POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A Transformer-based Approach to Predicting the Likeability of Books

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Gianluca Guarro

Advisor: Prof. Marco Brambilla

Co-advisor: Marco di Giovanni

Academic year: 2021-2022

1. Introduction

Are our computers capable of extrapolating what elements make a novel or work of literature compelling to the human reader? Instinctively, one would assume that the "cold" binary foundation of computer logic and analytics is at odds with the appraisal of creativity and the interpretation and "understanding" of any form of human art. However, recent advancements in Neural Networks have demonstrated that even the visual, musical, and literary arts can be modeled to an impressive degree by computers, and that the latter can not only interpret human artistic products, but even generate their own works in diverse artistic domains. The recent breakthroughs in this area of research provide the scientific basis and motivation to seriously consider and investigate the aforementioned question, that is whether and to what extent computers of the current generation and processing capability can analyze books and differentiate between successful and unsuccessful literature.
Beyond the purely scientific challenge presented by the posed question, the development of a system that can effectively discern between good literature and bad literature in an automated fashion would have immense practical value to book publishers. It is no secret that the publishing industry operates with an unprecedented level of competition. The advent of electronic submissions in the digital age has resulted in an explosion of manuscripts that editors need to process. As a result, many of these submissions are not adequately assessed and evaluated.
While there is great motivation to study such a question, the underlying problem to be solved is quite complex and seemingly affected by undefined or unidentified factors. When considering in fact the traditional book evaluation processes, it can be readily noted that even many best-sellers have been rejected several times by publishing houses before finally being accepted and permitted to find their audiences. The qualities that make a novel great are not clear cut and may involve aspects that are difficult to assess in a predictive perspective of likeability, such as the book novelty, style of writing, story line, and character development. In addition, a book success may depend on external factors such as resources available for its marketing, current and transient social trends, and competition from books released simultaneously.
Despite its obvious difficulties, recent research has demonstrated that artificial intelligence is able, to an effective degree, to distinguish good literature from bad literature. The goal of this thesis is to expand the research of book success

prediction systems by investigating the applicability and effectiveness of the novel transformer architecture [11] to the problem. Since its inception, the transformer architecture has become commonplace among natural language processing (NLP) researchers thanks to models like BERT [2]. BERT harnesses the full potential of a transformer encoder as it comes out-of-the-box pretrained on millions of English sentences via the masked-language modeling task. In this way, BERT is endowed from the start with a strong understanding of the English language. On the other hand, despite the transformer architecture widespread success across several natural language processing tasks, its applicability to the book likeability prediction problem has not been seriously investigated. Our work seeks to take a step in the direction of filling this gap.

2. Previous Work

Before the boom of Artificial Intelligence, literary experts conducted several studies aimed at extrapolating stylistic aspects and/or content characteristics from great authors and books in order to qualitatively understand the elements of successful writing. In recent years, a handful of studies have been carried out that aim to build statistical models capable of predicting the successfulness of a book from its text alone.

Ashok et al. [1] conducted the first computational study correlating writing style and quality in literature. Their work focused on the construction of an array of handcrafted features that could then be used to train a Support Vector Machine (SVM).

Maharjan et al. are the largest contributors, having published three papers on the topic. Their first paper [4] expands upon the work of Ashok et al. with the construction of additional handcrafted features, as well as, neural ones. Among these neural features are Recurrent Neural Network (RNN) features which were developed keeping RNN shortcomings of dealing with long sequences of text in mind. Additionally, they discovered that the RNN model would perform better when trained in a multitask setting. Their third paper [6] introduces the state-of-the-art Genre-Aware Attention Model which allows the classifier to dynamically weigh the various modalities, while, at the same time, learn genre embeddings that get baked into the final representation of the book.

Khalifa et al. [3] attempted to do away with handcrafted features and instead build an end-to-end Convolutional Neural Network (CNN)-based classifier using pretrained sentence embeddings. While they were able to achieve good results using this method, they found that they could boost the weighted F1 score by nearly 5 points by simply incorporating readability metrics.

