Reproducibility and automation of the APPRAISAL taxonomy

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, David Eyers
Department of Computer Science
University of Otago
New Zealand
[pradeesh,andrew,veronica,dme]@cs.otago.ac.nz

Abstract

There is a lack of reproducibility in the results of experiments that apply the APPRAISAL taxonomy. APPRAISAL is widely used by linguists to study how people judge things or people. Automating APPRAISAL could be beneficial for use cases such as moderating online comments. Past work on APPRAISAL annotation has been descriptive in nature, and the lack of publicly available data sets hinders the progress of automation. In this work, we are interested in two things; first, we are interested in how well humans can reproduce the annotation of APPRAISAL of the Australasian Language Technology Association (ALTA) data set. We employed four annotators, each with a similar cultural and linguistics background to reannotate the data set. Second, we are interested in measuring the performance of the existing automated approaches to APPRAISAL classification. Our results show a poor level of agreement at more detailed APPRAISAL categories (Fleiss κ = 0.059) and a fair level of agreement (κ = 0.372) at coarse-level categories. We find similar results when using automated approaches that are publicly available. Our empirical evidence suggests that, at present, automating APPRAISAL classification is practical only when considering coarse-level categories of the taxonomy.

1 Introduction

With the rising popularity of social media platforms, such as Twitter and Facebook, we are experiencing an unprecedented surge of unstructured textual discourse ready to be analysed (Gundecha and Liu, 2012). Supervised learning in Natural Language Processing (NLP) has helped us extract the richness of the information found in these texts for purposes such as sentiment analysis (Zhang et al., 2018), hate-speech detection (Schmidt and Wiegand, 2019), and question answering (Shah et al., 2019). The training task within supervised learning requires high-quality annotated data in order to perform well (Ramas et al., 2021).

Supervised learning can apply theories of evaluative language (Bateman et al., 2019). With high-quality annotated data, we are confident that the task of identifying phrases of evaluative language can be automated. Evaluative language allows us to analyse how we express our feelings, our assessments of people, situations and objects (Benamara et al., 2017). As evaluative language is such a large and intricate discipline, herein we restrict ourselves to just the APPRAISAL[1](Martin and White, 2003) taxonomy within it.

APPRAISAL gives linguists a systematic approach to evaluating language such as identifying and understanding how people make judgements about things (people and objects). The taxonomy has been widely used by linguists to analyse the language choices and attitudes used by writers in order to express their stances (Chen, 2022) in various media such as in news biographies, examiners' reports and tweets (Starfield et al., 2015; Ross and Caldwell, 2020; Su and Hunston, 2019).

Looking forward, APPRAISAL could be used to automate the moderation of online comments (Cavasso and Taboada, 2021); in spite of automation, human moderators continue to be required for the final judgement call on some comments (Ghosh et al., 2011).

The task of reading a comment (and analysing the language) is known to have a negative psychological impact on moderator's mental health (Steiger et al., 2021). Sullivan (2022) argued that the impact on human moderators could be lowered by reducing the number of comments they read. For example, if a comment is identified as having legal implications (based on APPRAISAL analysis),

[1]Small caps are used to distinguish technical, linguistic terms from their use in common parlance.

the comment could be automatically rejected, or flagged for legal review.

To date, much of the research in APPRAISAL annotation has been descriptive in nature (Fuoli, 2018; Fuoli and Hommerberg, 2015). Fuoli (2018) proposed a structured approach known as the *stepwise approach*, however it stops short of providing guidelines to future researchers and has not been quantified. If there are to be robust data sets for studying APPRAISAL then it is necessary to quantify the quality of the data sets already available, examine the practices involved in acquiring them, and strive to improve the techniques in a well grounded, measurable way.

We focus the scope of this work on just the JUDGEMENT subbranch of the APPRAISAL taxonomy because there is publicly available data provided by the Australasian Language Technology Association (ALTA) for their 2020 Shared Task Challenge (Mollá, 2020). Figure 1 shows the part of the APPRAISAL taxonomy that focuses on JUDGEMENT and where it fits in the hierarchy. In this paper, we investigate the following research questions: 1) *the reproducibility and the reliability of humans annotating* JUDGEMENT *sentences*, and 2) *the effectiveness of automated approaches to classify* JUDGEMENT.

