Adapting BERT for ‘Apples to Apples’ Gameplay

Authors

Arikka Cherniwchan (MacEwan University; cherniwchana7@mymacewan.ca)
Austin Countaway (MacEwan University; countawaya@mymacewan.ca)
Chunyang Ding (MacEwan University; dingc4@mymacewan.ca)
Brandon Funk (MacEwan University; funkb22@mymacewan.ca)
Calin Anton (MacEwan University; antonc@macewan.ca)
DOI: 10.1145/3774399.3774404

Introduction

“Apples to Apples” is a word association game where players match red noun cards with a given green adjective card in the hopes of winning the round through the selection of their card by the round judge. There are many possible selection criteria for determining which red card to play in a round. Combinations may focus solely on the direct association between a red and green card, on the humour elicited by a card pair, or through a wide variety of other selection criteria. The game’s difficulty comes from the anticipation and matching of the card played by a player to the criterion of the current round judge. Our aim was to create an agent capable of playing “Apples to Apples,” able to discover the current judge criterion and play the card in hand that best matches that criterion. To this end, we decided to explore the use of BERT pretrained models and evaluate their viability in word game agents. In this study, we leveraged fine tuned BERT models and developed a Naive Bayes classifier to simulate and predict judge personalities in “Apples to Apples.” Beyond these main contributions, this work also involved significant efforts in creating supporting tools for both training and testing processes. Specifically, we manually annotated nearly 1000 card pairs using a custom annotation program to reduce human error and fatigue and developed an environment testing tool to ensure platform consistency during training and testing. These contributions highlight our focus on both data quality and system robustness.

Background and Related Work

Since the introduction of transformers (Vaswani et al., 2017), they have been implemented and shown high performance in a variety of Natural Language Processing (NLP) tasks (Patwardhan, Marrone, & Sansone, 2023), becoming one of the go to architectures in NLP (Lin, Wang, Liu, & Qiu, 2022)). BERT (Devlin, Chang, Lee, & Toutanova, 2019) added greater functionality and performance to the base transformer on a variety of NLP tasks as well (Koroteev, 2021). BERT has shown strong performance in comparison to traditional machine learning text classification (Garrido-Merchan, Gozalo Brizuela, & Gonzalez-Carvajal, 2023) and has been used with high performance in tasks related to sentiment analysis (Sayeed, Mohan, & Muthu, 2023). Zhang et al. (Zhang et al., 2020) extend the application of BERT to semantic association, proposing a model, SemBERT, offering improved capability to the base BERT model in the areas of reading comprehension and language inference. Reimers and Gurevych (Reimers & Gurevych, 2019) offer another extension, SBERT, allowing for greater efficiency in sentence comparisons in a much shorter time. Previous success found

using BERT in NLP and semantic association tasks offers a strong motivation for adapting a similar pre-trained model for use in an “Apples to Apples” agent. While there has not yet been a published study adapting BERT to this purpose, a few instances of using pre-trained models for word games require comparable semantic similarity analysis. Koyyalagunta et al. (Koyyalagunta, Sun, Draelos, & Rudin, 2021) implemented an algorithm to generate clues for use in the game “Codenames” using various embedding methods, one of which was BERT. Although BERT underperformed compared to other word embedding models, their work highlights potential avenues to explore in adapting a pre-trained model to “Apples to Apples.” The work of Koyyalagunta et al. (Koyyalagunta et al., 2021) also offers some insight into potential areas of improvement when using a BERT pre-trained model with respect to contextual embeddings and fine-tuning. To adapt BERT to match the “Codenames” scoring metric, their team used an average of the BERT contextual embeddings for each word as that word’s new embedding. Contextual embeddings for BERT perform similarly to non-contextual models for large training sets but show increased performance with ambiguous words and words unseen in training (Arora, May, Zhang, & Re´, 2020). Maintaining the contextual embeddings during implementation in “Apples to Apples” may allow for a more versatile agent that can better work with untrained data. Additionally, their use of a pre-trained model as the word embedder without further task-specific fine-tuning may have led to suboptimal results in the model performance. Fine-tuning improves task performance, though excessive divergence from training and test sets may lead to suboptimal results as well (Zhou & Srikumar, 2022). Ensuring sufficient fine-tuning to increase task specificity while avoiding excessive divergence between training and test sets may increase agent performance. The winning card in “Apples to Apples” is often the one that the judge finds funny in some way. Playing this card involves detecting and ranking humorous associations between card pairs with respect to the judge’s sense of humour. BERT has shown success in the realm of humour detection using a generic pre-trained model and a corpus of annotated tweets (Mao & Liu, 2019), suggesting a model may be used for both semantic association and humour with changes to fine-tuning to reflect the judge. A particular challenge in creating a playing agent is adapting to previously unseen cards or card pairings. Zero-shot learning is one method that may ameliorate this problem. The capacity for BERT to understand context may allow for the incorporation of zero-shot learning, which has shown successful implementation in the word game “Taboo” (Isaak, 2022). Contextual understanding of each card would allow BERT to make associations with Adjective-Noun pairs, even if that pair had not been seen in training.