3. Task Dataset

An important characteristic of our target task is that it involves the classification of very long sequences. Generally speaking, neural networks designed to process sequential data, including RNNs, LSTMs, and transformers, are ill-suited for the processing of very long sequences. As a result, prior work on this problem has relied on both count-based textual features and/or some unorthodox and ad-hoc training procedure for allowing these neural networks to overcome to some degree this underlying issue. Given the nature of our problem, our work has proceeded along this same paradigm, in that one of its main contributions is an extensive investigation into how to make best use of transformer models when dealing with very long data sequences.

The dataset [4] used in this work is a benchmark dataset procured by Maharjan et al. The dataset consists of 1003 books across eight genres taken from Project Gutenberg. The average rating of the books on Goodreads was used to label the books as successful or unsuccessful. Specifically, if the book had an average rating of 3.5 or greater, it was considered successful, otherwise it was considered unsuccessful. In an effort to reduce noise or bias, certain heuristics were performed to appropriately choose the books, such as not allowing for books with too few ratings or books whose authors have already appeared twice in the dataset. The genre and success label distribution of the dataset can be seen in Table 1.

Genre	Unsuccessful	Successful	Total
Detective Mystery	60	46	106
Drama	29	70	99
Fiction	30	81	111
Historical Fiction	16	65	81
Love Stories	20	60	80
Poetry	23	158	181
Science Fiction	48	39	87
Short Stories	123	135	258
Total	349	654	1003

Table 1: Genre Distribution of Goodreads Maharjan Dataset

While working with this dataset, we had to take into account that the books typically began with a preamble containing information such as the copyright, translation or transcription note, an ASCII title page, etc. In an effort to concentrate our classifier training on the story itself, we studied methods to remove this information from our dataset. Noticing that the preamble often made frequent use of newline characters with respect to the rest of the text, we employed the CUSUM change detection algorithm [7] to identify precisely where the frequency of newline characters would change dramatically. According to our error analysis, this point correlated very well with where the book actually began. Thus, we used this method to trim the unwanted text from all the books in the dataset.

4. Our Work

In our work, we distinguish between first and second stage classifiers. The former process concerns the training of our BERT basic model and of the other BERT-like models that we have also investigated, whereas the latter concerns the training of models that make use of embeddings generated from the former. The motivation to study second stage classifiers comes as a direct result of working with very long sequences. Since transformer-based models typically have a low max sequence length limit, we needed to train our models by segmenting our books into multiple training samples. As a result, our BERT model does not attain a cohesive understanding of each book and instead averages its predictions on the segments to classify each book. The goal of the second stage classifier utilization was to compensate for this BERT limitation by allowing for a more holistic view of the books.

4.1. First Stage Classifiers

The development of the first stage classifier (i.e. transformer-based model) entailed many design decisions, that we investigated in an attempt to either better prepare the model for the target task, or to mitigate the potential issues arising from the segmenting / chunking of our dataset. These design decisions include a) the choice of the base transformer model; b) how to further pretrain the model (if at all); c) whether masking the characters in our dataset is useful; d) how to best tokenize our books into segments; e) whether to train the network in a multi-task setting with the genre; and f) if setting a max segment limit per book, or taking more fine-grained control of the sampling order to guarantee a more fair representation of shorter books, is useful. Decisions c, d, and f aim to target potential issues that surface from segmenting our samples into many parts.

Through Pointwise Mutual Information (PMI), we were able to verify that character names are the most discriminative words that distinguish the successful and unsuccessful classes in our dataset. Concerned about the risks of overfitting our model on character names, we passed our dataset through a named-entity recognition model to detect character names and subsequently mask them. Despite the results from PMI, character masking provided no significant benefit to the classification task, thus suggesting that BERT is robust to unique class identifiers.