In order to answer our research questions, we first employed four annotators, each with a linguistics background, to re-annotate the ALTA data set. Our experiments demonstrate significant levels of disagreement between the annotators at Level 4 of the APPRAISAL hierarchical taxonomy (Fleiss κ = 0.089) as opposed to the more consistent results at Level 2 of the taxonomy (κ = 0.558). We then compare the performance of three different systems submitted to the ALTA challenge. We observe a similar, relative effect: A κ score of 0.031 at Level 4 and 0.206 at Level 2.

Our qualitative analysis of the assessments shows that categorising the exact type of JUDGEMENT can be difficult, as what constitutes morality (a part of JUDGEMENT) is subjective—foresight and background context are required.

We aim to encourage further research into the application of the APPRAISAL taxonomy that is reproducible. This would collectively support our goal to build robust automated approaches to aid SFL practitioners in handling large datasets. To aid in future research, we have made our annotations and experimental data publicly available.[2]

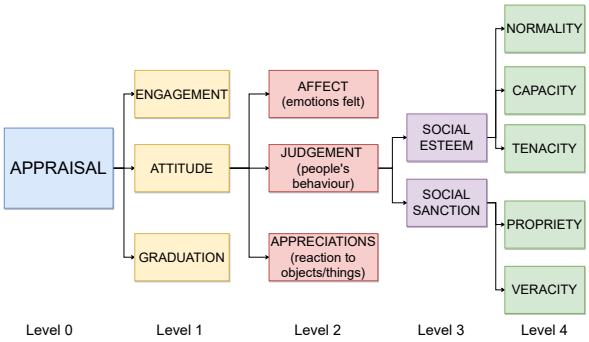

Figure 1: JUDGEMENT branch of the APPRAISAL hierarchical taxonomy (in context, and adapted from (Stewart, 2015, p. 3).)

2 Related Work

The APPRAISAL taxonomy was originally developed by Martin and White (2003) and it is used by linguists to study discourse across a diverse range of genres (Chen, 2022). APPRAISAL captures the evaluative meaning (*opinion*) of the person who wrote a piece of text toward another person or an object. This taxonomy is widely used for many tasks including analysing biographies (Su and Hunston, 2019), Donald Trump's tweets (Ross and Caldwell, 2020), investigating online reviews of South Park (Paronen, 2011), advertising messaging (Beangstrom and Adendorff, 2013) and PhD examiners' reports (Starfield et al., 2015).

There are five levels in the APPRAISAL taxonomy (Figure 1). The levels indicate the granularity of the categories. At Level 1, there are three categories: ATTITUDE (*emotions, ethics and aesthetics*), GRADUATION (*how* ATTITUDE *is being used in a sentence*) and ENGAGEMENT (*writer's openness for negotiation*). The ATTITUDE branch can be broken down further into AFFECT (*emotions*), JUDGEMENT (*ethics*) and APPRECIATION (*aesthetics*). The branch of JUDGEMENT is further divided into SOCIAL ESTEEM and SOCIAL SANCTIONS. SOCIAL ESTEEM deals with admiration and criticism of people, without any legal implication. This branch is further subdivided into NORMALITY (*how closely one follows the norm of the society*), CAPACITY (*how capable the person is*) and TENACITY (*how dependable the person is*). SOCIAL SANCTIONS on the other hand deals with the behaviour of a person that has a legal or

[2] https://github.com/prasys/appraisal-annotation-coling2022

Text	Classification
I am beyond mad that I lost track of a brown spider in my brown carpet. Where did you go?	CAPACITY
Feels like I lost my best friend #lost #fml #missingyou	NORMALITY
@alour @jkramon1313 you should force your neighbours to pay that! Those people have some nerve!!! #Victim	PROPRIETY
I feel like a burden every day that I waste but I don't know how to get out of this	CAPACITY,VERACITY
Instagram seriously sort your sh*t out. I spent ages writing that caption for you to delete it and not post it!! #fume #instagram	*None*
I am about to be a coward and I feel terrible. But I can't even face this	VERACITY

Table 1: Example Tweets from the ALTA data set (Mollá, 2020) and their classification under the JUDGEMENT branch of the APPRAISAL taxonomy.

moral implication. This branch of the taxonomy is further divided into VERACITY (*how truthful one is*) and PROPRIETY (*how ethical the person is*). Some examples of tweets and where they fit within JUDGEMENT are given in Table 1.

