Automatically Predicting Judgement Dimensions of Human Behaviour

Segun Taofeek Aroyehun
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
aroyehun.segun@gmail.com

Alexander Gelbukh
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
www.gelbukh.com

Abstract

This paper describes our submission to the ALTA-2020 shared task on assessing behaviour from short text, We evaluate the effectiveness of traditional machine learning and recent transformers pre-trained models. Our submission with the Roberta-large model and prediction threshold achieved first place on the private leaderboard.

1 Introduction

Language enables us to express evaluation of people, action, event, and things. This manifests as emotion and assessment of human behaviour and artefacts. The study of evaluative language has benefited from efforts in several disciplines such as linguistics, philosophy, psychology, cognitive science and computer science (Benamara et al., 2017). In linguistics, the appraisal framework of Martin and White (2003) provides a detailed classification scheme for understanding how evaluation is expressed and implied in language. In computer science, affective computing study evaluative language under the umbrella term of sentiment analysis with common tasks involving detection and classification of polarity and emotion, and aspect-based sentiment analysis, among others. Sentiment analysis has benefited from the availability of user-generated content on online platforms.

The theory of appraisal proposed by Martin and White (2003) has three categories of evaluative text: affect, judgement, and appreciation. These categories respectively model opinions in terms of emotions, norms, and aesthetics. Utterances are viewed as indicating positive ("praising") or negative ("blaming") disposition towards some object (person, thing, action, situation, or event). The judgement dimensions are normality, capacity, tenacity, veracity, and propriety. Each of the dimensions represents an answer to the following corresponding questions:

- Normality: How special?
- Tenacity: How dependable?
- Capacity: How capable?
- Veracity: How honest?
- Propriety: How far beyond reproach?

The corpus used in this paper is annotated with the above judgement dimensions.

Taboada and Grieve (2004) automatically categorized appraisal into affect, judgement, and appreciation using a lexical approach that groups adjectives according to their semantic orientation. Benamara et al. (2017) surveyed linguistic and computational approaches to the study of evaluative text. Their analysis suggested that appraisal is a richer and more detailed task amenable to computational approaches subject to availability of data. They envision that appraisal analysis can contribute to the advances in affective computing. Recently, Hofmann et al. (2020) showed that dimensions of appraisal can improve emotion detection in text. A similar observation was made by Whitelaw et al. (2005) who found appraisal phrases as useful features for sentiment analysis.

This paper investigates the capabilities of machine learning models in predicting the dimensions of human judgement expressed in short texts (tweets) as part of the ALTA-2020 shared task on assessing human behaviour (Mollá, 2020). The task aims to advance computational techniques for analysing evaluative language.

The use of neural networks has lead to significant performance improvements in NLP tasks. However, neural networks require a large amount of labeled data. On the contrary, the traditional machine learning models such as NBSVM are competitive in low-data regimes (Wang and Manning, 2012; Aroyehun

Label	Normality	Capacity	Tenacity	Veracity	Propriety
Proportion	0.11	0.16	0.11	0.015	0.18

Table 1: Frequency of each label in the training set as a fraction of the total number of examples.

and Gelbukh, 2018). The recently introduced contextual representation learning models (Peters et al., 2018; Devlin et al., 2019) are pre-trained with language modeling objective on a large and diverse collection of text. The learned representation can be transferred to downstream tasks via fine tuning (Howard and Ruder, 2018). We examine the effectiveness of using NBSVM and fine tuning a Roberta-large model (Liu et al., 2019) for predicting dimensions of judgement expressed in short text.

2 Methodology

Task. Given a short text predict one or more judgement dimensions expressed in the given text. This is a multilabel classification problem where the labels consist of the five judgement dimensions.

Data. We employed the data provided by the organizers of the ALTA-2020 shared task (Mollá, 2020). The training set has 198 tweets. Each example is annotated with the presence or absence of each of the judgement dimensions as outlined in Section 1. Table 1 shows the proportion of each label in the training set. The proportion ranges from 2% to 18%. The test set consists of 100 examples. About 50% each is used for the public and private leaderboards for the competition on Kaggle[1] In-class platform.

The private leaderboard is used for the final ranking, the scores are available after the completion of the competition while the public leaderboard is used by the competition participants to evaluate their models during the competition. In our experiment using the Roberta-large model, we created a validation set by randomly sampling 10% of the training set.

