.88	0.79	0.78	0.03	0.10	(0.76 - 0.88)	(0.82 - 0.94)
	LGBTQI+	1.00	0.99	0.98	0.98	0.02	0.01	(0.99 - 1.00)	(0.97 - 0.99)
Personal attributes	Leadership skills	0.85	0.84	0.89	0.88	0.00	0.00	(0.79 - 0.90)	(0.78 - 0.89)
	Time efficiency	0.80	0.83	0.76	0.79	0.04	0.04	(0.76 - 0.85)	(0.78 - 0.88)
	Commercial orientation	0.75	0.72	0.70	0.77	0.05	0.00	(0.70 - 0.81)	(0.66 - 0.77)
	Safety awareness	0.92	0.91	0.89	0.83	0.03	0.08	(0.88 - 0.96)	(0.87 - 0.95)
	High-pressure management	0.71	0.72	0.69	0.70	0.02	0.02	(0.66 - 0.77)	(0.66 - 0.76)
	Personal impact	0.83	0.85	0.86	0.82	0.00	0.03	(0.79 - 0.87)	(0.80 - 0.89)
	Result orientation	0.76	0.73	0.71	0.69	0.05	0.04	(0.70 - 0.81)	(0.69 - 0.78)
	Attention to details	0.78	0.78	0.75	0.77	0.03	0.01	(0.75 - 0.83)	(0.75 - 0.83)
	Independence	0.83	0.81	0.85	0.86	0.00	0.00	(0.79 - 0.87)	(0.77 - 0.85)
	Agility	0.84	0.82	0.72	0.75	0.12	0.07	(0.76 - 0.93)	(0.73 - 0.90)
Continuous education	Passion & Motivation	0.87	0.85	0.77	0.79	0.10	0.06	(0.83 - 0.91)	(0.80 - 0.89)
	Knowledge sharing	0.84	0.85	0.82	0.84	0.02	0.01	(0.79 - 0.88)	(0.80 - 0.90)
	Participation in training	0.87	0.84	0.81	0.85	0.06	0.00	(0.83 - 0.91)	(0.79 - 0.88)
	Seeking advice	0.71	0.72	0.78	0.77	0.00	0.00	(0.67 - 0.76)	(0.68 - 0.77)
	Self-reflection	0.70	0.69	0.76	0.68	0.00	0.01	(0.60 - 0.81)	(0.59 - 0.80)
Cognitive abilities	Analytic skills	0.77	0.74	0.82	0.78	0.00	0.00	(0.72 - 0.83)	(0.69 - 0.79)
	Problem understanding & solving	0.94	0.92	0.90	0.84	0.04	0.08	(0.91 - 0.97)	(0.88 - 0.95)
	Needs interpretation	0.78	0.82	0.71	0.79	0.07	0.03	(0.73 - 0.84)	(0.77 - 0.86)
	Innovation	0.90	0.91	0.86	0.89	0.04	0.02	(0.86 - 0.94)	(0.87 - 0.95)
Pursuit of job quality	Organisational environment	0.91	0.91	0.88	0.90	0.03	0.01	(0.86 - 0.96)	(0.86 - 0.96)
	Long-term contract	0.73	0.70	0.75	0.74	0.00	0.00	(0.64 - 0.81)	(0.61 - 0.78)
	Specified payment	0.79	0.77	0.71	0.75	0.08	0.02	(0.69 - 0.88)	(0.67 - 0.87)
	Unspecified payment	0.76	0.78	0.73	0.70	0.03	0.08	(0.70 - 0.83)	(0.71 - 0.85)
	Surrounding environment	0.61	0.66	0.65	0.70	0.00	0.00	(0.54 - 0.68)	(0.59 - 0.74)
Proof of qualification	Registration in institutions	0.84	0.84	0.87	0.81	0.00	0.03	(0.80 - 0.88)	(0.80 - 0.88)
	Writing skills	0.79	0.75	0.74	0.72	0.05	0.03	(0.74 - 0.83)	(0.70 - 0.80)
