Workshop Proceedings of the 19th International Conference on Intelligent Environments (IE2023)
G. Bekaroo et al. (Eds.)

doi:10.3233/AISE230014

Finding Beauty Products Chemicals Using Token Identification and Entity Recognition Transformer Model

Sarthak Tyagi[1,a], Karmabir Chakraborty[2,a], Venkata Ashish Nunna[3,a], Abhishek Singh[4,a]
[a] *Vellore Institute of Technology, Chennai (Tamil Nadu, India)*

Abstract. Transformer architectures and pre-training have facilitated building higher-capacity models and effectively utilizing this capacity for a wide variety of tasks. Various libraries consist of carefully engineered state-of-the-art Transformer architectures under a unified API. But still, there exists no such system which detects the chemicals present in the Beauty products reviews on online e-commerce websites. On the market, there are thousands of different cosmetic products, all with different combinations of ingredients. Methods such as token identification and entity recognition transformer model advancing in both model architecture and model pre-training are implemented to identify and categorize the named entities.

Keywords. Named Entity Recognition, Topic Modelling, BERT, Natural Language Processing

1. Introduction

The objective is to implement a token identification and entity recognition transformer model to identify and categorize named entities of different chemicals which will be integrated with our blogging website to inform the user about the type of blog, category of the blog and about what person, place or any other entity is mentioned in a particular blog. Our main objective is to use natural language processing techniques driven by advances in transformer architecture. Transformer architectures have facilitated building higher-capacity models and pretraining has made it possible to effectively utilize this capacity for a wide variety of tasks. The majority of NLP progress has been made through adaptations of widespread architectures. The original models used in natural language were recurrent: they maintained some state which got fed into the next part of the model, along with new input, at each step. Encoder-decoder architectures were further developed to overcome the complications of RNN(Recurrent Neural Network) for accurate machine translations and. As new inputs are fed in, the encoder updates the state until the final input, at which the last hidden state is taken into a numerical representation. The decoder is fed this representation and uses it to generate the output sequence. The decoder "packs in" each input and "unpacks" one output word at a time in the final hidden state. Although this was certainly a step in the right direction, the *information bottleneck* caused by the use of only one hidden state was a problem; the decoder only has access to a very reduced representation of the sequence. This is particularly problematic for long texts, as remembering and representing information

[1]Corresponding Author: Sarthak Tyagi (sarthak.tyagi2019@vitstudent.ac.in)

from far back in the sequence can be lost in the compression to the final representation. As a result, practitioners began to give the decoder access to all of the encoder's hidden states. This is known as *attention.* However sequential computations, requiring inputs to be fed in one at a time, prevents parallelization across the input sequence and increase computation time, To solve the limitations of the attention mechanism, the *transformer* took another step towards a free-form attention model. To do this, it removed the recurrent network blocks and allowed attention to engage with all states in the same layer of the network. This is known as self-attention and is shown below Figure 1. Both blocks have self-attention mechanisms, allowing them to look at all states and feed them to a regular neural network block. This is much faster than the previous attention mechanism and is the foundation for much of modern NLP practice.

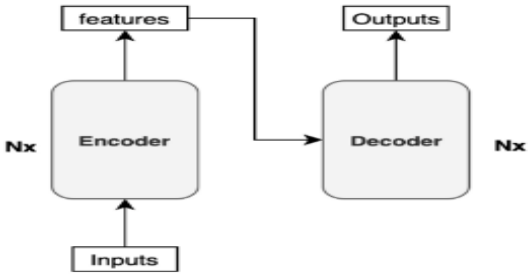

Figure 1. Encoder-decoder architecture of the original transformer

BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

We divided the work into 4 modules :

1.) **Front-end:** Created using typescript ReactJS and Material UI Toolkit with Axios which is used to communicate with the back-end REST API

2.) **Back-end** Built on NodeJS using ExpressJs with MongoDB as database and Mongoose as a DB helper

3.) **Data-mining** Used Selenium for data extraction.

