# Text-to-hashtag Generation using Seq2seq Learning.

Augusto Cesar de Camargo
Wesley Seidel Carvalho
Department of Computer Science, Institute of Mathematics and Statistics
University of Sao Paulo, Brazil
Número USP: 11891023, augustoc@usp.br
Número USP: 6544342, wesley.seidel@gmail.com

## Abstract

In this paper, we studied if models based on BiLSTM and BER can generate hashtags that can be used in Ecommerce websites. We processed a corpus of Ecommerce reviews and titles of products as inputs and we generated hashtags as outputs. We evaluate the results using four quantitatives metrics: NIST, BLEU, METEOR and a crowdsourced score. Word Cloud was used as a qualitative metric. Besides all computer metered metrics (NIST, BLEU and METEOR) showed bad results, the crowdsourced showed amazing scores. We concluded that the texts generated by the neural networks are very promising to be used as hashtags of products in Ecommerce websites [1]. The code for this work is available on https://github.com/augustocamargo/text-to-hashtag

## 1. Introduction

Hashtags were created by Cri Messina in 2007 [2]. They are pervasive and very important on the Internet today [3]. Almost every service on the Internet lets you tag something with hashtags.

Our main motivation was to generate hashtags automatically to tag products in Ecommerce websites. Those hashtags will be read by humans and indexed by Search Engines. The last ones are our main audience. Search engines are a big source of traffic for Ecommerce websites and because of this they are so important. There is a whole field of study called Search Engine Optimization (SEO) [4]. Later we go deeper in this subject in the topic 1.2.

LSTM and BERT [5][6] are used in a wide variety of tasks in Natural Language Processing (NLP) today. We decided to use LSTM and BERT because they achieve success in many NLP tasks.

### 1.2 Search Engine Optimization (SEO)

All Ecommerce websites depend on new visitors everyday . A big source of traffic for Ecommerce is the Search Engine Result Pages (SERPs) [8]. The non paid traffic originated from the links on those pages are known as Organic Traffic (OT). OT has zero cost and is an important source of new visitors. The task of optimizing all that is the main goal of SEO.

Ecommerce owners fight to be on the top links in the results page of a user searching for a product. To have better OT results they need a lot of links listed in the results page and not only few links in the top.

It is important to notice that Google imposed a dictatorship of non duplicated content. Any attempt to recycle the same content is penalised by their search algorithm. So it is vital for Ecommerce to generate more and more new unpublished content. The new unpublished content generates more links in the results page of Google and so more OT. That is our motivation: generate new and more hashtags for the products.

### 1.3 Metrics

We evaluate quantitatively the results using four metrics: NIST, BLEU (Bilingual Evaluation Understudy) , METEOR (Metric for Evaluation of Translation with Explicit ORdering) [9] and a crowdsourced one.

The crowdsourced[10] score was named metricF and the question for the anotator was: “Can this sentence be used as a hashtag in an Ecommerce website?”. The scale goes from 0 to 1:

- 0: meaningless or totally grammatically incorrect;
- 0.5: good context and average comprehensible.

We used 6% of the text sets to be checked by the human annotator for both models (BiLSTM and BERT).

- 1.0: perfect context and completely comprehensible.

Word cloud was used as a qualitative metric to have an overview of the major words used in the texts and with it we can display text data in graphical form [11].

## 2. Related work

In the last 3 years we saw growth in research on hashtags recommendation [12] [13], [14] [15]. Our approach differs from other works mainly because we use a corpus of Ecommerce reviews and not content coming from social media.

There are a lot of examples in the internet of content about how to implement neural networks for text translation and summarization similar to what we are doing here [16][17][18].

## 3. Methodology

We used a dataset of Ecommerce reviews provided by B2W Digital. Our main tool for coding was JupyterLab [16]. The dataset was pre-processed before being used. The input of the LSTM and BERT were prepared specifically for each model. We describe both models details below.