There is a substantial amount of work on building automated approaches to applying JUDGEMENT classification to text (Argamon et al., 2007; Bloom and Argamon, 2010; Whitelaw et al., 2005; Neviarouskaya et al., 2010; Taboada et al., 2011). Most of this work focuses on using a combination of machine learning approaches and a hand-built lexicon. For instance, Argamon et al. (2007), constructed lexicons from the seed words of Martin and White (2003) and apply Naive Bayes and Support Vector Machine (SVM) classifiers to the task. They obtain an F_1 score of 0.345. One of the major drawbacks of these past approaches is that the APPRAISAL terms in the lexicons were selected based on researchers' intuition, and it is unclear what these intuitions were. Worse, the lexicons do not appear to be publicly available, thereby making it impossible to reproduce the experiments.

The lack of publicly available APPRAISAL data resulted in a stall in research, but recently the Australasian Language Technology Association (ALTA) organised a shared task and encouraged participants to build an automated system to identify the subclasses of JUDGEMENT used in tweets (Mollá, 2020). They made their data publicly available.[3] The results of the shared task were underwhelming, with the best system obtaining an F_1 score of 0.155 (Aroyehun and Gelbukh, 2020). In reaction, we sought to determine whether humans found this task equally difficult.

Prior work in this area by Read and Carroll (2012), used two annotators to annotate a book corpus and obtained an F_1 score of 0.434 at Level 4 of the APPRAISAL taxonomy, as opposed to an F_1 score of 0.532 at Level 1 of the APPRAISAL taxonomy. Their data do not appear to be publicly available.

Ross and Caldwell (2020) discovered that tweets contain a higher proportion of JUDGEMENT words than AFFECT and APPRECIATION words. JUDGEMENT in tweets is especially insightful to an organisation interested in brand reputation, as an increasing number of consumers are using Twitter to recommend brands to their friends (Vidya et al., 2015). Knowing how a product is being judged can bring insights to an organisation on how that product might be improved. Collectively, these studies highlight the importance of evaluating the reliability of humans can perform the classification and compare the performance with existing approaches.

3 Data set and Annotation

3.1 Data set

We consider the data set from the ALTA 2020 Shared Task (Mollá, 2020). The data set was originally sourced from the SemEval 2018 AIT DISC data set (Mohammad et al., 2018). The ALTA data contains 300 tweets that have been annotated by two linguists, one from the University of Wollongong and the other from the University of New South Wales, and then verified by two other linguists from the same two universities. The data is then split into 200 tweets for training and the remaining 100 tweets for the testing portion. Each tweet was annotated with one or more categories of Level 4 of the APPRAISAL taxonomy. Some examples are given in Table 1.

[3] https://www.kaggle.com/c/alta-2020-challenge/

3.2 Annotation

In order to measure the reproducibility of the annotation process we re-annotated the test portion of the ALTA data set.

By using the test set it was possible to use the training set as guidelines for our annotators, and also to compare the performance of runs submitted to the shared task on multiple sets of annotations.

Annotation for JUDGEMENT is non-trivial and requires an understanding of linguistics. Because of this complexity we employed four human annotators each with a background in linguistics.[4] All annotators are native English speakers and New Zealanders. The detailed background of the annotators is as follows:

- a—an associate professor with 20 years of teaching experience in a languages department at a university in New Zealand,
- b—a graduate student in linguistics currently studying for a Master's in Linguistics at a university in New Zealand,
- c—a final year undergraduate student working towards a degree in applied linguistics at a university in New Zealand, and
- d—a professional language translator and interpreter in New Zealand with over 10 years of experience.