Agent Objectives

In “Apples to Apples,” a pivotal skill is the player’s ability to adapt to the current round judge, playing card pairs that best match the theme or judging method used by that judge in previous rounds. The creation of an "Apples to Apples” playing agent, therefore, necessitates mimicking the adaptation shown by human players to achieve high performance. Modelling various personalities that the judges may display offers one potential starting point to facilitate this adaptation. Refining the model by selecting common, broad personalities, collecting data for those personalities, and training the model on that data would help in the creation of archetypes to serve as a foundation for potential personality models. After the creation of the personalities, the agent would need some method of determining which personality archetype a judge best matches during gameplay. Modelling and predicting user personality has been used in a variety of applications to achieve high success, such as advertising applications (Shumanov, Cooper, & Ewing, 2022). Their team focused on Big Five personality traits in the realm of marketing, showing an overall increased effectiveness

of their advertising program after incorporating personality prediction. The use of deep learning (An & Levitan, 2018) and BERT-based (Lucky, Zain Nabiilah, Hendrik Jeremy, & Suhartono, 2023) models have shown similar success in predicting Big Five personality traits, further motivation for incorporating personality modelling and prediction in other platforms. A core facet of our approach was the belief that the incorporation of judge personality detection to an "Apples to Apples" agent would likely rely on more game-specific personality categories to apply to as many players as possible. The easiest personality to implement is a semantic association judge, favouring card pairs that directly relate. BERT models for semantic as sociation. (A. Rodriguez & Merlo, 2020) and (Delmonte & Busetto, 2022) have shown that comparisons using word embeddings and co sine similarity replicate the human association between words well, offering a good basis for a primary personality model. We reasoned that further judge archetypes would achieve success in focusing on various forms of hu mour due to the large part that humour plays in judging the winner of any round. Humour is a complex and adapting field, making it difficult to create an agent that can reflect the nuanced types found in humans. Nevertheless, nine broad categories determined through humour style questionnaires (Heintz & Ruch, 2019) offered a starting point for judging archetypes that can be subsequently narrowed down based on feasibility. The primary identified categories were fun/affiliative (often relying on shared experience), benevolent humour/self enhancing (laughing at self/life situations in a good-natured manner), sarcasm/aggressive (ridicule or mocking), nonsense (absurd or surreal), wit, irony, satire, cynicism, and self defeating. Creation of datasets for humour focused training would likely feature considerable overlap between these styles, though the work performed regarding the specific humour types of wit using wordplay (Palma Preciado, Sidorov, & Preciado, 2022) and irony (Potamias, Siolas, & Stafylopatis, 2020) provide insights and a starting point in the creation of specific archetypes for use in judge personality detection. The next logical step after creating our archetypes was to create a method to determine which archetype a judge most likely falls under. Naive Bayes classifiers offered one potential approach, having seen success in personality classification from text (Pratama & Sarno, 2015) and offering a simple but powerful classifier (Berrar, 2019).

