Classifying JUDGEMENTS using Transfer Learning

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

We describe our method for classifying short texts into the APPRAISAL framework, work we conducted as part of the ALTA 2020 shared task. We tackled this problem using transfer learning. Our team, "*orangutanV2*" placed equal first in the shared task, with a mean $F1$-score of 0.1026 on the private data set.

1 Introduction

Systemic Functional Linguistics (SFL) is a theory of language which examines the relationship between language meaning and the functions in a social context (Halliday, 1996). One popular framework that uses SFL is APPRAISAL (Martin and White, 2005). The APPRAISAL framework is based on the notion of uncovering the attitude of the author from the perspective of a potential listener or reader. It is used by linguists in analysing human behaviour from textual data (Ross and Caldwell, 2020; Starfield et al., 2015; Wu, 2013; Hommerberg and Don, 2015). Figure 1 shows an overview of the APPRAISAL framework.

The three main resources of the APPRAISAL framework are; ATTITUDE, ENGAGEMENT and GRADUATION (Martin and White, 2005). The ATTITUDE framework is then subdivided into three subsystems; AFFECT (emotions), APPRECIATION (evaluation of natural and semiotic phenomena) and JUDGEMENT (evaluation of people and their behaviour). The JUDGEMENT subsystem can be divided into two categories: SOCIAL ESTEEM and SOCIAL SANCTIONS. SOCIAL ESTEEM primarily involves admiration and criticism and SOCIAL SANCTION involves praise and condemnation.

SOCIAL SANCTIONS can be further divided into three subcategories: *normality* (how usual one is), *capacity* (how capable one person is) and *tenacity* (how dependable one is). As for SOCIAL SANCTION it can be further divided into two subcategories; *veracity* (how truthful one is) and *propriety* (how virtuous one is).

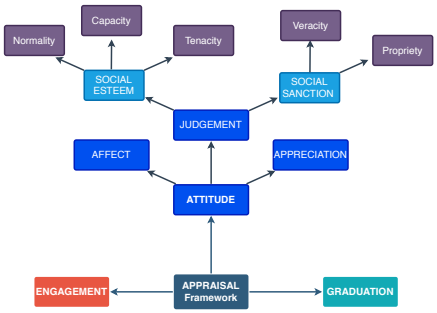

Figure 1: The APPRAISAL Framework (Adapted From (Martin and White, 2005))

The robustness of the APPRAISAL framework lies in its ability to be used in various different social contexts. It also offers linguists detailed strategies for realising the framework (Ngo and Unsworth, 2015). Since its debut, the APPRAISAL framework has been widely used to explore how language is being used in various different environments such as in analysing examiners' reports on doctoral theses (Starfield et al., 2015), Donald Trump's rhetoric tweets (Ross and Caldwell, 2020), people's perception on the outcome of the Brexit referendum (Bouko and Garcia, 2020) and in teaching English as a second language (ESL) (Ngo and Unsworth, 2015).

Currently, linguists manually classify sentences using annotation software as there is no automated classification technique that exists to automate the task (Fuoli, 2018). Thus, this problem sparked the interest of Australasian Technology Association (ALTA) to organise a shared challenge task to develop a model that can automatically identify and classify human behaviour (JUDGEMENT) ex-

Feature matched	Number
(None)	104
Normality	22
Capacity	31
Tenacity	21
Veracity	2
Propriety	33
Multiple Features	74

Table 1: Distribution and Pattern of Training Data

pressed in tweets on Twitter (Molla, 2020). The task was to classify tweets into either one or more (or none) of the five sub-categories of JUDGEMENT.

We present our participation in this challenge. We tackled this problem by utilising a pre-trained transfer learning model, ALBERT (Lan et al., 2019), as a classifier.

2 Data Set

The data set[1] provided by the organisers is a collection of 300 tweets from SemEval 2018 (Mohammad et al., 2018): 200 tweets for training and 100 for testing. The training data set consists of tweet ID and the labelled annotations of sub-categories of JUDGEMENT which the tweet belongs to. If the tweet does not contain any sub-categories it is marked as blank.

We analysed the training data to understand the distribution and the patterns of category use. Table 1 describes the pattern. The data set is not balanced between categories, particularly for *Veracity* where there are only have 2 examples in the training set. We have also found that there is 1 duplicate tweet in the training data and we promptly informed the organisers of this. Additionally we found 22 of the tweets in training data are in the testing data.

3 Methodology

First we handle class imbalances followed by pre-processing of our tweets. Then, we perform unsupervised classification by utilising ALBERT's pre-trained model. We chose ALBERT because it performs reasonably well on various different tasks such as offensive language detection (Zampieri et al., 2020), multiple-choice reading comprehension (Si et al., 2019), and question and answering (Khashabi et al., 2020). Finally, we employed the cosine similarity measure in order to correct the mistakes made by our classifier.