In addition to these four, we have the reverse engineered golden data set denoted (g).[5]

All annotators were aware of the Systemic Functional Linguistic (SFL) theory, which forms the basis of the APPRAISAL taxonomy. However, we also provided the set of guidelines from Martin and White (2003) and the 200 labelled tweets from the training set. Following the recommendations of Fuoli (2018), all annotators worked independently and were of similar cultural and ideological backgrounds. The annotators classified each tweet into one or more of the five categories of Level 4 of the APPRAISAL taxonomy in Figure 1, or marked the tweet as *None* if the tweet was not JUDGEMENT-bearing. Some tweets contained sensitive content, so we provided a *Skip* option, but it was not used. Each annotator was given two hours to complete the task but all finished in under an hour. We did not ask the annotators to identify JUDGEMENT-bearing terms and leave this for future work.

[4] We obtained ethical approval from University of Otago Ethics Committee (Approval No: D20/334).

[5] We contacted the organisers of ALTA Task in order to obtain the test set annotations, however the organisers were not able to release the data. We then reverse engineered the annotations by submitting controlled runs to the competition's automated scoring platform—something that was done with the full knowledge of the task organisers.

Level	Without g	With g
0–2	0.558	0.372
3	0.211	0.182
4	0.089	0.059

Table 2: Fleiss κ score between our annotators with and without g at different levels of the APPRAISAL taxonomy.

κ	Agreement
<0	Less than chance
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–0.99	Almost perfect

Table 3: Interpretation of κ scores (adapted from Landis and Koch (1977)).

4 Human Agreement Level

Annotating JUDGEMENT is difficult as sentences can contain varying degrees of objectivity and subjectivity. How these subjective sentences are finally classified can depend on the background of the annotator (Wiebe et al., 2005; Read and Carroll, 2012). To reduce subjectivity, we follow the recommendations of Read and Carroll (2012) by annotating and reporting the scores at a tweet level (sentence level). We calculate the agreement levels based on the Fleiss κ score and mean F_1 score.

Fleiss κ was chosen as it is appropriate when the given task is subjective (Waseem, 2016; Alm, 2011), but recent work by Delgado and Tibau (2019) recommends avoiding just κ when comparing the performance of automated approaches due to the *kappa paradox* (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990). The *kappa paradox* arises because the κ statistic accounts for agreement by chance—which is low for machine learning approaches that closely follow the patterns in the training data. It is, thus, important to consider F_1 when comparing the performance

Level	(a,b)	(a,c)	(a,d)	(a,g)	(b,c)	(b,d)	(b,g)	(c,d)	(c,g)	(d,g)
0–2	0.877	0.875	0.870	0.535	0.963	0.958	0.580	0.985	0.559	0.555
3	0.648	0.672	0.628	0.391	0.737	0.690	0.475	0.686	0.476	0.370
4	0.402	0.421	0.380	0.221	0.401	0.397	0.196	0.317	0.245	0.132

Table 4: Mean F_1 score between annotators at various levels of the APPRAISAL taxonomy.

of automated approaches. F_1 has also been used in previous work (Mollá, 2020; Read and Carroll, 2012; Argamon et al., 2007).

Table 2 shows the Fleiss κ score of our annotators when we include and do not include the golden data set[6]. Note that Levels 0, 1 and 2 of the APPRAISAL taxonomy are collapsed to Levels 0–2 because we have only annotated JUDGEMENT and below. To interpret the values of κ, we follow the guidelines by Landis and Koch (1977) that have been widely adopted in this research area and are reproduced in Table 3.

In line with the findings of Read and Carroll (2012), our annotators' agreement scores drop as the APPRAISAL classification moves from Level 0 towards Level 4. We explore some of the factors behind this in Section 7.

At Level 4, if we were to include the golden set, g, as an annotator, there is a substantial level of disagreement among the annotators. Similarly, at Level 3, (SOCIAL ESTEEM versus SOCIAL SANCTIONS), we see that our annotators still have a low agreement. However, at Level 0–2, we see moderate agreement between the annotators when the golden data set is not included, and a fair agreement level when the golden data set is included.

