Predicting Creativity with a Neural Network Approach

Xubo Cao
Stanford Graduate School of Business
xcao@stanford.edu

Zi Ying (Kathy) Fan
Stanford University
zyfan@stanford.edu

Abstract

Creative ideas can bring great success to individuals and organizations, while investing in creative ideas is often risky because it requires deviation from the normative. Individuals cannot always accurately predict the creativity of ideas, and the approaches taken by researchers in the psychology field are limited in their capabilities to process longer textual ideas. Our goal is to bring recent advances in natural language processing to the creativity field, by applying transformer-based models to creativity detection. We find that a fine-tuned DistilBERT model with a long short-term memory classifier can achieve a correlation score of 0.777 against human ratings. Adding hand-built features that were hypothesized to be useful in distinguishing creativity did not boost the model's performance, but the results are inconclusive because of the small number of features examined. A more systematic investigation of hand-built features is warranted.

1 Introduction

Can a neural network approach be used to predict the level of creativity in a textual idea?

Creative ideas can bring great success to individuals and organizations, while investing in creative ideas is often risky because it requires deviation from the normative. Therefore, the evaluation and selection of creative ideas play a critical role in creative performance and innovation. For example, book writers need to decide which story to write about, movie producers need to predict which screenplay is likely to be a hit, and managers need to evaluate the business value of employees' proposals.

Past research has shown that individuals cannot always accurately predict the creativity of ideas (Berg, 2016, 2019). Perhaps artificial intelligence that is capable of identifying creative ideas can assist people in this type of decision-making. The current non-neural approaches taken by the psychology field do not seem generalizable to longer texts, rendering them obsolete in applications such as the ones mentioned in the previous paragraph. Therefore, in this research, we explore the possibility of using transformer-based advancements from the natural language processing (NLP) field to assess creativity of longer textual ideas.

Creativity is an elusive construct and can be operationalized in various ways. In this section, we specify how creativity will be defined and assessed in our study. Despite earlier efforts to develop objective measures of creativity, most creativity researchers acknowledge its subjective nature and define creativity based on products that are perceived to be both novel and useful by appropriate observers (Amabile, 1982). The difference between novelty and usefulness will not be emphasized in our review, but our approach is influenced heavily by the implication that the creativity of a product (e.g., an idea or a poem) is largely determined by *observers' perception*. According to the consensual assessment technique (CAT), which has become the "gold standard" in creativity assessment, a group of qualified judges' average ratings can be used as the measure of creativity if a satisfactory level of agreement is reached among them (Amabile, 1982). Following this logic, our task is to develop a model that can predict the perceived creativity of textual ideas in a way that sufficiently agrees with scores produced by human raters.

We present a baseline that attempts to apply the semantic-distance-based approach from the psychology literature (Dumas et al., 2020), transformer-based models with either linear or recurrent neural network (RNN) classifiers, and models that combine hand-built features with the transformer-based classifier. We validate the claim that although the semantic distance based approach has achieved great success in context of the Alterna-

tive Use Tasks (AUT), it will perform poorly in our task, which consists of texts of longer length. We find that the DistilBERTRNN classifier performs the best and that the incorporation of hand-built features did not significantly improve the model. Despite that, we note in the future works section that the investigation of hand-built features is limited in this study, and more comprehensive exploration is warranted.

2 Prior literature

2.1 Non-neural network approaches to creativity prediction

In the creativity literature, scholars often take a theory-driven approach in which they use hand-built features (e.g., semantic distances, prototypicality of edge distributions) to evaluate creativity. The linguistic models involved in these studies are not the most up-to-date but relatively more interpretable than neural-network based models (e.g., word embedding models, hand-built word networks). This trend may be attributed to the emphasis on interpretability rather than predictive accuracy in psychology research. Across all studies reviewed in this space, semantic distances (measured by cosine similarity or the Jaccard Index) are central to the researchers' system design. According to the idea that creativity is based on remote associations, it has been hypothesized that the association of two semantically distant words contribute to perceived novelty (Heinen and Johnson, 2018). Therefore, the semantic distances between words may be positively correlated with the perceived creativity of the idea.