### 3.1 Dataset

We want to thank B2W Digital for providing the dataset of this work. Thanks to their kindness this work could be done. The language of the content in this dataset is Brazilian portuguese. We use an open corpus of product reviews [17]. It contains more than 132.373 Ecommerce customer reviews, 112.993 different users regarding 48.001 unique products. The data was collected from the Americanas.com website between January and May, 2018.

The use two fields of the total of fourteen present in the file:

- *review_title*, text format, introduces or summarizes the review content;
- *review_text*, text format, it is the main text content of the review.

### 3.2 Pre-processing

All digits and special characters were removed. We also padded punctuation with white spaces between punctuation and words.

#### 3.2.1 BiLSTM

Two special tokens, ‘<start>’ and ‘<end>’, were added to review_title in order to help the decoder know from which point it should start decoding and where to end the decoding process.

A sample from the processed corpus is:

1. review_title: *<start> produto muito bom <end>*
2. review_review: *excelente qualidade , chegou dentro do prazo , recomendo .*

#### 3.2.2 BERT

We used an auto-regressive model for text generation using many iterations. Each iteration generates an output word that is incorporated into the input of the next iteration. We will generate the output sentence from left to the right, word by word. BERT is used in each iteration to generate the output word. We will do a classification task to refine BERT in order to fine-tune it to generate words. It is good to emphasize that this model was adapted to generate words, as it is usually used for classification tasks.

A sample of the processed corpus for BERT is:

ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] [MASK]
ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] melhor

ótimo filme , um dos melhores filme feito até hoje , o amor antigamente era tão profundo e sincero . [SEP] melhor[MASK]
ótimo filme , um dos melhores filme feito até hoje , o amor antigamente era tão profundo e sincero . [SEP] melhor filme

ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] melhor filme [MASK]
ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] melhor filme .

ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] melhor filme . [MASK]
ótimo filme , o amor antigamente era tão profundo e sincero . [SEP] melhor filme . [SEP]

## 3.3 Proposed Models

Several models were tested in preliminary experiments and we describe the two that led to the best results: one for LSTM and one fort BERT. All the details about the models can be seen here: https://github.com/augustocamargo/text-to-hashtag

### 3.3.1 BiLSTM

In Figure 1, we can see the details of the implemented BiLSTM model. Keras do not officially implement Attention so we used an outside implementation based on *'Bahdanau Attention'* [19] [20]. We follow the approach bellow to create the model:

- *We are using 'Teacher Forcing' technique for faster training of our model. In the teacher forcing method, we also pass the target data as the input to the decoder. For example, if we are going to predict 'hello', then we will pass 'hello' itself as an input to the decoder. This accelerates the training process* [19].
- *We use an inference model to predict our output sequences, using the weights from a pre-trained model. In other words, the model generalizes what it has learned during the training process to handle never-seen-before data* [19].

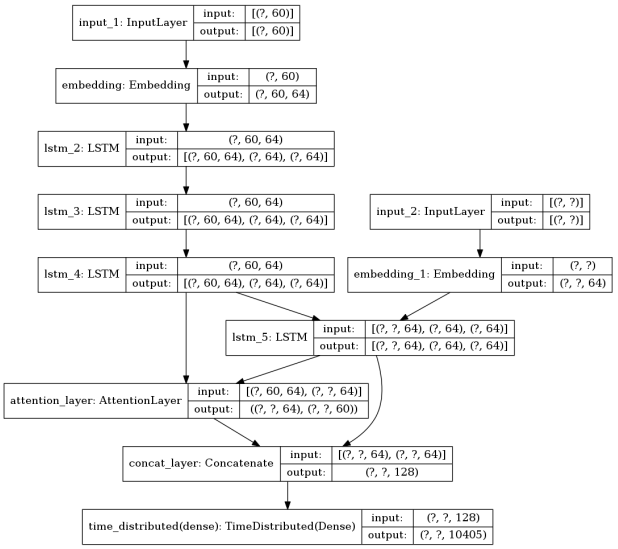


Figure 1: BiLSTM - model graph and summary.

### 3.3.2 BERT

In Figure 2 we can see the summary of our BERT implementation.