Table 4 shows the results when using the mean F_1 score. There we see the scores across annotators are very high at Level 0–2 (with the exception of comparing our annotators with g). As with the κ scores, we see that the scores go down as we traverse through the APPRAISAL taxonomy.

The difference between F_1 and κ scores demonstrates the importance of reporting both scores. Reporting F_1 alone gives the impression that this task is easy (has a high level of agreement) but the F_1 score does not take into account classifications that could have occurred by chance. The κ scores suggest that the task is difficult (has a low level of agreement) at Level 4 of the taxonomy. This sheds light as to why the automated approaches that were built by the participants of the ALTA task performed poorly (Mollá, 2020), as we find even humans disagree at Levels 3 and 4.

[6]We built the collective agreement (intersection) set but found the scores dropped further.

5 Classifiers' Agreement Level

In this section, we discuss our experiments for measuring the effectiveness of automated approaches. We consider all three systems that were used to submit runs to the ALTA 2020 Shared Task because the source code to these is publicly available, and we were able to gain access to the submissions. We did not compare to earlier systems as neither they, nor their word lists, nor the data they trained on are publicly available.

The systems we used are as follows:

- NLP-CIC (Aroyehun and Gelbukh, 2020): An ensemble of logistic regression and ROBERTA classifier.
- OrangutanV2 (Parameswaran et al., 2020): An ensemble of two ALBERT classifiers.
- NITS (Khilji et al., 2020): An ensemble of XGBoost and decision tree classifiers that use pre-trained BERT embeddings.

Mean F_1 scores of each tweet against g are reported in Table 5. The performance of these systems against g is lower than the performance of our annotators against g (e.g., (a,g) in Table 4).

This is hardly surprising, since the systems were trained on a limited amount of data, but the annotators were able to draw from years of experience. Nonetheless, NLP-CIC and OrangutanV2 are able to distinguish JUDGEMENT from non-JUDGEMENT fairly accurately at Level 0–2. NITS may be failing due to the use of XGBoost which is susceptible to outliers, and certainly produces results that differ from the transformer-based models of NLP-CIC and OrangutanV2.

Level	NLP-CIC	OrangutanV2	NITS
0–2	0.605	0.558	0.384
3	0.407	0.389	0.258
4	0.157	0.155	0.132

Table 5: Mean F_1 scores of the automated approaches at each level of the APPRAISAL taxonomy.

6 Performance of Humans and Classifiers

We then proceed to compare the performance of human annotators with the automated approaches. We are interested in knowing at which level of the APPRAISAL taxonomy is there a significant difference between the performance of a human and of a machine.

We picked the best performing human annotator who obtained the highest F_1 score against the golden data set at every level to be evaluated (c).

We determined that scores of c for each tweet are not normally distributed by running the Shapiro–Wilk test (Royston, 1992) and obtaining a p-value of 0.0466 (not normally distributed at the significance level of $p < 0.05$).

Knowing that the data are not normally distributed we chose the Wilcoxon signed-rank test (Woolson, 2007) to test for significant differences between c and each automated system. We ran the test after removing ties, which minimises type one error (McGee, 2018).

Level	NLP-CIC	OrangutanV2	NITS
0–2	0.198	0.221	0.314
3	0.102	0.113	0.214
4	0.038*	0.042*	0.095

Table 6: p-values of Wilcoxon signed-rank test. Asterisks (*) denote statistical significance at ($p < 0.05$).

We present our results in Table 6 which shows the p-values from each test. Our analysis shows that there is a statistically significant difference between c and two of the three systems (NLP-CIC and OrangutanV2) at Level 4 of the APPRAISAL taxonomy. However, we find that there are no statistically significant differences at Levels 0 through to 3. Our evidence suggests that at levels closer to 0, the performance of machines and humans are comparable. However, we cannot confidently say so, as our effect size ($d = 0.15$) is very small. We report our Fleiss κ scores in Table 7. The agreement levels are low, a similar finding to that of Delgado and Tibau (2019).