(Dumas et al., 2020) used semantic distance in vector-space models (VSM) to measure originality in the AUT and examined the psychometric properties of this measure against human ratings. The AUT is a commonly used test for creativity assessment, in which participants are asked to come up with as many alternative uses of a certain object (i.e., brick) as possible. This test is designed to measure the divergent thinking ability, an essential component of creative thinking, of individuals. Note that the AUT is a measure of a person's creative potential (similar to an intelligence test), which is different from the task of interest in our paper (i.e., assessing the perceived creativity of an idea). That being said, the AUT involves scoring creative products and therefore is closely related to our research question.

The authors calculated the semantic distance between the prompt (e.g., "brick") and the response (e.g., "it can be used as a weapon") as a measure of the originality of the response. Specifically, the authors extracted word vectors for the prompt and the response from a VSM. When the text included more than one word, the word vectors were summed up with an idf weighting. The cosine similarity between the prompt and response vectors was then calculated as the originality measure. The authors employed four different pretrained VSMs in the study, including two latent semantic analysis (LSA)-based models (TASA and EN100k LSA), word2vec, and GloVe 840B. They then compared the scores produced by each VSM, with respect to their internal reliability and correlation with human-rated originality. Based on their results, (Dumas et al., 2020) recommended using the GloVe-based system for creativity assessment because it reached the strongest agreement with human raters and had high internal consistency.

(Beaty and Johnson, 2021) took a similar semantic approach to measuring originality scores in the AUT but extended the method of (Dumas et al., 2020) in two ways. First, it utilized five different semantic spaces in its system, including three continuous bag-of-word models, the TASA model, and GloVe embeddings. These models were combined to generate a latent variable, which partially mitigates some biases presented in the individual models and leads to predicted scores that were better correlated with human ratings of creativity. In addition, this paper also explored combining word vectors via element-wise multiplication and found that it performed better than additive approaches in earlier works.

2.2 Neural network approaches to humor prediction

Due to the sparsity of relevant NLP literature in the creativity space, we also survey neural network approaches to what we consider a related task: humor prediction. Humor can be considered as a type of creative text, sharing characteristics such as being highly subjective and requiring world knowledge to understand or evaluate (Hossain et al., 2020). Indeed, social psychologists have argued that creativity and humor are deeply connected in that they both involve "appropriate violations of norms" and require balance between surprisingness and appropriateness (Lu et al., 2019).

In contrast to the creativity literature, we see that in papers concerning humor detection, recent works are stepping away from theory-driven features and moving towards neural approaches, especially BERT-based (Bidirectional Encoder Representations from Transformers) models. Whereas earlier works in the literature like (Yang et al., 2015) use humor theory to create heavily hand-engineered, domain-specific algorithms for humor detection, (Hossain et al., 2019, 2020; Weller and Seppi, 2019) provide evidence that pre-trained BERT-based models, with relatively little engineering and additional training, can perform better. Hossain et al. collect datasets of modified news headlines which are rated with a score between 0-3. They examine the dataset in the context of various theories which characterize humor in terms of linguistic features such as joke length, topic incongruity, semantic distance of words, and joke structure. They then provide a few baseline models that make use of these linguistic features and demonstrate that the models can distinguish funny vs non-funny samples from the dataset. Non-neural classifiers include logistic regression, random forest, and support vector machine; some of the feature sets include n-gram features and features based on GloVe embeddings. In addition, they test out a neural model, a single-layer bidirectional LSTM (long short-term memory) with GloVe embeddings, which outperforms the non-neural methods. Finally, they demonstrate that a fine-tuned BERT model outperforms all of these prior methods. (Weller and Seppi, 2019) similarly fine-tunes BERT on a dataset of jokes from Reddit's r/Jokes thread and report that it performs better than other convolutional neural network based models, as well as models using hand-build features and word2vec. These prior works give us good confidence that it is productive to start explorations of the creativity prediction problem using a transformer-based architecture.