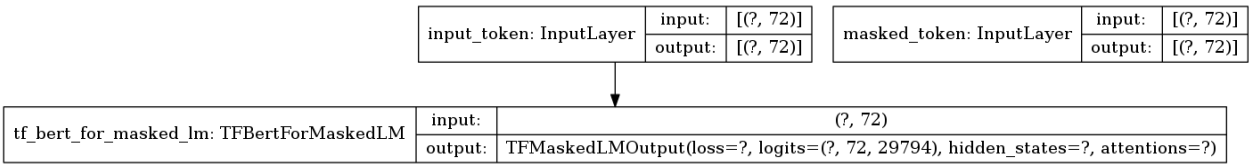


Figure 2: BERT - model graph and summary.

.

### 3.4 Experiments

We run 17 experiments for the BiLSTM and 25 for the BERTfor and present here the best experiment for each model. All the details about the experiments below can be found in here: https://github.com/augustocamargo/text-to-hashtag

In Table 1, all parameters used to execute the experiments for BiLSTM and BERT are presented.

Table 1: Setup of the experiments.

| | BiLSTM | BERT |
|---|---|---|
| Training set | 81,746 | 81,746 (306,974)* |
| Validation set | 17,517 | 17,517 (66,014)* |
| Test set | 17,517 | 17,517 (65,657)* |
| Epochs | 16 of 20** | 5 of 5 |
| Batch | 128 | 128 |
| GPU | NVIDIA Tesla K80 - 12 Gb RAM | NVIDIA Tesla V100S - 32 Gb RAM |
| Execution time | 60 min | 225 min |

* Size of the set after the pre-processing of the corpus for the BERT model.

The experiments were run using TensorFlow-GPU 2.3.1 and Keras 2.4.3.

## 4. Results and Discussion

Text samples generated by the models are shown in Tables 7 and 8.

*Generated* texts and *Predicted* texts are interchangeable terms. It happens the same to *Original* and *Review_title* fields in the corpus.

In Figure 3 and 4 we can see both models performances while being trained. In Table 2 we see the optimization results.

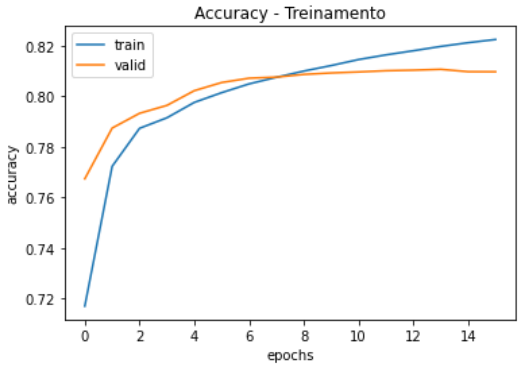


Figure 3: BiLSTM - training accuray.

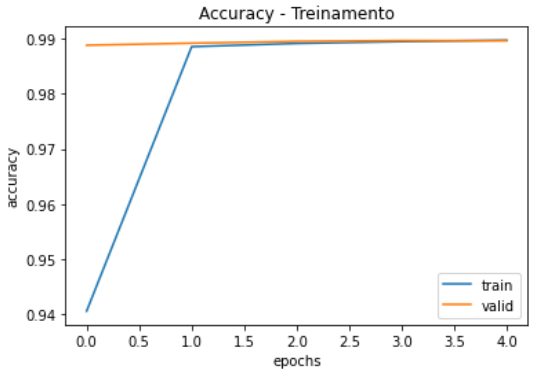


Figure 4: BERT - training accuracy.

Table 2: Results for training and test sets.

| | BiLSTM | BERT |
|---|---|---|
| Accuracy Validation | 0.809 | 0.990 |
| Accuracy Test | 0.801 | 0.990 |

As seen in Figure 3 and 4 both experiments do not overfit.

Below in Table 3 It is clear that the BERT model is much more creative (1,491) than the BiLSTM one (135), since it generates 10x more different sentences than the second.

Table 3: Quantity of unique sentences of Original vs Predicted texts.

| | BiLSTM - sentences (unique) | BERT - sentences (unique) |
|---|---|---|
| Original sentences | 6,994 | 7,947 |
| Predicted sentences | 135 | 1,491 |

In Table 4 we see that BERT is much more literate than BiLSTM since its vocabulary is almost 9x bigger than the second.