We also explored two different tokenization algorithms. One that would ensure sentences did not get split between segments and one that would tokenize books with a moving window allowing for a defined amount of overlap. Moreover, we also studied the effect of truncating longer books to a defined max number of segments so that they are not overrepresented. Our experiments show that the two tokenization algorithms perform similarly on our dataset. Moreover, the insight BERT can extrapolate from each book quickly saturates. That is, we perceive no advantage in using more than 20 segments per book during our training.

Among all the BERT-inspired models, DistilBERT yielded the best performance on our

task. This comes as somewhat of a surprise, since the DistilBERT model was designed simply to be smaller and faster to train than other transformer-based models, rather than to seek better performance [9]. Moreover, also as a partial surprise, other models that either claimed to be a better version of BERT (RoBERTa) or advantageous when dealing with very long sequences (BigBird and Longformer) actually performed poorly on our task. These models were larger in size than BERT. We are led to deduce from these contrasting performance records, although we do not have a definitive proof for this conclusion, that such a difficult task as ours, for which such a small dataset as the one at our disposal is available, is actually better addressed by a smaller model with more built-in bias like DistilBERT.

Maharjan et. al report improved performance on the book likeability prediction task when training their models to simultaneously identify / predict the book genre. We have drawn inspiration from their results and modified BERT (DistilBERT) to this multitask setting. Despite not having come across other research that trains BERT to predict classes from two sets of labels simultaneously, our results show that our target task is benefitted by this multitask setup. We believe that simultaneously predicting book genre acts as a form of regularization for the likeability task.

While the investigated transformer-based models have been pretrained on millions of sentences of the English language, they have not necessarily been pretrained to understand literary English. We have therefore explored techniques that further pretrain these models on text within the same domain as our target dataset, making them more adept at handling our target task. In particular we have applied both the within-task and in-domain pretraining paradigms [10]. The former implies pretraining by using the same text as our classification dataset, whereas the latter requires an additional dataset whose text is obtained from a similar distribution as the text of the classification dataset. To explore the latter approach, we procured our own dataset of 2600 books from Project Gutenberg. While our results are not entirely conclusive, they suggest that further pretraining is at least slightly beneficial for the target task.

In comparison with the use of BERT in its simplest form (no pretraining, single-task), the utilization of the design decisions validated by our experiments in combination, to build the first stage classifier, allowed us to achieve a boost in performance from a weighted F1 (W-F1) score of 62.89% to a W-F1 of 72.15% on the test set. This result outperforms other architectures from other work designed to process sequential input as well as our own initial strong baselines.

4.2. Second Stage Classifiers

The classifier model developed so far makes predictions on segments of a book and aggregates them together to make a prediction for the whole book. We have attempted in this fashion to build and investigate a classifier that can consider a representation of the whole book in unison. As a building block, this study makes use of embeddings of the segments of our book extracted from our fine-tuned transformer model. Our research includes A) the training of hierarchical sequential networks, RoBERT and ToBERT , over our segment embeddings; B) the aggregation of the segment embeddings to create book embeddings and subsequently training a shallow feedforward network and an SVM over them; and C) the training of multimodal architectures that not only incorporate our BERT-based book embeddings but also utilize handcrafted features and other useful neural features. Our hierarchical sequential networks like RoBERT and ToBERT [8] train an LSTM layer and transformer encoder respectively over our segment embeddings. Despite the promising performances published in their original paper, these models underperformed our base DistilBERT model. On the otherhand, we were able to achieve our best W-F1 score of 73.63% by averaging the segment embeddings per book and using them to train a support vector machine (SVM). We were able to achieve a similar performance of 73.57% by training Maharjan's Genre-Aware Attention multimodal network using our DistilBERT book embeddings, sentiment concept features, and readability metrics. While the original paper for the Genre-Aware Attention Model [6] reports state-of-the-art results of 75.4% using their own combination of handcrafted and neural features, we found that model to be a difficult

and somewhat unpredictable network to train, with its performance being highly dependent on initial conditions and hyperparameters. Lastly, to allow for an easy comparison between our best performing model, our baselines, and the best performing models from prior works, we coalesce their results in Table 2. Model names in normal, italicized, or bolded text represent, respectively, our baselines, the best performing models of prior work, and our best performing model.