Finally, surveying the existing humor datasets, (Annamoradnejad, 2020) found that the datasets were small and prone to artifacts that meant even simple feature-based models could succeed at humor detection. Therefore, they contribute a new dataset to the field and also train a larger model on this dataset. In contrast to (Hossain et al., 2020) and (Weller and Seppi, 2019), they did not use a BERT model; rather, they used BERT only to the extent of creating sentence embeddings for their tokenized data and fed these inputs to a neural net model consisting of hidden layers and a concatenation layer. The specific design of their model was motivated by humor theories that characterize the typical linguistic structure of a joke (ex setup and punchline). From these observations, they decided to construct parallel hidden layers which process each example both as a whole, as well as on a per-sentence basis. Those outputs are then concatenated to form the final output. They found that the model performed better than non-neural baselines such as decision tree, SVM, and Naive Bayes. It also achieves a higher F1 score than XLNet, despite the model having roughly a third of the number of parameters and hidden layers as XLNet. In this context, XLNet is meant to serve as a benchmark model that can be considered as even better than BERT with regards to some other text classification tasks. These results suggest that incorporating the observations from existing theories into the model design, rather than into hand-crafted features, could be a productive complementary technique to using neural network models.

3 Data

3.1 Data collection

A corpus of written ideas is needed for model training, validation, and testing. In this study, we reuse data collected for three experimental studies in an earlier research project on creativity forecasting (Berg, 2019). In each of the three studies, participants were asked to write one or two creative ideas that can be adapted to solve a real-life problem. In study 1, 300 MTurk participants each provided one idea for the design of a new piece of fitness equipment; in study 2, 294 students from a US university each provided two ideas for keeping people engaged while using self-driving cars, resulting in 588 ideas; in study 3, 315 MTurk participants each provided one idea for a new travel experience to offer to consumers.

The ideas were evaluated by judges composed of ordinary consumers and experts in related fields. Each fitness-equipment-related idea was rated by 30 consumers and 10 fitness experts recruited online, like personal trainers and athletic coaches; each self-driving-car-related idea was rated by 30 consumers and 6 students in Mechanical Engineering who had just completed a course on self-driving car technology; each travel-related idea was rated by 30 consumers and 10 relevant experts, like travel

agents and tour guides (Berg, 2019). All judges rated the ideas on their novelty, usefulness, and overall creativity on a 7-point scale. The judges' average ratings are used as the measure of creativity.

Because expert ratings and consumer ratings are highly correlated, in all our experiments, we aggregate the two types of ratings by taking the average score of all judges' ratings. In addition, since the novelty, usefulness, and overall creativity scores end up being highly correlated, we only test our models on predicting the overall creativity score. Finally, due to the limited sample size, we conduct experiments in a "domain-general" setting, where we combine and shuffle the examples across the three studies, resulting in a sample size of 1203. Across this dataset, we note that the overall creativity scores are relatively normally distributed, with a mean of 3.903 and standard deviation of 0.768. The "domain-general" setting has the additional benefit that the diversity of the topics across the studies can help prevent our model from overfitting to a particular topic, potentially resulting in models that can used across various domains.

We set aside 203 examples from this merged dataset as the test set. The remaining 1000 examples are split into training and validation sets during our 5-fold cross validation training procedure.

3.2 Text preprocessing

For the semantic distance baseline model, we use the nltk module to remove stop words and punctuation marks in the ideas in order to reduce noise before subsequent vectorization.

For transformer-based models, we preprocess the text by removing any extraneous symbols (ex '/' forward slashes) and using the DistilBERT tokenizer. We add the [CLS] token to the beginning of each sample and the [SEP] token to the end. Sequences with less than 512 tokens are padded with the [PAD] token to create example texts of uniform length; samples longer than 512 tokens (including [CLS] and [SEP]) were truncated.

4 Approach

Our approach begins with implementing a semantic distance baseline model, and we expect that this model will fail to predict creativity scores that correlate with the true scores. On the other hand, our core hypothesis is that a transformer-based model can predict perceived creativity with a relatively high accuracy. We fine-tune a DistilBERT model and train it with a simple linear classifier in order to validate this hypothesis. We then explore two variations that could potentially boost the performance of this model: using a RNN classifier, and adding hand-built features.