Table 4: Size of dictionary of Original vs Predicted texts.

| | BiLSTM - Words | BERT - Words |
|---|---|---|
| Original Dictionary | 3,100 | 4,244 |
| Predicted Dictionary | 102 | 907 |
| Difference | -2,998 / - 2,939.2% | - 3,337 / - 367,9% |

The BiLSTM model is more articulated than BERT because it generates sentences with 1-3 words while BERT only generates 1-2 words as seen in Table 5. We can see that the BiLSTM model has a greater variance in the generated titles size (51.77%) than BERT (29.43%). In Figure 5-7, we have the histogram of words/sentences. One more insight from Table 5: BiLSTM is smarter than BERT, as more complex sentences were formed from fewer words.

Table 5: Size of words/sentences of Original vs Predicted texts.

| | BiLSTM - Average, SDev and %CV (words/sentence) | BERT - Average Words - Average, SDev and %C (words/sentence) |
|---|---|---|
| Original text | 2.632 ± 1.647 %CV: 62.57% | 2.964 ± 1.866 %CV: 62.96 |
| Predicted text | 2.117 ± 1.096 %CV: 51.77 | 1.784 ± 0.525 %CV: 29.43 |

In Table 6-8 we can see generated text samples and their respective scores. NIST, BLEU and METEOR statistically have had the same behavior: really bad results. But metricF, our crowdsourced score, showed very good results: 0.810 for BiLSTM and 0.797 for BERT. It just demonstrates what we already know: those computed metrics can not capture semantics as well as crowdsource (humans) can.

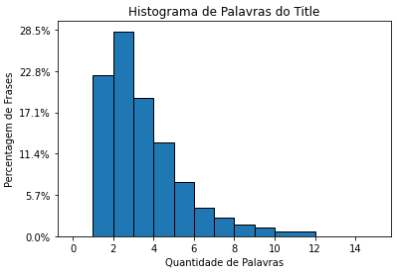


Figure 5: Histogram of original sentences.

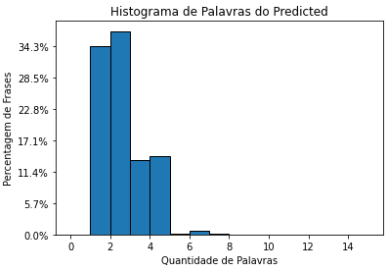


Figure 6: BiLSTM - histogram of predicted sentences.

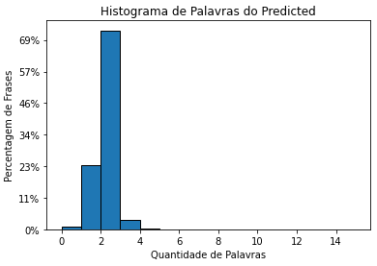


Figure 7: BERT - histogram of predicted sentences.

We also asked for the annotator of the metricF to analyze human generated text (Original) and give them a score. BiLSTM corpus scores 0.810 ± 0.349 and BERT corpus 0.912 ± 0.273.

Since computed metrics (NIST, BLEU and METEOR) are not in the same scale we applied the min-max normalization [21] on the results:

Table 6: Scores - Values.

| | Average of the Score | |
|---|---|---|
| | BiLSTM | BERT |
| NIST | 0.066 ± 0.164 | 0.058 ± 0.140 |
| BLEU | 0.046 ± 0.121 | 0.058 ± 0.119 |
| METEOR | 0.107 ± 0.218 | 0.091 ± 0.188 |
| metricF | 0.794 ± 0.349 | 0.706 ± 0.389 |

Table 7: Scores - %CV.

| | %CV of the Score | |
|---|---|---|
| | BiLSTM | BERT |
| NIST | 247.6% | 243.1% |
| BLEU | 263.9% | 205.5% |
| METEOR | 203.1% | 209.5%. |
| metricF | 43.8% | 55.1% |