	Oral & Presentation skills	0.81	0.81	0.80	0.79	0.01	0.02	(0.75 - 0.87)	(0.75 - 0.87)
	Residency	0.96	0.89	0.84	0.85	0.12	0.04	(0.92 - 0.99)	(0.85 - 0.94)
	Tertiary degree	0.88	0.91	0.88	0.85	0.00	0.06	(0.81 - 0.95)	(0.83 - 0.98)
	Years of industry experience	0.84	0.84	0.81	0.82	0.03	0.02	(0.79 - 0.88)	(0.79 - 0.88)
	General IT skills	0.70	0.71	0.73	0.69	0.00	0.02	(0.65 - 0.76)	(0.66 - 0.77)
Professional standards	Policy & Regulation familiarisation	0.89	0.87	0.83	0.81	0.06	0.06	(0.85 - 0.93)	(0.92 - 0.91)
	Background check	0.92	0.96	0.88	0.90	0.04	0.06	(0.88 - 0.95)	(0.93 - 0.99)
	Ethical conduct	0.64	0.68	0.62	0.67	0.02	0.01	(0.60 - 0.68)	(0.64 - 0.72)
	Awareness of confidentiality	0.85	0.87	0.88	0.80	0.00	0.07	(0.80 - 0.91)	(0.82 - 0.92)
Aesthetics	Detecting defects & debugging	0.81	0.83	0.80	0.78	0.01	0.05	(0.76 - 0.85)	(0.78 - 0.87)
	Refined design	0.93	0.92	0.86	0.87	0.07	0.05	(0.89 - 0.97)	(0.88 - 0.96)
	Maintenance of environment	0.84	0.86	0.81	0.79	0.03	0.07	(0.78 - 0.91)	(0.80 - 0.93)
Courage	Change management	0.87	0.88	0.79	0.82	0.08	0.06	(0.83 - 0.91)	(0.84 - 0.92)
	Risk management	0.85	0.83	0.77	0.78	0.08	0.05	(0.80 - 0.91)	(0.78 - 0.88)
	Execution from concept	0.66	0.70	0.68	0.63	0.00	0.07	(0.60 - 0.72)	(0.64 - 0.76)
	Travelling & Driving	0.85	0.88	0.81	0.82	0.04	0.06	(0.79 - 0.90)	(0.82 - 0.93)
	On-call availability	0.73	0.74	0.77	0.69	0.00	0.05	(0.68 - 0.77)	(0.69 - 0.78)
	Conflict management	0.77	0.74	0.69	0.71	0.08	0.03	(0.73 - 0.82)	(0.70 - 0.79)
	Working in harsh environment	0.77	0.73	0.70	0.68	0.07	0.05	(0.70 - 0.85)	(0.66 - 0.80)
Resource management	Capital & Budget management	0.75	0.77	0.71	0.72	0.04	0.05	(0.70 - 0.81)	(0.72 - 0.83)
	Configuration management	0.83	0.84	0.77	0.82	0.06	0.02	(0.79 - 0.88)	(0.80 - 0.89)
	Resource allocation	0.73	0.71	0.68	0.62	0.05	0.09	(0.68 - 0.78)	(0.66 - 0.76)
Hiring procedure	ASAP orientation	0.70	0.69	0.70	0.66	0.00	0.03	(0.64 - 0.76)	(0.64 - 0.75)
	Quality selection process	0.81	0.84	0.87	0.78	0.00	0.06	(0.75 - 0.86)	(0.78 - 0.89)
Subject knowledge	Computer science subject knowledge	0.81	0.79	0.76	0.77	0.05	0.02	(0.73 - 0. 88)	(0.71 - 0.86)
	Medical science subject knowledge	0.69	0.71	0.60	0.63	0.09	0.08	(0.61 - 0.78)	(0.62 - 0.80)