4.1 General reasoning

As mentioned earlier, although psychology researchers have succeeded to measure the originality of responses in the AUT using semantic distances between text vectors, we hypothesize that this success cannot be replicated in our setting. In the AUT, participants are asked to generate as many creative uses for an object (e.g., "fork") as possible within a certain amount of time. The responses are mostly single words or short phrases (e.g., "eat pasta", "conduct electricity"), which are relatively easy to vectorize. In our data however, participants' responses are much longer and more open-ended, with an average length above 100 words. These longer sequences are likely to include more noise than responses in the AUT. Meanwhile, representing these longer sequences with single vectors will inevitably lead to great information loss. For these reasons, a semantic distance method may not generalize well to longer texts. In contrast, transformer-based models have the capacity to process longer sequences. In addition, two features may contribute to their potential in creativity assessment. First, creativity may lie in some complex relationships between words that are beyond the linear addition of simple features. transformer-based models, built upon a neural structure, have the flexibility to capture non-linear relationships between predictors. Second, creativity is often highly contextual, such that an idea that is novel and useful in one domain may be ordinary or irrelevant in another domain. Transformer-based models like BERT are context-aware and are therefore suitable for creativity assessment. Indeed, previous studies have shown that transformer-based models outperform many other models in humor detection, including systems based on semantic distances and other hand-built features. Considering that humor and creativity are related psychological constructs, both of which are highly subjective and involve appropriate deviation from norms, we hypothesize that transformer-based models have similar potential in creativity detection as in humor detection (Lu et al., 2019).

Although we believe that a fine-tuned DistilBERT model is already promising in creativity assessment, we further hypothesize that combining BERT-based models with other systems may boost their performance. First, we test whether using a RNN classifier to process DistilBERT representations can lead to a better performance compared to a linear classifier. Recurrent layers adds extra flexibility to the system, which may help capture additional abstract relationships that are not learned by the transformer. Also, we test whether combining hand-built features (like semantic distance) from the psychology literature with BERT models can lead to better performance. If a DistilBERT model fails to learn the relationship between certain hand-built features and creativity through fine-tuning, manually adding this feature to the system can potentially increase its performance.

4.2 Feature masking

In order to understand how the BERT-based models make predictions, we run feature masking experiments to probe the models' behavior, in which we remove a certain type of linguistic cues from the text in the testing set and examine its influence on model performance. These follow-up studies will also allow us to make more informed decisions in selecting hand-built features. Specifically, we conduct two types of feature masking experiments: text scrambling and word switching. In the text-scrambling experiment, we shuffle the test data by randomizing the order of words within each idea. Consequently, the sentence structure was destroyed, while word frequencies and some other linguistic properties (e.g., length of the idea) were preserved. It has been documented that text scrambling may not influence BERT-based models' performance in prediction tasks (Ettinger, 2020; Rogers et al., 2020). We test whether this finding can be replicated in the context of creativity assessment with DistilBERT.

In another experiment, we test the extent to which the model relies on superficial linguistic properties like word lengths and word count in the prediction task. For example, there may be a natural artifact in the dataset where more creative ideas use longer, richer words, and are more developed; while less creative ones use shorter, simpler words, and include fewer words. The model might pick up on these patterns even though they are not critical components of creativity. In this experiment, we replace each word in the text with an irrelevant word of the same length. The replacement words (Table 1) were chosen arbitrarily and expected to be neutrally associated with perceived creativity. Words greater or equal to 11 characters in length were replaced with the same word.

Length	Replacement	Length	Replacement
1	a	7	ancient
2	an	8	accident
3	and	9	accidents
4	andy	10	accidental
5	antic	$\geq$11	accidentally
6	accent		

Table 1: Replacement words used per word length.

4.3 Metrics

For all our models, we use Pearson correlation coefficient r between the predicted scores and human ratings as the primary performance metric. Correlation coefficients are commonly used as the metric for regression tasks. We choose Pearson correlation over Spearman correlation here because the former is more standard in psychometrics and can be directly transformed into an R-square value, which quantifies the proportion of variance captured by the measure. Further, most previous studies in automatic creativity assessment reported Pearson r. Our results will be more comparable with these studies by using the same metric.

5 Models

5.1 Semantic distance model

Our non-neural baseline is a conceptual replication of (Dumas et al., 2020)'s semantic-distance-based system. In this model, we represent each idea and the corresponding topic ("fitness equipment", "self-driving car", and "travel") with two vectors and calculate the cosine similarity between the two vectors. We apply the 300-dimension GloVe 6B embedding to vectorize the input text. Words that are not present in the word embedding dictionary are discarded. Element-wise multiplication is then used to combine multiple word vectors into a single vector that represents the topic or the idea. Although (Dumas et al., 2020) used weighted addition to combine multiple vectors, a later study showed that the multiplicative composition approach works better (Beaty and Johnson, 2021). Finally, the cosine similarity between the topic vector and the

idea vector is used as the predicted creativity score of the idea.