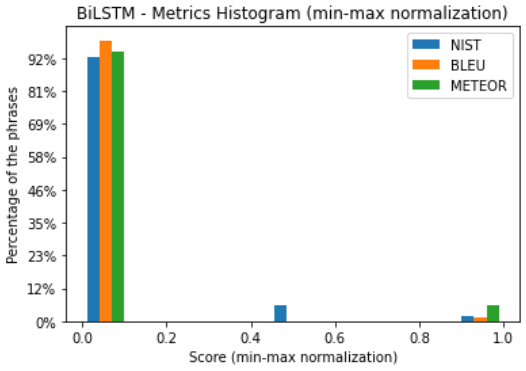


Figure 8: BiLSTM - scores histogram.

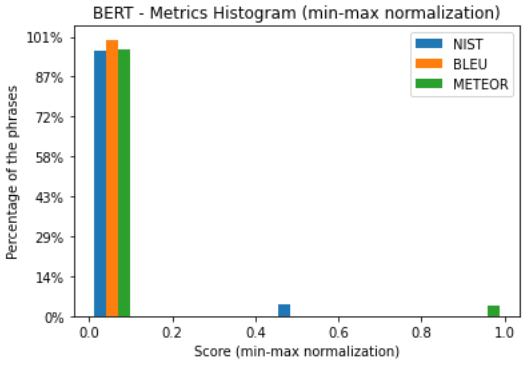


Figure 9: BERT - scores histogram.

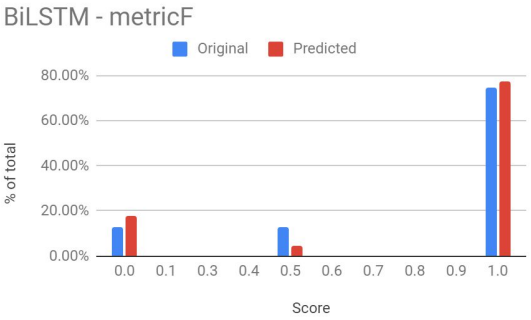


Figure 10: BiLSTM - metricF histogram

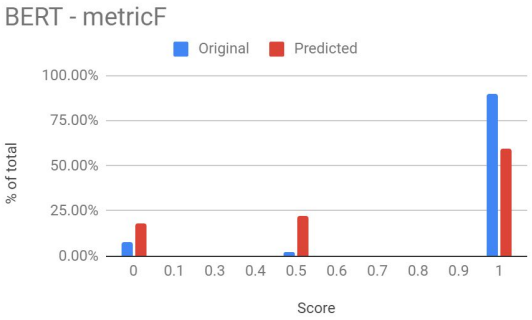


Figure 11: BERT - metricF histogram

Table 7: BiLSTM - sample of texts and metrics.

| NIST | BLEU | METEOR | metricF | Review | Original | Predicted |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1.0 | atendeu todas minhas expectativas qualidade ótima entrega ótimo antes do prazo nota para o produto e para americanas | americanas melhor loja | excelente |
| 0 | 0 | 0 | 1.0 | isso é um descaso com o cliente já tem um mês que comprei e nunca chegou já até quebrei meu cartão americanas nunca mais compro aqui | nunca mais compro nessa loja | nao recebi o produto |
| 0 | 0 | 0 | 0.5 | a parte da frente que tem o unicórnio veio faltando a peça simplesmente não estava dentro da embalagem como pode ser resolvido esse problema absurdo o presente de natal da minha filha | veio peça faltando | não gostei do produto |
| 2.000 | 1.000 | 0.992 | 0.5 | fui enganado não recebi o produto americanas não entregou reclamei mas não recebi paguei e não recebi | não recebi o produto | não recebi o produto |
| 0 | 0 | 0 | 0.5 | finalizei a compra e ainda nao chegou | a entrega demora muito | não é o que eu esperava |
| 0 | 0 | 0 | 0 | não funcionou no iphone no iphone plus nem no iphone x | não funcionou | gostei do produto |
| 0 | 0 | 0 | 0 | tem que ficar com o celular muito próximo ao adaptador se se um pouco mais de metro já começa a falhar assim é melhor usar um cabo p pra ouvir música | falha muito | bom custo benefício |
| 0 | 0 | 0 | 0 | simplesmente descartável use uma vez e o tecido se separa do elástico sinceramente não recomendo | material péssimo | gostei do produto |