Level	Without g	With g
0–2	0.120	0.206
3	0.094	0.105
4	0.007	0.031

Table 7: Fleiss κ score between automated approaches and with g at various levels of the APPRAISAL taxonomy.

This may be because these models are based on BERT, and thus they share similar characteristics and produce similar results.

7 Qualitative Analysis

Observing the high level of disagreement among annotators and g (mean κ at Level 4 being 0.199), we examined each disagreement in turn. They all fall into two categories:

- Category 1—our assessors do not agree with each other, and g chooses *None*,
- Category 2—our assessors agree with each other, but not with g. This is only seen at Level 4 of the taxonomy.

Consider the following Category 1 example:

> "*Absolutely love @unqualified but can't listen to it during my commute on the subway because I burst out laughing and people stare! .*"

Annotators a, c and d marked this NORMALITY while a and b marked it CAPACITY (a marked it twice). The golden assessor, g, marked it *None*. When we apply the APPRAISAL taxonomy manually, we observed that the first part of the sentence *"absolutely love @unqualified but can't listen to it during my commute on the subway"* being AFFECT, however in the second part *"because I burst out laughing and people stare"* is JUDGEMENT.

We believe this sentence does contain JUDGEMENT. In many societies, laughing loudly on a train is considered rude, as is listening to music loudly. We agree with the majority of our annotators that this warrants classification as NORMALITY, and if we follow the methodology of Fuoli

(2018), this tweet falls into both categories. All of the automated approaches also predict NORMALITY. Our analysis shows three instances of Category 1. Although, from our visual inspection, these sentences contain JUDGEMENT, we are not sure why g chose *None* (and we have no way to investigate this).

As an example of Category 2:

> "*So disappointed in myself for spending £50 on an outfit for meeting a boy #noselfcontrol #nervous* ."

Annotator g tagged this NORMALITY, disagreeing with all the other annotators who marked it CAPACITY. There are two ways that the sentence can be interpreted, for our annotators, they view the person as not being *capable* of controlling their impulses to purchase an outfit for their date. As for g, the annotators plausibly view it as NORMALITY as it was normal for people to be in their best outfit and behaviour when they are out on a date, thus it was *normal* to spend money on a dress. We believe that both of these categories are correct, especially as there are no well-defined criteria to distinguish between CAPACITY and NORMALITY. To address this issue, we suggest clarifying annotation guidelines given to the assessors, and the criteria used to distinguish the different JUDGEMENT categories. All of the automated approaches predicted PROPRIETY. This is likely to be due to the lack of data—there is only one example in the training data set that has a sentence similar to the sentence above: "*Incredibly shocked and disappointed with @united customer service. Really making me rethink flying with them in the future. #unhappy*". That sentence is marked as PROPRIETY. Our findings reflect the findings of Tayyar Madabushi et al. (2019), who have demonstrated that BERT fails to generalise properly when training and test data are significantly dissimilar even though these data sets are very similar in nature. There are four instances of category 2 disagreement.

Another explanation for discrepancies between our annotators and g is that annotating tweets is tricky because the tweeter and the annotator are subject to different cultural and personal views of what JUDGEMENT is. The data we use consists of tweets from different individuals and we have different individuals assessing them, thus we can expect high levels of disagreement. By comparison, the work of Ross and Caldwell (2020), applies APPRAISAL theory to tweets from one individual (Donald Trump), and so the assessors are able to better understand the message behind the tweet. This could be the reason why much of the previous work on APPRAISAL has focused solely on a particular topic or person and not a generalised situation. One way to address discrepancy among annotators is for the annotators to meet and discuss differences—and to improve the assessment guidelines by writing clear-cut criteria for distinguishing difficult cases. We hope to see more diversified data sets released in the future so that we can validate the generalisability of automated JUDGEMENT approaches.