4.) **Natural Language Processing**:

a.) Used NMF (Non-negative matrix factorization) for Topic Modelling

b.) BERT (Bidirectional Encoder Representations from Transformers) used for the Named Entity Recognition model for detecting chemicals/ ingredients in a cosmetic product

2. Literature Review

[1] In 2017, Vaswani et al. proposed a new neural network architecture named Transformer. This architecture was modern and it quickly revolutionized the natural language processing world. Models like GPT and BERT that rely on the Transformer architecture have fully outperformed the previous state-of-the-art networks. It outperformed the earlier approaches by such a wide margin that all the recent cutting-edge models seem to be relying on these Transformer-based architectures. In this paper, an overview and explanations of the latest models are provided. The auto-regressive models such as GPT, GPT-2, and XLNET, as well as the auto-encoder architectures such as BERT and a lot of post-BERT models like Roberta, ALBERT, ERNIE 1.0/2.0 are covered. Models that use RNNs, GRUs and LSTMs are not perfect; the inherent recurrent structure makes them very hard to parallelize on multiple processes, and the treatment of very long clauses also becomes problematic due to the diminishing gradient. [2] This paper provides a comprehensive and systematic review of research in the field of summary summarizing published from 2008 to 2019. There are 85 journal journals and seminars that are the result of the release of selected studies and analysis studies to explain research topics/trends, data sets, advanced analysis, features, strategies, methods, assessments, and problems in this field of research. The results of the analysis provide an in-depth description of the topics/trends on which their research is based; provide references to public databases, pre-processing and features used; describes the techniques and methods commonly used by researchers as a comparison and methods to improve methods. [3] In this paper, a general end-to-end approach to sequence learning that makes minimal assumptions about the sequence structure is presented. This paper used a method that uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of fixed dimensionality, and then another deep LSTM that decodes the target sequence from the vector. The main outcome is that on an English-to-French translation task from the WMT'14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. [4] In this paper, the dominant sequence transduction models are actually based on complex RNN or CNN in an encoder-decoder configuration. The models with the best performance also connect the encoder and decoder through an attention mechanism. This model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, which is an improvement over the existing best results, including ensembles by over 2 BLEU.

Here it is shown that the Transformer can actually generalize well to other tasks by being applied successfully to English constituency parsing both with large and limited training data. Attention-based models are going to have an exciting future and the plan is to apply them to other tasks. The plan is also to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle some large inputs and outputs such as images, audio and video. [5] HunFlair is a NER tagger that is integrated into the widely used NLP framework Flair. Its task is to recognize five biomedical entity types, reach or overcome state-of-the-art performance on a wide set of evaluation corpora, and it is trained in a cross-corpus setting to avoid corpus-specific bias. [6] In this research paper, the author presents their novel approach to provide ontological-based concept extraction of information to the user. [7] In this research paper, the authors have worked on the

project by classifying the Casual-Conversation subreddits post by their tag names which correctly classifies the particular topic of the post. Some of the ML algorithms used are LightGBM, XGBoost, Naïve Bayes, Logistic regression, and Linear SVM. [8] In this paper, categorical news headlines and summaries for a Turkish news agency website have been collected using web scraping methods and the classification of the testing data was done using the "one-hot encoding" method along with deep learning and vector learning methods. 90% accuracy has been achieved using this classification technique. [9] The author divides the type of classifiers used to predict the sentiment of a text document into 2 types, Lexicon based approach which used a pre-defined dictionary of words and tries to match it with the text document provided by us and then estimates the overall polarity of the document ex: (SentiWordNet and Word Sense Disambiguation). A stacked model is also used which combines both lexicon and ML-based algorithms such as naïve bayes and SVM. [10] In this paper the author presents his work on detecting the events and classifying them into categories of event or not event, The main goal is to identify the entity of a tweet of a particular topic and linking them to external knowledge sources such as DBpedia etc. The categories chosen for classification for sports and politics. There are two configurations for label prediction, The first configuration classifies if the tweet is event based or not and second configuration in which non-event based are treated as a separate category. The underlying Entity in the text is first identified using NERD-ML which is NE API which is then passed through external knowledge resources called YAGO and DBpedia to replace the entity recognized in the tweet. [11] In this paper, The author has utilized the pre-trained embedding, sub-word embedding, and closely related languages of languages in the code mixed corpus to create a meta-embedding. They then used a Transformer to encode the code mixed sentence and use Conditional Random Field to predict the Named Entities in the code-mixed text. In contrast to classical Named Entity recognition where the text is monolingual, our approach can predict the Named Entities in code-mixed corpus written both in the native script as well as Roman script. Our method is a novel method to combine the embeddings of closely related languages to identify Named Entity from Code-Mixed Indian text is written using a native script and Roman script in social media. [12] Summarization based on text extraction is inherently limited, but generation-style abstract methods have proven challenging to build. In this work, they proposed a fully data-driven approach to abstractive sentence summarization. Our method utilizes a local attention-based model that generates each word of the summary conditioned on the input sentence. [13] As a new way of training generative models, Generative Adversarial Nets (GAN) that uses a discriminative model to guide the training of the generative model has enjoyed considerable success in generating real-valued data. [14] Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. The authors demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task. In contrast to previous approaches, The authors make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. [15] Recently pre-trained models have achieved state-of-the-art results in various language understanding tasks which focuses on training the model with several simple tasks to grasp the co-