Methodology

Data Preparation

To ensure high-quality training data, we developed a custom annotation program (Figures 1 and 2) for manual training of card combinations. The program was designed to be user friendly, featuring progress saving and resuming capabilities. Although initially intended for use by additional testers, time constraints limited its deployment. Moreover, to maintain consistency across different platforms during training and testing, we created an environment testing tool (Figure 3). This tool ensures compatibility by checking operating system details, Python versions, software dependencies, available resources, and hardware acceleration options. Its intuitive interface allows users to quickly identify any system in compatibilities, enabling smoother execution of training and testing processes.

Figure 1: *Screenshot of the custom annotation program for manual card labeling. The interface is designed for ease of use, supporting progress saving and resuming capabilities.*

Figure 2: *Screenshot of the custom annotation program for manual card labeling.*

Pre-trained Models

For this study, we utilized pre-trained BERT models available via the Hugging Face library. These models were selected for their state-of-the-art performance in natural language understanding tasks.

Fine-tuning Process

To adapt the pre-trained models to the "Apples to Apples" gameplay, fine-tuning was performed on a custom dataset of 800+ labeled pairings could be judged; second, to evaluate whether a BERT model would be capable of reflecting a more niche and difficult to define personality in the form of irony or puns/wordplay combinations, which were manually annotated by our team. This dataset captured the nuances of four distinct judge personalities (semantic association, sentiment analysis, irony detection, and pun detection).

Figure 3: *Screenshot of environment checker.*

Dataset Similarity and Isolation

While the training dataset overlapped thematically with the test set (e.g., similar topics and card categories), we strictly held out the test data during training to ensure no data leakage occurred.

Pre-trained vs Fine-tuned

Pre-trained BERT models provided general language understanding capabilities, while fine-tuning adjusted the models to prioritize personality-specific card selection tasks, enabling them to better simulate judge behavior

Selection of Personality Subtypes

The general ‘archetypal’ personalities chosen for implementation were Semantic Association, Sentimental, and Humorous. A judge with the Semantic Association personality would prefer red cards that had a close conceptual or contextual relationship with their given green card, creating a coherent, logical pairing. A judge with a Sentimental personality favours combinations that elicit a strong positive or negative emotional sentiment. The Humorous archetypal personality would look for any combinations that were considered generally humourous. These three personality types serve as the foundation for the typical range of personality traits one would expect to encounter in “Apples to Apples.” We divided the humour archetype into two additional niche types, in the form of ‘Irony’ and ‘Puns/Wordplay.’ The motivation for this sub division was twofold: first, due to the subjective nature of different types of humour, to create more specific categories in which card

Choice of Base Models and Datasets

Four pre-trained models were selected to represent the judge personality types, each trained on datasets relevant to its respective personality. These models serve as the ‘base’ for the Semantic Association, Senti mental, Puns/Wordplay, and Ironic personal ities. The models utilized for each personality type are ”all-mpnet-base-v2” (Reimers & Gurevych, 2019), ”sentiment-roberta-large english” (Hartmann, Heitmann, Siebert, & Schamp, 2023), ”roberta-small-pun-detector v2”, and ”twitter-roberta-base-irony” (Barbieri,

Camacho-Collados, Espinosa Anke, & Neves, 2020), respectively. All models were sourced from Hugging Face, a well-established machine learning community which offers a wide range of models, datasets, and applications. These four specific models were chosen because they offer a general yet comprehensive foundation for analyzing text in relation to each personality type and the ability to provide relevant scores for each. An additional dataset for irony was used from gimmaru/tweet eval irony to further train our irony model on pre labelled data. As all data provided so far to the models was based on more general pieces of text, we created a program to generate randomly paired card combinations to manually rank each based on our interpretations of the four personality types. This served not only to train each model on approximately 800 points of data in the specific context of “Apples to Apples" but also to provide subjective data from human players, which was viewed as especially important for detecting puns and irony.