Table 8: BERT - sample of texts and metrics.

| NIST | BLEU | METEOR | metricF | Review | Original | Predicted |
|---|---|---|---|---|---|---|
| 0.033 | 0.033 | 0.106 | 1.0 | comprei e não recebi o produto ! minha a avaliação vai para a americanas que não tem comprometimento com o cliente ! decepcionada ! | o produto não foi entregue | não recebi |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1.0 | jogo excelente , com gráficos ótimos e jogabilidade muito boa . recomendo a todos . | ótimo | excelente |
| 0 | 0 | 0 | 1.0 | ótimo custo benefício este aparelho atendeu perfeitamente minhas necessidades | aparelho bom | ótimo custo |
| 0 | 0 | 0 | 0.5 | o disco de fatiar não é bom para batata , fica muito fina e ao fritar gruda na panela mesmo com óleo bem quente . | produto excelente ! | maravilhos |
| 0 | 0 | 0 | 0.5 | cachos maravilhosos e duradouros , super recomendo ! | produto excelente ! | maravilhos |
| 0 | 0 | 0 | 0.5 | achei confortável , design grande , pelo numero ser  não ficou apertado , ajustou perfeitamente aos meus pés . | gostei | goste do |
| 0 | 0 | 0 | 0 | expeliarmus wingardium leviosaaa expectro patrono to na conta da minha mãe | gostei do exemplo | muito bom |
| 0 | 0 | 0 | 0 | produto funciona como o original da gree , além de esteticamente também ser igual ao original . | gostei ! | produto de |
| 0 | 0 | 0 | 0 | podia ter v . forno muito simples , bonito , mas só v em um forno a gás e lamentável . | forno bem simples | forno muito |

The word clouds in Figures 12-14 show us that words from generated sentences for both BiLSTM and BERT have very similar distributions. Although, they are a little bit different from the Review words. It may suggest that the Attention model really works.


Figure 12: Review_text - word cloud.

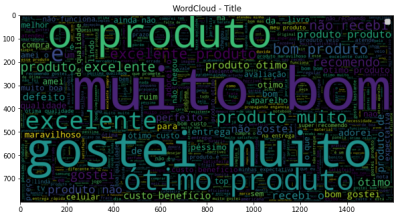


Figure 13: BiLSTM - predicted word cloud.

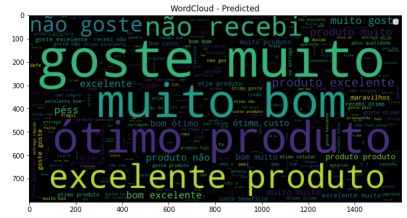


Figure 14: BERT - Predicated word cloud.

## 5. Conclusions

The implementation of the BiLSTM model was completely straightforward, but the BERT model was very counterintuitive and demanded a lot of tests, debugging and back and forth coding.

Our experimental results showed that LSTM and Transformers applied to text generation are useful to this type of task. Our metricF (crowdsourced) showed that the generated texts could really be used as hashtags in the Ecommerce websites: 0.794 for BiLSTM and 0.706 for BERT. All computured metrics (NIST, BLEU and METEOR) showed bad results and did not address the needs of this task.

The models used here (BiLSTM and BERT) are tied in almost all metrics when considering the standard deviation.
After so many experiments, we believe that the BiLSTM model is in the limit of its performance with this corpus.

Sentences generated by the models when used as hashtags can be confused with human ones. metricF scores for computer generated sentences and for human generated sentences are tied when considering standard deviation,

Generating sentences with BERT is not usual but is possible. Considering the potential of this model, we believe that more effort on fine-tuning it can have a huge payback.

Future works include different approaches and new architectures for generating sentences with BERT.