8 Recommendations & Limitations

Our analyses show that: 1) Annotating Level 4 of JUDGEMENT categories is challenging as there is ambiguity in the interpretation of the text; and 2) The evaluation and the reasoning presented in the APPRAISAL literature are very rarely complete. The latter can be addressed by making the data sets publicly available and also sharing the assumptions and annotation guidelines. It is paramount to have these guidelines as it helps reproduce research results.

Moreover, sharing these rules would make it easier to automate approaches that classify APPRAISAL-bearing sentences. By addressing these gaps, we believe that annotating Level 4 of APPRAISAL would be clearer, although how the rules differ or are similar from one data set to another is yet to be seen.

Our work has limitations. Our data set is small and focuses solely on Twitter and the JUDGEMENT branch of APPRAISAL. We find that performing discourse analysis can be challenging especially when it is related to judging the morality of a person from a single tweet. This can surface different perspectives based on different assumptions based on a tiny piece of text (Lachmar et al., 2017).

We evaluated our automated approaches using systems built for the ALTA Shared Task. All of these models were based on BERT and we believe that these models can be further improved by fine-tuning, which has been shown to improve performance elsewhere (Xin et al., 2021). Another plausible cause of the poor performance of these models is that the data set that was trained on and evaluated on was small, and thus was likely to contain bias.

As a direction for future work, it would be inter-

esting to revisit some of the techniques from earlier work mentioned in Section 2. Of course, the lexicons would need to be reproduced as closely as possible, but we hypothesise that the deep learning models could be further improved by using such techniques.

9 Conclusions

In this study, we investigated two topics: (1) reproducibility and the reliability of a popular SFL taxonomy, APPRAISAL, focusing on JUDGEMENT annotation, and (2) the effectiveness of automated approaches to assessing JUDGEMENT. To carry out our investigation in a systematic manner, we employed four linguists to carefully re-annotate the publicly available ALTA 2020 Shared Task data set and used three publicly available, automated approaches. We then performed experiments quantifying and evaluating the performance of our annotators and automated approaches.

We find a low level of agreement when annotating JUDGEMENT despite using annotators with linguistics backgrounds. We obtained a Fleiss κ score of 0.059 when using Level 4 (most detailed categories) within the APPRAISAL's hierarchical taxonomy as opposed to 0.372 when using Level 2 (coarse-grained categories).

We find a similar pattern with the automated approaches. We obtain a Fleiss κ score of 0.031 at Level 4 and 0.206 at Level 2 of the taxonomy. Although the low κ score in automated approaches is attributed to the nature of κ statistical penalising agreement not occurring by chance, our F_1 score (0.605 at Level 2 and 0.155 at Level 4) supports our earlier findings that humans find classifying JUDGEMENT to be difficult. Furthermore, we find that there is no statistical significance between the performance of our best performing annotators and of the best performing system when working with Level 2 category of the APPRAISAL taxonomy, thus we argue that automation of JUDGEMENT is possible at this level as automated systems are already performing at human levels.

Our analyses sheds light on the challenges in reproducibility of APPRAISAL annotation. We believe that the poor scores of human annotators and automated approaches are due to a multitude of factors including the lack of publicly available data sets (examples), the absence of details such as prior assumptions made by the annotators, and the lack of generally available clear and concise annotation guidelines.

We have publicly released our data and analysis to encourage more research into APPRAISAL. We believe that the application of APPRAISAL to tweets and other discourse will enable the application of APPRAISAL to other domains (such as eCommerce).

Acknowledgements

We wish to thank the New Zealand eScience Infrastructure (NeSI) for providing the infrastructure that we used to conduct our experiments.

References

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Shlomo Argamon, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2007. Automatically determining attitude type and force for sentiment analysis. In *Language and Technology Conference*, pages 218–231. Springer.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2020. Automatically predicting judgement dimensions of human behaviour. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 131–134.

John A Bateman, Daniel McDonald, Tuomo Hiippala, Daniel Couto-Vale, Eugeniu Costechi, et al. 2019. Systemic-functional linguistics and computation: new directions, new challenges. *The Cambridge Handbook of Systemic-Functional Linguistics*.