Models	**Test W-F1**
Most Frequent	0.506
Stratified	0.542
BERT One Randomized Chunk [3]	0.660
Bag of Words Logistic Regression	0.665
Tf-idf Logistic Regression	0.670
Word2Vec RNN	0.686
Emotion Flow Model [5]	0.690
Doc2Vec SVM	0.691
CNN with Readability [3]	0.720
Maharjan Multimodal With RNN [4]	0.735
SVM w/ BERT-based Features	0.736
Genre-Aware Attention [6]	0.754

Table 2: Model Comparison

5. Bibliography

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers-for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, 2017.
- Suraj Maharjan, Manuel Montes, Fabio A González, and Thamar Solorio. A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3381–3391, 2018.
- Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844, 2019.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

6. Conclusion

In conclusion, we found transformer-based models like BERT to be a suitable building block for the realization of a book likeability classifier, even more so when supplemented by the use of second stage classifiers which build off of BERT output, outperforming strong baselines. This result comes despite the fact that our task entails the classification of very long sequences, a domain for which transformer models are typically bottlenecked. In order to realize the true potential of transformer models on the task, we found worthwhile to explore several design decisions, such as a proper choice of the base model, further pretraining the model using a within-task approach, and modifying the transformer to work in a multitask setting. To our surprise, first-stage transformer models that advertise great progress in performance, such as RoBERTa, BigBird, and Longformer, failed in our task to live up to their claimed capabilities, while DistilBERT, whose principal focus was primarily to be just smaller and easier to train than BERT, was actually the most well suited in performance to the task among the base transformer models investigated. Moreover, we were able to show that fine-tuning BERT, so that it simultaneously predicts both book genre and likeability significantly boosts performance in the likeability task. We were also successful in coalescing the BERT embeddings of the book segments together by taking their mean, as evidenced by the enhanced performance of our SVM model. On the other hand, hierarchical sequential models underperformed with respect to our base model, in contrast with the results published in the associated original paper. While our results do not surpass the state-of-the-art, we have been able to achieve a high performance without using count-based features or a multi-

tude of neural features. This fact makes our network in some ways easier to apply, although at the expense of the considerable amount of computational effort needed for an adequate level of BERT training.

7. Acknowledgements

I would like to extend my gratitude to the several people who have been part of my journey during the completion of my master's thesis.
To my superadvisors, Professor Brambilla and Marco Di Giovanni: It has been an honor to work under your guidance and direction. The discussions from our weekly meetings inspired me to keep pushing my work further and further.
I must also thank my parents, Drs. Clorinda Donato and Sergio Guarro and my siblings, Marcello Guarro and Adriana Romero. Without their unconditional love and support, I would have never made it this far.
A special thanks goes to my relatives in Terni: Elisabetta, Vilma, Francesca, and Riccardo, who helped me stay sane and productive in my Polytechnic online studies when I escaped to the Umbrian countryside during the pandemic. I am forever grateful that allowed me to develop strong bonds with all four of you.
Last but not least, I would like to thank all the friends that I met along the way. Because of all of you, I will forever cherish the time I spent in Milan.

References

[1] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764, 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Muhammad Khalifa and Aminul Islam. Will your forthcoming book be successful? predicting book success with cnn and readability scores. *arXiv preprint arXiv:2007.11073*, 2020.

[4] Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, 2017.

[5] Suraj Maharjan, Sudipta Kar, Manuel Montes-y Gómez, Fabio A González, and Thamar Solorio. Letting emotions flow: Success prediction by modeling the flow of emotions in books. *arXiv preprint arXiv:1805.09746*, 2018.

[6] Suraj Maharjan, Manuel Montes, Fabio A González, and Thamar Solorio. A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3381–3391, 2018.

[7] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[8] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844, 2019.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[10] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.