We have made our system's source code publicly available on Github.[2]

3.1 Handling Class Imbalance

Due to a low number of training examples for *Veracity*, we removed this category from our examples (and, consequently, results). Early experiments showed that this led to a significant performance improvement. We did not make any adjustments to any other categories.

3.2 Pre-Processing Data

We experimented with various pre-processing strategies of including stemming, removing mentions, hashtags and URLs. From our early experiments, we found that by removing mentions from the tweets, and keeping the text as is, yielded the best performance.

3.3 ALBERT Transfer Learning Classifiers

We used huggingface's[3] implementation of ALBERT. We then added a sigmoid classifier (for binary classification) or softmax classifier (for multi-label classification) on top of the model to predict the probability of a category. We built three separate classifiers using this model; a binary classifier for SOCIAL SANCTIONS (C_s), a binary classifier for SOCIAL ESTEEM (C_e) and a multi-label classifier to classify the potential categories the tweet belongs to (C_m).

First we feed our pre-processed data into C_s and C_e. Once the texts get classified to be either both or one of the categories, we continue to feed it to C_m in order to get the potential granular categories.

We evaluated the performance of our classifier by splitting our training data into 70% for training, 10% for validation and 20% for evaluation purposes. We used the Adam Optimizer with a learning rate of 10^{-5} for 50 epochs. We set the batch size to be 64. We used our validation set's mean $F1$ score as an early stopping criterion. We stop the training if the score does not increase for 15 consecutive epochs, or the maximum number of epochs has been reached.

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

[2] https://github.com/prasys/alta2020-appraisal

[3] https://huggingface.co/

Figure 2 shows the recall and precision scores of the three different classifiers on our evaluation set. We set the class probabilities confidence level to be 0.5 in order to maximise precision and recall scores. Our C_s classifier in Figure 2a obtained both recall and precision score of 73.68% and C_e classifier in Figure 2b obtained both recall and precision score of 93.71%. As for our multi-label classification in Figure 2c we obtained a precision score of 56.25% and a recall 42.86%.

From visual inspection of the training data and our result, we observed that the C_m classifier can be further improved by adding personal pronoun detection of third person pronouns. We encoded this feature as a binary value. We used Google Natural Language Processing API[4] to extract the pronouns. Then we append the values in the final layer of our model before the softmax classifier.

3.4 Document Cosine Similarity

In the test set we observed the presence of 22 prelabeled tweets from the training set. To correct classifier mistakes, we used the Universal Sentence Encoding (Cer et al., 2018) to perform cosine similarity between the training data and the test data. Our solution was generic. We converted tweets into a high dimension vector representation, computed the cosine similarity with the training data, and those above a given threshold were considered to be the correct answer. We set the threshold to 1 in order to catch only exact matches.

4 Results

Kaggle was used as the platform for run submission. In Kaggle, the test data provided to us by the ALTA organisers is split into public (public leaderboard) and private (private leaderboard). The private portion serves as a validation portion in order for the organisers to determine the effectiveness of all systems. The scores are evaluated using mean $F1$. We summarise and present our results in Table 2.

4.1 Discussion

The objective set by the organisers of the shared task at ALTA was to create a baseline for this task.

(a) C_s Classifier

(b) C_e Classifier

(c) C_m Classifier

Figure 2: Recall and Precision Scores of C_s, C_e & C_m Classifiers

System	Public Score	Private Score
C_m (Baseline)	0.19333	0.06133
C_m + Cos. Similarity	0.20333	0.08133
C_s + C_e + C_m + Cos. Similarity	**0.21333**	0.08133
C_s + C_e + C_m + Cos. Similarity + Pronoun	0.20000	**0.10266**

Table 2: System Evaluation

[4] https://cloud.google.com/natural-language/docs

Although our system placed 2nd on both private and public leader boards, statistical tests (run by the challenge organisers) showed no statistically significant difference between the scores of our team and the team that got a slightly higher score. Both were declared joint winners.

Our further investigations suggest that our system performed well at identifying SOCIAL SANCTIONS, probably because the important words in the tweets appear close to each other in the vector space.

Equally, we are not performing well at classifying SOCIAL ESTEEM. Although, our binary classifier is able to classify tweets belonging to SOCIAL ESTEEM with a high degree of accuracy, we are not able to classify them accurately at a sub-categorical level. This prompted us to look deeper into the problem and to offer several ways to improve this task – which we discuss in subsections 4.2 and 4.3.