5.2 DistilBERTLinear

To test our main hypothesis that a transformer-based neural model can better predict the creativity score, we fine-tune a PyTorch DistilBERT model on the dataset and train a linear classifier to make predictions. We choose DistilBERT, since it is a lightweight variant of BERT that achieves similar performance while offering better computational efficiency (Sanh et al., 2019). Specifically, the authors of DistilBERT state that it is 40% smaller (in terms of the number of model parameters) and 60% faster, while retaining 97% of BERT's language understanding capacities.

We initialize the model with 'distilbert-base-uncased' pre-trained weights using huggingface's transformer module (Wolf et al., 2020). We take the output embeddings above the [CLS] token as the input to the downstream classifier, since we found that this embedding led to better performance over other pooled representations. The linear classifier consists of a linear layer, a dropout layer to mitigate potential overfitting, and a final linear predictor. We train the model with dropout value of 0.2, batch size 8 and an Adam optimizer of 3e-05. These parameters were tuned via gridwise hyperparameter search, where we choose the model with the highest average validation Pearson correlation over 5-fold cross validation. Each fold was trained for a maximum of 10 epochs, with early stopping if the model's loss changes by less than 10% from the previous epoch or if the training correlation outperforms validation correlation by more than 0.3.

5.3 DistilBERTRNN

To test our hypothesis that an RNN classifier may improve the model's performance, we initialize the DistilBERT model as previously and build a classifier with one hidden layer of dimension 768, a linear layer, a dropout layer, and a final linear layer. The hidden layer uses bidirectional LSTM cells. The model was trained with the same procedure as above; type of cell (LSTM vs GRU), hidden dimension, bidirectionality, and the number of RNN layers were similarly chosen via hyperparameter search.

5.4 DistilBERTLinear-Feature models

We then test our alternative hypothesis that combining hand-built features with DistilBERT embeddings could aid the model. More concretely, we use the concatenation of some hand-picked features and BERT representations as the input of the final classifier layers and examine whether the addition of these features lead to better results.

Two types of hand-picked features are tested in our study. In the DistilBERTLinear-Semdis model, we concatenate a set of semantic-distance measures with BERT representations. Although, the results of our baseline model show that the semantic distances between the topic vector and a single idea vector do not correlate with perceived creativity (see the Results section), previous studies also used semantic distances between word pairs within the text to detect creativity or humor (Gray et al., 2019; Yang et al., 2015). To implement this approach, we calculate 12 different types of within-sequence semantic distances measures and incorporate all of them into our DistilBERTLinear model. For the first 3 measures, we use GloVe to extract the semantic distances between all content word pairs within a sequence and pool them using one of the three functions: [mean, max, min]. In other words, we calculate the average word distance, max word distance, and minimum word distance within a written idea. We then apply a two-step method to generate the other 9 features. We split the sequence into sentences using nltk's sentence tokenizer and extract all word distances within each sentence using GloVe. These word distances are first pooled at the sentence-level and then pooled again at the sequence level using the aforementioned functions. Compared to directly pooling word distances within the whole sequence, this two-step approach assumes that semantic distances could be locally meaningful. All of the 12 measures are significantly correlated with perceived creativity, justifying the use of these features. However, these correlations are also likely mediated by lengths of ideas. In other words, the inclusion of these features does not guarantee an increase in model performance. Therefore, we also test a DistilBERTLinear-Length model.

Our feature masking experiments show that the DistilBERT model already takes word count into consideration (see the Results section). Thus, it adds little value to directly incorporate word count into the baseline DistilBERT model. Instead, in

the DistilBERTLinear-Length model, we test an alternative approach: we increase the length of all ideas to over 512 tokens by repeating the content of the idea, which will then all be truncated to 512 tokens by the tokenizer. Consequently, this manipulation obscures the original length of the idea and forces DistilBERT to rely on signals other than word count to make predictions. This text padding is different from the default token padding in the DistilBERT tokenizers in that here, we fill the sequence with actual words that will not be filtered out by attention masks. We fine-tune a DistilBERT model with these padded texts in hopes that it is more sensitive to cues other than word count. We then concatenate the original word count with these BERT embeddings to examine whether the two-step approach can lead to better results.