Naive Bayes Approach

We attempted to incorporate a Naive Bayes classifier to enhance the ability to classify judge personalities based on their exhibited card selections. Naive Bayes was chosen due to its simplicity and efficiency, granting our model the ability to analyze multiple features simultaneously while attempting to determine a specific judge’s personality type. The features provided to the classifier were the individual ratings generated by each personality model for a given card combination, aggregated into a single feature vector consisting of five ratings: Semantic Association, Positive Sentiment, Negative Sentiment, Pun, and Irony. The feature vector, after being created by a certain personality model, was labelled with the name of that personality. These labelled feature vectors were used to train a Naive Bayes classifier, which would then be used to predict the most likely personality type of future card combinations. The feature vectors of the winning red and green cards chosen by the judge were aggregated for every round where that judge appeared. The goal was to have a changing overall feature vector that the agent would use to predict the most likely personality that a given judge possesses - reasoning that the card combinations from certain personalities would show patterns over time.

Card Combination Ranking and Selection

The approach to ranking and selecting the most suitable card combinations for each judge went as follows. Both red and green cards were tokenized using the tokenizer functions and embedding layers associated with each personality model. The obtained encodings were then processed by the model to produce logits, whose raw scores indicate how strongly the card combination aligns with a specific personality type. The model architecture necessitated treating each personality output as a multiclass problem rather than binary. Therefore, logits were then passed to a SoftMax function, which converted them into probabilities indicating how likely (or not) a given combination adhered to a particular personality type. Once the probabilities for each personality trait were determined, the card combinations were ranked based on the highest probability of matching the target personality type. The combination with the highest confidence score was chosen as the best match for that judge, as high confidence scores indicate the best possible alignment with the desired personality type in each hand. The Semantic Association model followed a slightly different process due to using SBERT architecture. Instead of using logits and the Soft Max function, the model computes the cosine similarity between the green and red card embeddings. This measures how closely related the two cards are in vector space, with higher similarity scores indicating a stronger semantic connection. The combination with the highest cosine similarity score is then selected.

Testing

Testing the “Apples to Apples” playing agent proved to be time intensive and difficult due

to inherent subjectivity in the game. The approach to testing, therefore, consisted of both human and automated testing to generate further data on the viability of the methods used. The first testing came from the datasets used for training. Splitting the datasets with training and testing allowed for quick and superficial analysis of the overall performance of the models on labelled data. Human testing consisted of multiple rounds of gameplay using a hand of seven red cards and a single green card. The tester then selected the red card from their hand that would best correspond to each of the personality models - association, positive sentiment, negative sentiment, irony, and pun. The same hands of red cards and green cards given to the human tester were then given to the agent, where each of the personalities would choose one card from the hand to match the green card. The amount of human/playing agent agreement on card selection was then tallied. Automated testing provided another challenge in that finding an agent that portrayed a certain personality type would be needed but not available. Therefore, automated testing did not allow for the evaluation of personality effectiveness. Instead, we focused testing on the overall viability of an agent playing “Apples to Apples” by examining situations where the agent should win. To this effect, we gathered additional testing models with similar personalities but pre-trained on different data sets to act as the judges each round. Each round then had eleven players total - the five agent personalities, five complements (not fine-tuned), and random. The winners of each round and the amount that each personality won when their complement was the judge was tallied. Finally, a control test was conducted to examine the overall effect that the randomness of hand assignment would have on the outcomes in “Apples to Apples.” There are often situations where there aren't any strong red cards to play for a given green card for the round judge personality. To examine how much this would impact an agent over time, similar automated testing was performed, but with the same personalities instead of complement models, i.e. there were two copies of each personality each round for a total of ten players, with the judge rotating through distinct personalities. Again, the winners of each round and the amount that each personality won when their duplicate was the judge were tallied.

Results

The result of the dataset testing (Table 1) gave the accuracies obtained by the senti ment, irony, and pun models.

Personality	Testing Accuracy
Association	-
Sentiment	0.84
Irony	0.69
Pun	0.92

Table 1: The association model was not able to undergo the same testing due to the continuous nature of the training label.

Human testing (Table 2) on 48 data points showed some differences in pre-trained and fine-tuned models. The pre-trained association model showed 23% accuracy in selecting the same card as a human tester, compared with 29% for the fine-tuned model. Positive sentiment showed relatively similar results at 21% and 23% for the pre-trained and fine-tuned models, respectively, while negative showed 25% and 35%. Irony (13%, 10%) and pun (17% for both) showed comparatively lower accuracy scores.