occurrence of words or sentences. However, besides co-occurring information, there exists other valuable lexical, syntactic and semantic information in training corpora, such as named entities, semantic closeness and discourse relations. In order to extract the lexical, syntactic and semantic information from training corpora, The authors propose a continual pre-training framework named ERNIE 2.0 which incrementally builds pre-training tasks and then learn pre-trained models on these constructed tasks via continual multi-task learning which captures lexical, syntactic and semantic aspects of information in the training data. Experimental results demonstrate that ERNIE 2.0 model outperforms BERT and XLNet on 16 tasks. Our work focus on extracting and recognizing chemical entities from beauty product reviews using novel data augmentation techniques and supervised modelling techniques.

3. Dataset

Dataset for the chemicals present in the beauty products on the web isn't available as such. The data was extracted using selenium from the website seknd.com. Even then to improve the training and testing accuracy of the model we augmented the data. [16] For data augmentation there are four main operations which we applied i.e. Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), Random Deletion (RD).

1. Synonym Replacement: Synonym replacement is a technique in which we replace a word by one of its synonyms. We randomly select n words, and replace them by their synonyms. This function can then be used in an apply function on a data frame for example. To create a larger diversity of sentences, one could try to replace 1 word, then 2, then 3, and so on.

2. Random Insertion (RI): Identifying and extracting synonyms for some randomly chosen words that are not StopWords in the sentence. Inserting this identified synonym at some random position in the sentence.

3. Random Swap (RS): Randomly choose two words in the sentence and swap their positions. This may not be a good idea in a morphologically rich language like Hindi, or Marathi as it may entirely change the meaning of the sentence.

4. Random Deletion (RD): Randomly remove each word in the sentence with probability p.

Table 1. Concise summary of the dataset

# Training golden set	# Training +Augmented set	# Test
5000	10000	500

4. Methodology

We briefly explain the model architecture in the following section and using Figure 2.

4.1. Data Extraction

Data extraction was done using a Python library Selenium a tool which is used for automating web browsers to do a number of tasks. One such is web-scraping to extract useful data and information that may be otherwise unavailable. Through Selenium's methods, web elements were located by HTML tags, attributes, or text content, and data was extracted using getText(), getAttribute(), or getPropertyValue() methods. Dynamic

web elements were also handled, including those that load or change content dynamically, by using explicit or implicit waits, or handling AJAX requests. Selenium's features for handling common web interactions, such as form filling, button clicking, scrolling, and capturing screenshots, were utilized to simulate user interactions with the web pages and extract data as needed. (It's important to note that all web scraping activities were conducted in compliance with the website's terms of service and legal requirements, ensuring respect for the website's policies and adherence to applicable laws and regulations, including copyrights and intellectual property rights.)