Table 2: AUC values and respective 95% confidence intervals (IC) & Drop from Machine Ni tested on Nj.

man annotators a good alternative to crowdsourcing and expert annotation. Regarding whether or not to differentiate granularity levels and whether or not to resort to human annotation frameworks from non-NLP disciplines in the human annotation process, our suggestion is that researchers should make the decision based on specific research purposes. We hope this study could serve as a point to drive reflection upon the science of annotation within our NLP community.

Acknowledgement

We are grateful for the support from Emsi Burning Glass Inc, PostAc®, and ANU CV Discovery Translation Fund2.0. Our thanks also go to Prof. Inger Mewburn, Dr. Will Grant, and the anonymous paper reviewers for their insightful comments on this paper. We thank Dr. Lindsay Hogan and Chenchen Xu for offering us advise on the technical and legal requirements involved in this study. We appreciate the anonymous annotator's contribution in our coders' workshop. Finally, the first author would like to thank Australian Government Research Training Program International Scholarship for supporting her PhD studies.

References

Andreotta, M., Nugroho, R., Hurlstone, M., Boschetti, F., Farrell, S., Walker, I., and Paris, C. (2019). Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, 51(4):1776–1781.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., and Ridella, S. (2012). The 'k' in k-fold cross validation. In *Proceedings of the 2012 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 441–446.

Australian Government (2020). Fair work: Minimum wages. Accessed: 2021-07-23.

Bhatia, V. (2014). *Analysing genre: Language use in professional settings*. Routledge, London, UK.

Chang, C. J., Amershi, S., and Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346.

Chen, L., Liang, J., Xie, C., and Xiao, Y. (2018). Short text entity linking with fine-grained topics. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 457–466.

Chen, L., Suominen, H., and Mewburn, I. (2020). A machine-learning based model to identify phd-level skills in job ads. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 72–80.

Cocos, A., Qian, T., Callison-Burch, C., and Masino, A. J. (2015). Crowd control: effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of biomedical informatics*, 69:86–92.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Da San Martino, G., Yu, S., Barron-Cedeno, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5636–5646.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley Sons, London, UK.

Hallgren, K. (2013). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–24.

Heinisch, P. and Cimiano, P. (2021). A multi-task approach to argument frame classification at variable granularity levels. *Information Technology*, 63(1):59–72.

Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salantera, S., and Salakoski, T. (2006). Towards automated classification of intensive care nursing narratives. *Studies in health technology and informatics*, 124:789–794.

Hovy, E. and Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–16.

Johnson, M., Anderson, P., Dras, M., and Steedman, M. (2018). Predicting accuracy on large datasets from smaller pilot data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 450–455.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications., New York, USA.

Landis, R. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Maswana, S., Kanamaru, T., and Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand*, 2:1–11.

Mewburn, I., Grant, W. J., Suominen, H., and Kizimchuk, S. (2020). A machine learning analysis of the non-academic employment opportunities for phd graduates in australia. *Higher Education Policy*, 33(4):799–813.

Miller, C. (1984). Genre as social action. *Quarterly journal of speech*, 70(2):151–167.

Moreno, A. I. and Swales, J. M. (2018). Gstrengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, 50:40–63.

Muller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly, Newton, US.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards Data Science*, 26:220–227.

Ng, A. (2020). Coursera: Machine learning by stanford university. Accessed: 2021-07-23.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Qin, Y. P. and Wang, X. K. (2009). Study on multilabel text classification based on svm. In *Proceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 333–304.

Racz, A., Bajusz, D., and Heberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4):1111.

Song, H., Tolochko, P., Eberl, J., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., and Boomgaarden, H. (2020). In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4):550–572.

Suominen, H., Pyysalo, S., Hiissa, M., Ginter, F., Liu, S., Marghescu, D., Pahikkala, T., Back, B., Karsten, H., and Salakoski, T. (2008). Performance evaluation measures for text mining. In Song, M. and Wu, Y., editors, *Handbook of Research on Text and Web Mining Technologies*, pages 724–747. IGI Global, Hershey, USA.

Swales, J. M. (1990). *Genre analysis: English in academic and research setting*. Cambridge University Press, Cambridge, UK.

Tange, H. J., Schouten, H. C., Kester, A. D., and Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association*, 5(6):571–582.

Vijayan, V. K., Bindu, K. R., and Parameswaran, L.-h. (2017). A comprehensive study of text classification algorithms. In *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1109–1113.

Wang, T. Y. and Chiang, H. M. (2011). multi-label text categorization problem using support vector machine approach with membership function. *Neurocomputing*, 74(17):3682–3689.

Yang, B., Sun, J. T., Wang, T., and Chen, Z. (2009). Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 916–926.

Zhang, K., Lan, L., Wang, Z., and Moerchen, F. (2012). Scaling up kernel svm on limited resources: A low-rank linearization approach. *Artificial intelligence and statistics*, 22:1425–1434.