Personality	Amount Correct	
	Pre-trained	Fine-tuned
Association	11	14
Positive	10	11
Negative	12	17
Irony	6	5
Pun	8	8

Table 2: Human Expert Round Testing

Automated testing (Table 3) revealed that after 1000 rounds of testing, the complement judge accounted for 34% of the association model wins, 28% of the positive sentiment wins, 23% of the negative sentiment wins, 21% of the irony wins, and 15% of pun wins. Further, the association model won 12% overall, the positive

sentiment 11%, 10% for negative sentiment, 8.9% for irony, and 9.4% for pun. The control player, which played a random card every round, won 8.3% of the rounds. These total 594 rounds. The remaining 406 rounds were won by the complement models.

Personality	Rounds Won	
	Complement Judge	Total
Association	39	115
Positive	31	109
Negative	24	104
Irony	19	89
Pun	14	94
Random	-	83

Table 3: Automated Game Round Testing

Personality	Rounds Won	
	Complement Judge	Total
Association	25	72
Positive	27	62
Negative	26	61
Irony	12	61
Pun	9	53

Table 4: Automated Game Round Testing - Control

The control automated testing (Table 4) revealed that after 500 rounds of testing, the complement judge accounted for 35% of the association model wins, 44% of the positive sentiment wins, 43% of the negative sentiment wins, 20% of the irony wins, and 17% of pun wins. Further, the association model won 14% overall, the positive sentiment 12%, 12% for negative sentiment, 12% for irony, and 10% for pun of a total of 309 rounds. The remaining 191 rounds were won by the complement models.

Naive Bayes data showed an accuracy of about 25% for determining the personality based on a given feature vector of a ranked card combination.

Discussion

Our results show that while there is better than random selection in the use of BERT based models to play “Apples to Apples,” further refinement in methods for reflecting human personality and judgement is needed. Testing performed on items set aside from the team-made training dataset shows that there is relatively good prediction accuracy by the models for detecting the presence of their respective personalities in a red and green card combination. The effectiveness of the models dropped quite a bit when asked to choose between multiple options, however. Further, some models showed much higher capability in card selection than others. Random selection would result in an overall success rate of around 14%, which the fine-tuned association and sentiment models were able to defeat but which the irony and pun models were not. This may reflect the more subjective nature of irony and pun detection and may indicate the need for a more significant number of human testers to capture this nuance. One notable limitation of our work was the under-utilization of the annotation and environment testing tools. These tools were specifically designed to be distributed to external testers for broader data collection, but time constraints restricted their use to the internal team. In future work, refining these tools for wider deployment could facilitate larger-scale testing and validation efforts The automated testing shows that despite having similar personalities, the large amount of chance involved in “Apples to Apples" may still result in a win about 34% of the time. This is supported by the results of the Automated Game Round Testing Control group, where the highest contribution of ex act copies only accounted for a maximum of 44% of any model’s wins. This suggests that if a player is dealt a hand that does not have any good cards for a given judge’s personality, they may not be able to win the round despite knowing precisely what that personality is. The problem compounds as there is only one

new card drawn per round, and if a player cannot recycle their hand, they are effectively stuck. The result is a difficult situation for human players to overcome, reflecting the results of a playing agent attempting the same. The automated testing further revealed poor performance from the irony and pun players, even for the control. This likely stems from the relatively rare nature of a truly 'ironic' or 'punny' combination, resulting in an almost random selection from the two models. Further adaptability of a playing agent would be needed to find success moving forward - such as having alternative, broader personalities in the humour category to revert to should the more niche personalities (e.g. irony and pun) be unable to find a satisfactory match. Incorporating additional human testers would allow for greater nuance capture for niche personality types. An additional benefit would be to ask testers to perform the same task as the models and select the card that best matches each model. This would give a direct comparison between the human player and the playing agent based on criteria selected by a human judge. Finally, Naive Bayes showed limited success with our given implementation, likely due to similar confounding principles as above. The feature vector for any one card combination would have high variance, and only with aggregation from multiple rounds would a pattern emerge in the rankings seen from models. Our implementation trained on single card combinations, and therefore each label was associated with a high degree of variance in each feature. Future implementations would likely benefit from training the Naive Bayes classifier on aggregate data instead.