6 Results and discussion

The Pearson correlation from the semantic distance model was .06 ($p > .05$), -.03 ($p > .05$), and -.03 ($p > .05$) for each sample respectively, meaning that the model was unable to predict creativity scores that correlated with the true human rater's scores. This validates our hypothesis that this baseline method cannot be directly generalized to predicting the creativity score of longer sequences.

Our DistilBERTLinear model achieved a test set correlation of 0.772. To contextualize how good of a performance this represents, we compare it to "human performance," which is the average performance of individual human raters. To obtain the human performance benchmark, we treat the rating of each human rater on each idea as a data point and calculate the Pearson correlation between these individual ratings and the averaged scores of the other raters. Since the human performance is 0.450 and our model performs at 0.772, we conclude the transformer-based model indeed is successful at performing the task.

The DistilBERTRNN model achieved a test set correlation of 0.777. Considering that the DistilBERTRNN model has approximately 17% more trainable parameters than the linear model, this is not a drastic improvement. Indeed, during hyperparameter search, we noticed that the variation in performance was quite low; more layers (deeper) or more hidden units did not necessarily help the performance beyond a certain point. Since the presence of recurrence seems to have minimal effects on the model's performance, we believe it is possible that in our model, the amount of information encoded by the DistilBERT embeddings may be a bottleneck for the downstream classifier's performance. It is also possible that due to our limited dataset size or the quality of our data, the maximum achievable correlation is near the 0.77 range and learnable by a linear classifier; therefore, a more powerful RNN classifier could demonstrate similar ceiling performance. Regarding the results of feature masking, text scrambling had essentially no influence on the model's performance, implying that the model is able to learn how to predict creativity scores without paying attention to syntactic information. This finding is consistent with the results observed in previous studies (Ettinger, 2020; Rogers et al., 2020). The model was also able to make reasonable predictions even after word switching, indicating that linguistic properties like word lengths and word counts play an non-trivial role in the model's prediction.

Our DistilBERTLinear-Semdis model achieved a test set correlation of 0.772, which is almost the same as the pure DistilBERTLinear model. This lack of meaningful improvements may indicate that our measure of semantic distances includes little information that has not been learned by a finetuned DistilBERT (e.g., the correlation between our semantic distance measure and perceived creativity is likely accounted for by word count).

Our DistilBERTLinear-Length model achieved a test set correlation of 0.765, which is slightly worse than the pure DistilBERTLinear model. Note that because padded texts include more real tokens and thus take more storage space, we were forced to use a batch size of 2 instead of 8 like in other models, which might explain the decrease in performance. Nevertheless, because the choice of batch size had relatively trivial effect in our other hyperparameter searches, we conclude that our DistilBERTLinear-Length model did not outperform a baseline DistilBERTLinear model. Surprisingly, even without the addition of the word count feature, DistilBERT can still evaluate the creativity of padded text with a Pearson r of 0.731, indicating that DistilBERT was not fully tricked by the padding manipulation and was able to filter out some repeated information.

We summarize the results from our models in Table 2.

7 Conclusion

Most of our hypotheses are supported by the results. We demonstrate that while semantic distance

Model	Correlation
Semantic distance	0.06, -0.03, -0.03
Human performance	0.450
DistilBERTLinear	0.772
DistilBERTLinear, scrambled	0.774
DistilBERTLinear, word replacement	0.571
DistilBERTRNN	**0.777**
DistilBERTLinear-Semdis	0.772
DistilBERTLinear-Length	0.765

Table 2: Summary of model performance on test set.

cannot be directly applied to predicting creativity in longer texts such as those of our dataset, transformer-based models like DistilBERT with either linear or RNN classifier can achieve correlation scores up to 0.777, greatly exceeding human performance. On the other hand, our hypothesis that adding hand-built features to a DistilBERT model can boost its performance was not supported. Our current results show that adding within-sequence semantic distances to DistilBERT models led to little improvement. Yet, these results should not be regarded as conclusive. The discrepancy between our results and previous psychology research speaks to the difficulty of generalizing the method to longer free text, which is nevertheless still possible and worth exploring.