4.2 Lack of Training Data

The primary difficulty in achieving higher accuracy in classifying tweets is the limited amount of training data available (Lu et al., 2014). Whilst acknowledging the fact that annotating a large set of data manually is challenging (Ciravegna et al., 2002), we propose that a smaller data set such as the one being used for this task should be tailored to be a specific topic rather than being spread across multiple topics. For instance, if the topic were the recent New Zealand elections, we may be able to improve the performance of the classifier by augmenting it with domain knowledge obtain from news sources or the Wikipedia (Yangarber et al., 2000; Gabrilovich and Markovitch, 2006). This is similar to how humans used domain knowledge to resolve ambiguity in evaluating the APPRAISAL framework with Trump's tweets (Ross and Caldwell, 2020).

4.3 The Annotation Process

Identifying expression of APPRAISAL in a piece of text is not as straightforward as some discourse analysis tasks (Mauranen and Bondi, 2003). Although Martin and White (2005) discuss the framework in detail and provides examples, there is the potential for the "Russian doll syndrome" (Thompson, 2014), where classifying into one category can be interpreted as indirectly classifying into other categories. This creates a problem in providing a reliable annotation. We show two examples from the training data set, with the provided categories in bold—

> "*@Gennneral thanks gen!! Love you miss you happy birthday natong duha* ❤️ ❤️ ❤️." ("**None of the above**")

> "*@priny_baby happppy happpppyyyyy happpppppyyyyy birthday best friend!! Love you lots #chapter22* ❤️ ❤️ ❤️." ("**Normality**")

In the first example, the annotators classified it as being none of the 5 categories, whereas in the second example this was not the case. In both cases, our system predicted *Normality*. Our deep learning model was not able to accurately distinguish between these two tweets. We hypothesise that humans face a similar difficulty with these two tweets, and may not choose deterministically. One way of addressing ambiguity is to follow Fuoli (2018) and to use a step-wise method to ensure reproducibility of annotations. We plan to explore this further as part of our future work.

5 Conclusion and Future Work

We presented our approach in order to automatically identify and classify JUDGEMENT expressed in textual segments. We competed in the ALTA 2020 challenge under the team name of "*orangutanV2*" and placed equal first. Our best-performing system used a combination of transfer learning and document cosine similarity.

Despite setting a baseline for future work, we believe that there is still much work to be done in this area. As part of future work we are planning to tackle this problem in several ways, including:

- Looking at human level performance; and
- Experimenting with various different transfer learning models.

Acknowledgements

We would like to thank Google Cloud for their support in providing the infrastructure to conduct our experiments. We would also like to thank the ALTA organisers (especially Dr. Diego Molla-Aliod) for organising the challenge, and promptly replying to our enquiries.

References

Catherine Bouko and David Garcia. 2020. Patterns of Emotional Tweets: The Case of Brexit After the Referendum Results. In *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, pages 175–203.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*.

Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli, and Yorick Wilks. 2002. User-system cooperation in document annotation based on information extraction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2473(October):122–137.

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

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the National Conference on Artificial Intelligence*.

Michael AK Halliday. 1996. Literacy and linguistics: A functional perspective. In *Literacy in society 339*, page 376.

Charlotte Hommerberg and Alexanne Don. 2015. Appraisal and the language of wine appreciation: A critical discussion of the potential of the Appraisal framework as a tool to analyse specialised genres. *Functions of Language*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. In *arXiv preprint arXiv:2005.00700*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *arXiv preprint arXiv:1909.11942*, pages 1–17.

Zhongqi Lu, Yin Zhu, Sinno Jialin Pan, Evan Wei Xiang, Yujing Wang, and Qiang Yang. 2014. Source free transfer learning for text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

J. R. Martin and P. R.R. White. 2005. *The Language of Evaluation*.

Anna Mauranen and M. Bondi. 2003. Evaluative language use in academic discourse. *Journal of English for Academic Purposes*.

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

Diego Molla. 2020. Overview of the 2020 alta shared task: Assess human behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association.*

Thu Ngo and Len Unsworth. 2015. Reworking the appraisal framework in ESL research: refining attitude resources. *Functional Linguistics*, 2(1):1–24.

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

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? *arXiv preprint arXiv:1910.12391*.

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. *Linguistics and Education*, 31:130–144.

Geoff Thompson. 2014. Affect and emotion, target-value mismatches, and Russian dolls. In *Evaluation in context*, volume 47, page 66. John Benjamins Amsterdam.

Hai-bin Wu. 2013. Appraisal Perspective on Attitudinal Analysis of Public Service Advertising Discourse. *English Language and Literature Studies*.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic acquisition of domain knowledge for Information Extraction. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *arXiv preprint arXiv:2006.07235*.