Conclusion

The creation and implementation of an "Apples to Apples" playing agent poses significant challenges, many of which are insurmountable due to the chance involved in the red cards provided to players. Different personality models showed varying success depending on how niche their selection criteria were, with broad selection categories showing better than random chance of correctly identifying the ideal card from a hand for a given judge. Future work in the creation of playing agents for word games would need a greater degree of focus on alternative selection criteria based on the current hand rather than an ideal personality match.

References

An, G., & Levitan, R. (2018). Lexical and acoustic deep learning model for personality recognition. In *Interspeech 2018* (pp. 1761–1765). doi: 10.21437/Interspeech.2018-2263

A. Rodriguez, M., & Merlo, P. (2020, November). Word associations and the distance properties of context-aware word em beddings. In R. Fernandez & T. Linzen ´ (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 376–385). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.conll-1.30/ doi: 10.18653/v1/2020.conll-1.30

Arora, S., May, A., Zhang, J., & Re, C. (2020, July). Contextual embeddings: When are they worth it? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2650–2663). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.236/ doi: 10.18653/v1/2020.acl-main.236

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020, November). TweetEval: Unified bench mark and comparative evaluation for tweet classification. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1644–1650). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.148/ doi:

10.18653/v1/2020.findings-emnlp.148

Berrar, D. (2019). Bayes' theorem and naive bayes classifier. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schonbach ¨ (Eds.), *Encyclopedia of bioinformatics and computational biology* (p. 403-412). Oxford: Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780128096332204731 doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1

Delmonte, R., & Busetto, N. (2022, June). Measuring similarity by linguistic features rather than frequency. In H. Bunt (Ed.), *Proceedings of the 18th joint acl - iso workshop on interoperable semantic annotation within lrec2022* (pp. 42–52). Marseille, France: European Language Resources Association. Re trieved from https://aclanthology.org/2022.isa-1.6/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171– 4186). Minneapolis, Minnesota: Association for Computational Linguistics. Re trieved from https://aclanthology.org/N19-1423/ doi: 10.18653/v1/N19-1423

Garrido-Merchan, E. C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. (2023, Apr.). Comparing bert against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, *2*(4), 352–356. Retrieved from https://ojs.bonviewpress.com/index.php/JCCE/article/view/838 doi: 10.47852/bonviewJCCE3202838

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Market ing*, *40*(1), 75-87. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167811622000477 doi: https://doi.org/10.1016/j.ijresmar.2022.05.005

Heintz, S., & Ruch, W. (2019). From four to nine styles: An update on individual differences in humor. *Personality and Individual Differences*, *141*, 7-12. Retrieved from https://www.sciencedirect.com/science/article/pii/S0191886918306421 doi: https://doi.org/10.1016/j.paid.2018.12.008

Isaak, N. (2022, 06). *A zero-shot classification approach for a word-guessing challenge.* doi: 10.48550/arXiv.2206.13099

Koroteev, M. (2021, 03). *Bert: A review of applications in natural language processing and understanding.* doi: 10.48550/ arXiv.2103.11943

Koyyalagunta, D., Sun, A., Draelos, R. L., & Rudin, C. (2021, September). Playing codenames with language graphs and word embeddings. *J. Artif. Int. Res.*, *71*, 319–346. Retrieved from https://doi.org/10.1613/jair.1.12665 doi: 10.1613/jair.1.12665

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111-132. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666651022000146 doi: https://doi.org/10.1016/j.aiopen.2022.10.001

Lucky, H., Zain Nabiilah, G., Hendrik Jeremy, N., & Suhartono, D. (2023, Feb.). A three-order ensemble model for user level big five personality prediction on twitter dataset. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2), 283–292. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/

view/2630

Mao, J., & Liu, W. (2019). A bert based approach for automatic humor detection and scoring. In *Iber lef@sepln.* Retrieved from https://api.semanticscholar.org/CorpusID:199448318