In addition, our feature masking experiments provide some primary evidence for how DistilBERT models performed the prediction tasks. Consistent with previous research, the text scrambling experiment shows that DistilBERT models do not rely on the order of words to make creativity predictions. The word switching experiment shows that some length related features like word counts and word lengths are important predictors used by DistilBERT. Finally, the fact that DistilBERT was able to make highly accurate predictions based on padded text may imply that the model is able to filter out repetitive information and capture the true length of the idea.

8 Future work

Future work that directly extends this study could perform experiments in a "domain-specific" setting where the three dataset samples are not combined, in order to further study the model's ability to generalize across different creative topics and its mechanisms for identifying creative content within each topic. Although we deemed the novelty and usefulness scores to be highly correlated to the overall creativity metric, future work could also explore predictions on the novelty and usefulness scores independently. Finally, we have only searched for a small set of hand-built features in this study. Although our attempts to add hand-built features based on semantic distance and idea length did not result in higher correlation scores, it is still possible that other creativity-related features can be used to improve the model. Future work can benefit from taking a more rigorous approach to feature building and explore the effect of hand-built features in a more comprehensive manner.

Another line of work could lie in making additional enhancements to the model. For example, (Hossain et al., 2019) points out that one characteristic of humor in their dataset is the presence of a coherent but surprising ending to a phrase. We think that the emotion of surprise could also be a trait of creative ideas, and it may be possible to combine the current model with methods from NLP works that classify emotions (ex anger, surprise, and happiness). Another possible idea is to use the findings of (Annamoradnejad, 2020) to influence how we present each training example to the model. Just like how a sentence in a joke can be non-humorous in of itself and requires the context of other sentences to create humor, single phrases from a creative text may seem ordinary unless understood in the context of the idea as a whole.

Lastly, a direction of future work could be to gather a dataset and establish a benchmark metric or set of tasks for evaluating creativity machine learning models in a more robust manner. Currently, the tasks, datasets, and evaluation benchmarks tend to borrow heavily from the psychology side of creativity and be limited in applicability or comparability, compared to standard benchmarks in NLP research. In the humor literature, papers cited in the prior works section of this report actively make contributions in the dataset collection area as well, and it would be fruitful to see similar advancements for creativity, in order to better facilitate the intersection of the field with NLP.

Acknowledgments

We would like to thank Dorottya Demszky for her mentorship during our project.

Authorship Statement

Xubo collected the dataset and performed the feature-masking, semantic-distance baseline, and hand-built features experiments. Kathy set up the code pipelines and performed experiments on the linear and RNN models. Both authors contributed to the background research, results analysis, and report-writing equally. Our code is available at https://github.com/kathyfan/cs224u-creativity.

References

Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013.

Issa Annamoradnejad. 2020. Colbert: Using BERT sentence embedding for humor detection. *CoRR*, abs/2004.12765.

Roger E. Beaty and Dan R. Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behav Res*, 53.

Justin M. Berg. 2016. Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Administrative Science Quarterly*, 61(3):433–468.

Justin M. Berg. 2019. When silver is gold: Forecasting the potential creativity of initial ideas. *Organizational Behavior and Human Decision Processes*, 154:96–117.

Denis Dumas, Peter Organisciak, and Michael Doherty. 2020. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S. Christian, John Patrick, Laura Huang, Yoed N. Kenett, and Kevin Lewis. 2019. "forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist*, 74.

David J.P. Heinen and Dan R. Johnson. 2018. Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, 12.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. *CoRR*, abs/1906.00274.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry A. Kautz. 2020. Stimulating creativity with funlines: A case study of humor generation in headlines. *CoRR*, abs/2002.02031.

Jackson G. Lu, Ashley E. Martin, Anastasia Usova, and Adam D. Galinsky. 2019. Creativity and Humor Across Cultures. In *Creativity and Humor*, pages 183–203. Elsevier.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Orion Weller and Kevin D. Seppi. 2019. Humor detection: A transformer gets the last laugh. *CoRR*, abs/1909.00252.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.