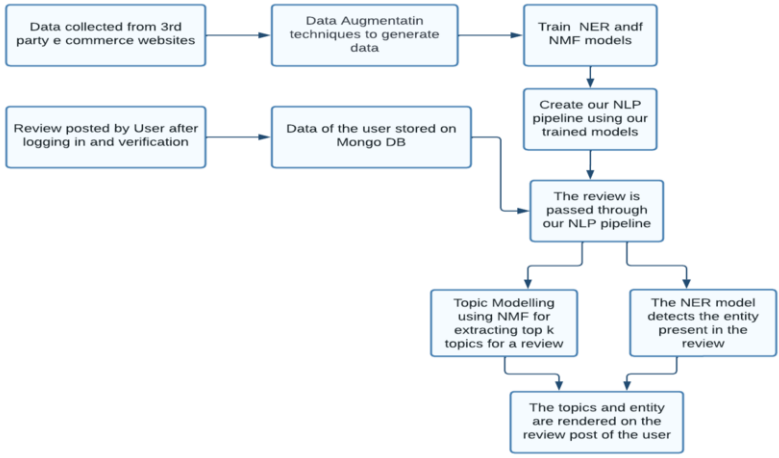

Figure 2. Methodology of the whole model architecture

4.2. Data Cleaning

The user reviews extracted using selenium from the SEKND website contain a lot of stop words whose removal is necessary to start with the applying NLP techniques. We then look for the most frequent ingredients (chemicals) that are present in the reviews which are necessary for building our tf-idf and implementing our topic modelling algorithm We also apply the word2vec algorithm to convert the text corpus into a feature space of similar words occurring closer to each other using different similarity and distance metrics. Data cleaning is a critical step in preparing the dataset for training a BERT model for named entity recognition (NER) of chemicals in user reviews. The quality of the data used for training directly affects the accuracy and performance of the model. In this research project, several data cleaning techniques were employed to ensure the integrity and reliability of the dataset. First, the collected user reviews were thoroughly inspected for errors, inconsistencies, and irrelevant information. This involved visually inspecting the reviews to identify and correct any data entry errors, such as misspelled or misplaced chemical names. Inconsistent or irrelevant information, such as emojis, special characters, or HTML tags, were also removed to ensure that the dataset contained only relevant text data for training the BERT model. Next, data validation techniques were applied to verify the accuracy and consistency of the chemical names in the dataset. This involved checking for expected patterns, rules, and formats of chemical names. For instance, chemical names were validated against

established chemical nomenclature rules, such as IUPAC, to ensure that the names were accurate and consistent. Data consistency and standardization were also ensured during the data cleaning process. Inconsistent naming conventions, units of measurement, or formatting of chemical names were checked and corrected to ensure consistency and comparability across the dataset. Chemical names were standardized to a common format, such as using the IUPAC naming convention, to ensure accurate identification and consistent training of the BERT model. The above stated processes were performed iteratively, with multiple rounds of inspection and validation to ensure that the dataset was thoroughly cleaned and reliable for training the BERT model.

4.3. Models Used

a.) Non-negative matrix factorization

Used NMF for Topic Modelling. Non-negative matrix factorization (NMF or NNMF), also non-negative matrix approximation is a group of algorithms in multivariate analysis and linear algebra where a matrix is factorized into two subsequent matrices such that all three matrices have no negative elements as shown in Figure 3.

The preprocessed text data is converted into numerical representations for input to the NMF algorithm. This is typically achieved using a term-frequency-inverse-document-frequency (TF-IDF) vectorizer or word embeddings, such as Word2Vec or GloVe, to represent the text data in a continuous vector space. We have utilised to Word2Vec embeddings to train our NMF model. The NMF algorithm is then applied to the feature matrix obtained from the previous step. NMF is an unsupervised learning algorithm that factorizes the feature matrix into two non-negative matrices: a topic-term matrix and a document-topic matrix. Hyperparameter tuning, such as the number of topics, iterations, and initialization method, is performed to optimize the model's performance.

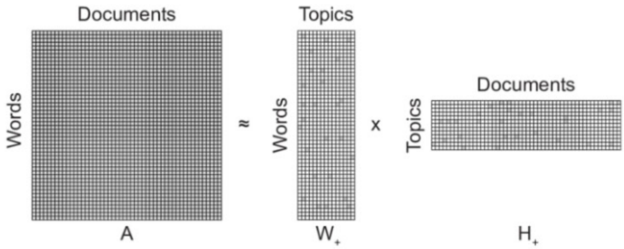

Figure 3: [17] Conceptual illustration of non-negative matrix factorization (NMF) decomposition of a matrix consisting of m words in n documents into two non-negative matrices of the original n words by k topics and those same k topics by the m original documents

b.) Bidirectional Encoder Representations from Transformers

The training of the BERT [18] model for NER of chemical substances in product reviews involves several key steps. Firstly, a comprehensive dataset of product descriptions, containing information about various chemical substances used in the

products, is collected by extraction and the augmentation techniques mentioned in the previous sections to mitigate bias, underfitting and to increase the diversity of the dataset, Such techniques results in a larger and more varied training set. The collected dataset is preprocessed to prepare it for training. This includes tokenizing the text into individual words or subwords, and encoding the text into numerical representations that can be fed into the BERT model. Special care is taken to handle rare or out-of-vocabulary (OOV) words, as they may occur in chemical names or product descriptions. The BERT model is then fine-tuned using the preprocessed dataset and on the augmented+ original dataset. Fine-tuning involves training the model on the dataset, using the labeled data for NER annotations as ground truth. The model learns to recognize and label chemical substances in the text, leveraging the contextualized representations provided by BERT, which captures the contextual information of words in the text as shown in Fig 4 where the embedding procedure is demonstrated in detail. During the training process, hyperparameter tuning is performed to optimize the model's performance. This includes experimenting with different learning rates, batch sizes, and epochs, as well as adjusting BERT-specific hyperparameters, such as the maximum sequence length. For our experiments we have used the BERT-base model with a batch size of 32, learning rate of 1e-5 and 5 epochs.