## References


[1] A. Belz and E. Reiter, "Comparing Automatic and Human Evaluation of NLG Systems," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006 [Online]. Available: https://www.aclweb.org/anthology/E06-1040

[2] "The hashtag at 10 years young." [Online]. Available: https://blog.twitter.com/en_us/topics/product/2017/the-hashtag-at-ten-years-young.html. [Accessed: 20-Dec-2020]

[3] S. Giannoulakis and N. Tsapatsoulis, "Evaluating the descriptive power of Instagram hashtags," *Journal of Innovation in Digital Ecosystems*, vol. 3, no. 2, pp. 114–129, Nov. 2016.

[4] "What is SEO?" [Online]. Available: https://moz.com/learn/seo/what-is-seo. [Accessed: 13-Dec-2020]

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv [cs.CL]*, 11-Oct-2018 [Online]. Available: http://arxiv.org/abs/1810.04805

[6] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv [cs.CL]*, 26-Jul-2019 [Online]. Available: http://arxiv.org/abs/1907.11692

[7] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv [cs.CL]*, 12-Jun-2017 [Online]. Available: http://arxiv.org/abs/1706.03762

[8] "Ecommerce SEO Guide - Drive More Organic Traffic To Online Store," 08-Sep-2020. [Online]. Available: https://newsdio.com/ecommerce-seo-guide/. [Accessed: 13-Dec-2020]

[9] K. Wołk and D. Koržinek, "Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking," *arXiv [cs.CL]*, 12-Jan-2016 [Online]. Available: http://arxiv.org/abs/1601.02789

[10] Wikipedia contributors, "Crowdsourcing," *Wikipedia, The Free Encyclopedia*, 27-Nov-2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Crowdsourcing&oldid=990983866. [Accessed: 21-Dec-2020]

[11] C. A. DePaolo and K. Wilkinson, "Get Your Head into the Clouds: Using Word Clouds for Analyzing Qualitative Assessment Data," *TechTrends*, vol. 58, no. 3, pp. 38–44, May 2014.

[12] D. Yang, R. Zhu, and Y. Li, "Self-Attentive Neural Network for Hashtag Recommendation," *Journal of Engineering Science and Technology Review*, vol. 12, no. 2), pp. 104–110, Apr. 2019.

[13] Y. Li, T. Liu, J. Hu, and J. Jiang, "Topical Co-Attention Networks for hashtag recommendation on microblogs," *Neurocomputing*, vol. 331, pp. 356–365, Feb. 2019.

[14] Y. Li, T. Liu, J. Jiang, and L. Zhang, "Hashtag Recommendation with Topical Attention-Based LSTM," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3019–3029.

[15] M. Kaviani and H. Rahmani, "EmHash: Hashtag Recommendation using Neural Network based on BERT Embedding," in *2020 6th International Conference on Web Research (ICWR)*, 2020, pp. 113–118.

[16] "Project Jupyter." [Online]. Available: https://jupyter.org/. [Accessed: 21-Dec-2020]

[17] *b2w-reviews01*. Github [Online]. Available: https://github.com/b2wdigital/b2w-reviews01. [Accessed: 12-Dec-2020]

[18] J. Wagner, R. S. Wilkens, M. Idiart, and A. Villavicencio, "The brWaC Corpus: A New Open Resource for Brazilian Portuguese," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018 [Online]. Available: https://www.researchgate.net/publication/326303825_The_brWaC_Corpus_A_New_Open_Resource_for_Brazilian_Portuguese. [Accessed: 19-Dec-2020]

[19] H. Patel, "Neural Machine Translation (NMT) with Attention Mechanism," *Towards Data Science*, 05-Jun-2020. [Online]. Available: https://towardsdatascience.com/neural-machine-translation-nmt-with-attention-mechanism-5e59b57bd2ac. [Accessed: 12-Dec-2020]

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv [cs.CL]*, 01-Sep-2014 [Online]. Available: http://arxiv.org/abs/1409.0473

[21] "Max Normalization." [Online]. Available: https://www.sciencedirect.com/topics/computer-science/max-normalization. [Accessed: 13-Dec-2020]