Palma-Preciado, V. M., Sidorov, G., & Preci ado, C. P. (2022). Assessing wordplay pun classification from JOKER dataset with pretrained BERT humorous models. In G. Faggioli, N. Ferro, A. Han bury, & M. Potthast (Eds.), *Proceedings of the working notes of CLEF 2022 - conference and labs of the evaluation forum, bologna, italy, september 5th - to - 8th, 2022* (Vol. 3180, pp. 1828–1833). CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-3180/paper-142.pdf

Patwardhan, N., Marrone, S., & Sansone, C. (2023, 04). Transformers in the real world: A survey on nlp applications. *Information*, *14*, 242. doi: 10.3390/ info14040242

Potamias, R. A., Siolas, G., & Stafylopatis, A. . G. (2020, Dec 01). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309-17320. Retrieved from https://doi.org/10.1007/s00521-020-05102-3 doi: 10.1007/s00521-020-05102-3

Pratama, B. Y., & Sarno, R. (2015). Personality classification based on twitter text using naive bayes, knn and svm. In *2015 international conference on data and software engineering (icodse)* (p. 170-174). doi: 10.1109/ICODSE.2015.7436992

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1410/ doi: 10.18653/v1/D19-1410

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). Bert: A review of applications in sentiment analysis. *HighTech and Innovation Journal*. Retrieved from https://api.semanticscholar.org/CorpusID:264954476

Shumanov, M., Cooper, H., & Ewing, M. (2022, Jan). Using ai predicted personality to enhance advertising effectiveness. *European Journal of Marketing*, *56*(6), 1590–1609. Retrieved from https://doi.org/10.1108/EJM1220190941 doi: https://doi.org/10.1108/EJM1220190941

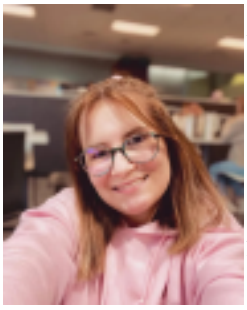
Vaswani, A., Shazeer, N., Parmar, N., Uszkor eit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st inter national conference on neural information processing systems* (p. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020, Apr.). Semantics-aware bert for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9628-9635. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/6510

doi: 10.1609/aaai.v34i05.6510

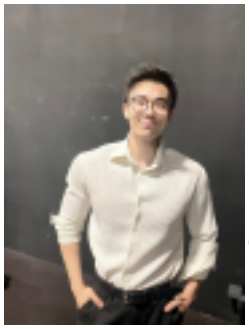
Zhou, Y., & Srikumar, V. (2022, May). A closer look at how fine-tuning changes BERT. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1046– 1061). Dublin, Ireland: Association for Computational Linguistics. Re

trieved from https://aclanthology.org/2022.acl-long.75/ doi: 10.18653/v1/2022.acl-long.75



Arikka Cherniwchan is an undergraduate student at MacEwan University, majoring in Computer Science. Her interests include AI, Machine Learning, Video Game Design and Development, and Software Design and Development.

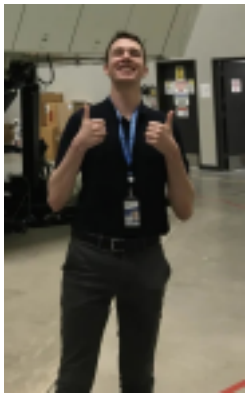
Austin Countaway is an undergraduate Computer Science student at MacEwan University. His interests include natural language processing and emergent behaviour in multi-agent systems using artificial intelligence, as well as video game development and script writing.



Chunyang Ding is an undergraduate Computer Science student at

e.

MacEwan University, with a minor in Statistics. He is interested in AI and data-driven methods, and was recognized at DataFest with the Best Use of External Resources award.



Brandon Funk is a recent graduate of MacEwan University. He is currently working as a developer at an Edmonton startup, where he's part of a project that brings together his interests in aviation and software development.

Calin Anton holds an MSc from the University of Bucharest, Romania and a Ph.D. from the University of Alberta, Canada, both in Computer Science. He has been teaching in different roles at postsecondary institutions for more than 25 years. For the last 15 years, he has taught at MacEwan University in Edmonton, Canada, where he is currently an associate professor of Computer Science. His current interests reside in Computer Security and Artificial Intelligence.