Data Pre-processing. We clean the text of each tweet by removing punctuation marks, digits, and repeated characters. We normalize URLS and usernames (tokens that starts with the @ symbol). Hashtags are converted to their constituent word(s) after removing the # symbol.

NBSVM. Wang and Manning (2012) proposed a support vector machine (SVM) model that uses the naive bayes log-count ratio as features. NBSVM is a strong linear model for text classification. In our implementation we use the logistic regression classifier in place of the SVM. The features are based on word n-grams (unigrams and bigrams). We experiment with and without the data pre-processing step. In the multi-label classification setting, we train a binary classifier per label with the same classifier settings.

Roberta-large. An optimized BERT (Devlin et al., 2019) model trained for longer and on larger and more diverse text collection totalling 160GB. In addition, the pre-training tasks did not include next sentence prediction and the tokenizer is based on BPE (Liu et al., 2019). We fine tune the model on the data provided by the task organizers without the data pre-processing step. We used the simpletransformers library [2] for our experiment. The classifier is a linear layer with sigmoid activation function. The hyperparameters are: maximum learning rate of $4e-5$, number of epochs is 20 with early stopping on the validation loss using a patience of 3, batch size of 64, the model parameters are optimized using AdamW with a linear schedule and a warm up steps of 4 and the maximum sequence length is 128.

Prediction threshold. Lipton et al. (2014) studied the difficulty of relating the maximum achievable F1 score with the decision thresholds on predicted conditional probabilities. They observed that selecting predictions that maximize the F1 score is a function of the conditional probability assigned to an example and the distribution of conditional probabilities for other examples. Following this observation, we choose decision threshold for each label to track the distribution of conditional probabilities on the validation set without reference to the gold labels, to avoid overfitting. The default decision threshold is 0.5 and we find that the conditional probabilities are significantly less. We apply this heuristic to the model outputs of the Roberta-large model. Specifically, we set 0.2 as the decision

[1] https://www.kaggle.com/

[2] https://simpletransformers.ai/

Method	Public leaderboard	Private leaderboard	Average
NBSVM	**0.16000**	0.00000	0.08000
NBSVM w/ prep.	0.16000	0.00000	0.08000
Roberta-large	0.11666	0.06666	0.09166
Roberta-large w/ threshold	0.14285	**0.15466**	**0.14876**

Table 2: Mean F1 score on the public and private test sets. Average is the unweighted mean of the scores on the private and public leaderboards as they are approximately 50% each of the test set.

threshold for the *capacity* label and 0.1 for the remaining labels.

3 Results

Table 2 shows the results obtained on the test set split into two equal halves as the public and private leaderboards. With the NBSVM model, we achieved the best score of 0.16 on the public leaderboard. The application of data pre-processing step did not impact the performance of the NBSVM model, probably because the tokens removed are not relevant lexical units for the task. Following this observation, we did not apply the pre-processing step to our experiments with the Roberta-large model. The Roberta-large model obtained a relatively lower score on the public leaderboard and appears to generalize better on the other half of the test set as shown by the scores on the private leaderboard. There is a significant performance improvement due to the decision thresholding on the Roberta-large model outputs. With this strategy, we achieved the best overall score on the ALTA-2020 competition.

4 Conclusion

We address the task of automatically predicting judgement dimensions in the context of the ALTA-2020 shared task. We evaluated the performance of a strong linear classifier, NBSVM with n-grams as features and a recent pre-trained language model, Roberta-large. We observed that the NBSVM achieves our best score on the public leaderboard but it did not generalize to the private test set. The Roberta-large model with decision thresholding strategy showed consistent performance on both the public and private leaderboards. With this model, we achieved the best overall score on the competition.

While we achieved better performance with the Roberta-large model, we think that the statistical power (Card et al., 2020) of the test set is limited due to the small sample size (100 examples). As such, it is difficult to differentiate performance improvement by chance from substantial model advantage. We hope to test our approaches on a larger test set in order to examine the robustness of our approaches.

Acknowledgments

The authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

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.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. *arXiv preprint arXiv:2003.14155*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize f1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Diego Mollá. 2020. Overview of the 2020 ALTA Shared Task: Assess Human Behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Sida I. Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

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.