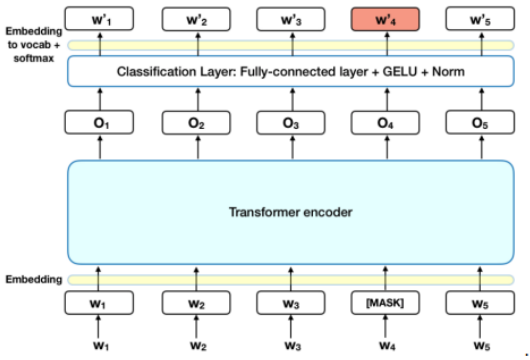

Fig 4: Illustration BERT Transformer architecture

5. Result

As we have established, the objective was to create a website which, after user verification and log-in, allows users to search or look for the different chemicals present in the beauty product they use. The work is in the form of a web application. It initiates after a log-in procedure. The web app provides a platform to also show reviews of various products. The user can also look for products based on some keywords present in the reviews. We achieved an accuracy of 86% even after training our model on training on augmented data, the accuracy of the model increased by 2.5% which is 88.5%.

Table 2. Model Results

Model	Accuracy on the extracted dataset	Accuracy on extracted + Augmented
Bert-base	0.86	0.89

We see that by augmenting just a fraction of our original dataset, augmentation improves the current performance on the extracted dataset and restricts overfitting on the smaller size of the dataset.

6. Conclusion and Future Work

In this work, we have shown that Transformer architectures and pre-training can help in the facilitation of the building of higher-capacity models and effectively utilize this capacity for a wide variety of tasks. Beauty products reviews on online e-commerce websites. On the market, there are thousands of different cosmetic products, all with different combinations of ingredients. Methods such as token identification and entity recognition transformer model advancing in both model architecture and model pretraining are implemented to identify and categorize the named entities. The work is in the form of a web app for which the front end has also been built as the final output.

We propose a fine-tuned Bert model for our named entity recognition task to identify the chemicals used in a particular product review, Capturing these entities and topics generated by our trained NMF algorithm makes it easier for the user on our application to navigate user blogs and find desired results. We also generate an augmented dataset using various novel techniques to mitigate bias and prevent the model from overfitting, our results suggest that the augmentation techniques used boost our results and make way for future works to incorporate such an algorithm in a low-resource setting.

In Future work, we aim to fine-tune other large language models for the named entity and span detection of chemical substances identification on online e-commerce websites and benchmark our results on different metrics and a larger corpus to make it generalisable for this task. Our initial study's contribution should provide directions for future research on improving current data augmentation techniques that preserve the original context of a sentence or document and increase the diversity of the entire dataset to improve results and increase model robustness over varied lengths of data.

References

[1] A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 179- 183, DOI: 10.15439/2020F20.

[2] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi,ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.05.006

[3] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27 (2014).

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, Illia Polosukhin, 2017. Attention is All you Need. https://arxiv.org/abs/1706.03762

[5] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, Alan Akbik, HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition, *Bioinformatics*, Volume 37, Issue 17, 1 September 2021, Pages 2792–2794, https://doi.org/10.1093/bioinformatics/btab042

[6] Ahmed, Adeel & Saif, Syed. (2017). DBpedia-based Ontological Concepts Driven Information Extraction from Unstructured Text. International Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.080954.

[7] Heijden, Guido & Aslam, Suhaib. (2018). Classifying r/CasualConversation Reddit Posts By Flairs.

[8] F. Ertam, "Deep learning based text classification with Web Scraping methods," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1-4, doi: 10.1109/IDAP.2018.8620790.

[9] P. Harjule, A. Gurjar, H. Seth and P. Thakur, "Text Classification on Twitter Data," 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 2020, pp. 160-164, doi: 10.1109/ICETCE48199.2020.9091774.

[10] Edouard A., Cabrio E., Tonelli S., Le Thanh N. Semantic linking for event-based classification of tweets International Journal of Computational Linguistics and Applications, 12 (2017).

[11] R. Priyadharshini, B. R. Chakravarthi, M. Vegupatti, and J. P. McCrae, "Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding," *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 68-72, doi: 10.1109/ICACCS48705.2020.9074379.

[12] Rush, A.M., Chopra, S. and Weston, J., 2015. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[13] Yu, L., Zhang, W., Wang, J. and Yu, Y., 2017, February. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[14] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

[15] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020). ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8968-8975

[16] Jason Wei, Kai Zou, 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. ArXiv:1901.11196

[17] Kuang, Da & Brantingham, P. & Bertozzi, Andrea. (2017). Crime Topic Modeling. Crime Science. 6. 10.1186/s40163-017-0074-0.

[18] Ming-Wei Chang, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805