Tracy Beangstrom and Ralph Adendorff. 2013. An appraisal analysis of the language of real estate advertisements. *Southern African Linguistics and Applied Language Studies*, 31(3):325–347.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Kenneth Bloom and Shlomo Argamon. 2010. Unsupervised extraction of appraisal expressions. In *Canadian Conference on Artificial Intelligence*, pages 290–294. Springer.

Luca Cavasso and Maite Taboada. 2021. A corpus analysis of online news comments using the appraisal framework. *Journal of Corpora and Discourse Studies*, 4:1–38.

Muxuan Chen. 2022. An appraisal analysis of Sina Weibo texts about reforms of undergraduate education. *Scientific and Social Research*, 4(1):1–16.

Domenic V Cicchetti and Alvan R Feinstein. 1990. High agreement but low kappa: II. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Matteo Fuoli. 2018. A stepwise method for annotating appraisal. *Functions of Language*, 25(2):229–258.

Matteo Fuoli and Charlotte Hommerberg. 2015. Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, 10(3):315–349.

Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176.

Pritam Gundecha and Huan Liu. 2012. Mining social media: a brief introduction. *New directions in informatics, optimization, logistics, and production*, pages 1–17.

Abdullah Faiz Ur Rahman Khilji, Rituparna Khaund, and Utkarsh Sinha. 2020. Human behavior assessment using ensemble models. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 140–144.

E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. 2017. # mydepressionlookslike: Examining public discourse about depression on twitter. *JMIR Mental Health*, 4(4):e8141.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Monnie McGee. 2018. Case for omitting tied observations in the two-sample t-test and the Wilcoxon-Mann-Whitney Test. *PloS one*, 13(7):e0200837.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Diego Mollá. 2020. Overview of the 2020 ALTA Shared Task: Assess human behaviour. *ALTA 2020*, page 127.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*, pages 806–814.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2020. Classifying judgements using transfer learning. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 135–139.

Tiina Paronen. 2011. *Appraisal in Online Reviews of South Park: A study of engagement resources used in online reviews*. Ph.D. thesis, University of Jyväskylä.

Jose Garrido Ramas, Giorgio Pessot, Abdalghani Abujabal, and Martin Rajman. 2021. Identifying and resolving annotation changes for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 10–18.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in English. *Language resources and evaluation*, 46(3):421–447.

Andrew S. Ross and David Caldwell. 2020. 'Going negative': An APPRAISAL analysis of the rhetoric of Donald Trump on Twitter. *Lang. Commun.*, 70:13–27.

Patrick Royston. 1992. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, 2(3):117–119.

Anna Schmidt and Michael Wiegand. 2019. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, pages 1–10. Association for Computational Linguistics.

Asad Ali Shah, Sri Devi Ravana, Suraya Hamid, and Maizatul Akmar Ismail. 2019. Accuracy evaluation of methods and techniques in web-based question answering systems: a survey. *Knowledge and Information Systems*, 58(3):611–650.

Sue Starfield, Brian Paltridge, Robert McMurtrie, Allyson Holbrook, Sid Bourke, Hedy Fairbairn, Margaret Kiley, and Terry Lovat. 2015. Understanding the language of evaluation in examiners' reports on doctoral theses. *Linguist. Educ.*, 31:130–144.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the*

2021 CHI conference on human factors in computing systems, pages 1–14.

Martyn Stewart. 2015. The language of praise and criticism in a student evaluation survey. *Studies in educational evaluation*, 45:1–9.

Hang Su and Susan Hunston. 2019. Language patterns and attitude revisited: Adjective patterns, Attitude and Appraisal. *Functions of Language*, 26(3):343–371.

Mark Sullivan. 2022. Facebook is expanding its tools to make content moderation less toxic.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. Association for Computational Linguistics.

Nur Azizah Vidya, Mohamad Ivan Fanany, and Indra Budi. 2015. Twitter sentiment to analyze net brand reputation of mobile phone providers. *Procedia Computer Science*, 72:519–526.

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

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. BERxiT: Early exiting for BERT with better fine-tuning and extension to regression. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 91–104, Online. Association for Computational Linguistics.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.