Project Title:

Regional Language Toxic comment classification

MSc Project Research

Yashkumar Parikh

Student ID No.: 1138765

Supervised by

Dr. Jinan Fiaidhi, PhD, ISP, IEEE Senior Member, PEng (Full Professor)

Department of Computer Science

Lakehead University

Ontario, CANADA

Reference: COMP9800_Spring_Summer_Fall2022_1138765

Date: 29/07/2022

Abstract

Social media sites are gaining popularity day by day. They are best for communication, business, entertainment, and many other things. After more than a decade, social media have become very influential. On the flip side, fake news, hate speech, and online trolls are the biggest concerns because of social media. So, a solution to curb this issue is needed, especially in regional languages. Many social media platforms support regional languages. This paper will provide a machine learning-based solution to this problem. The focus of this paper is to classify comments written in regional languages. Firstly, a dataset has been created in Gujarati, Hindi, English, Marathi, and Punjabi languages. After that, different machine learning and deep learning models are applied to the multilingual dataset. At last, a comparison of all model performances was made.

Keywords

Hate speech, Troll, Transfer Learning, BERT, RNN, CNN, Word Embedding

Introduction

Over the years, the popularity of social networking sites has skyrocketed. The internet is unquestionably a blessing for humanity. Users of social networks can use them to connect with new people, exchange their thoughts with like-minded people, and keep in touch with old friends and co-workers, among other things. People can make new friends from around the world through social networking. Popular social networking sites are designed user friendly. Most websites are so simple to access that they only require a basic understanding of the internet. In addition, social media is open to everyone and has a dominant influence on how people behave. Professional social networks aid in the development of a personal online brand. People can use these networks for advertising their abilities, accomplishments, and previous experience, allowing co-workers', other peer groups, and potential employers to identify them. Another crucial reason for social networking sites' success is that they make it easier for businesses to reach out to potential clients.

On the flip side, fake news, hate speech, and online trolls are the biggest concerns due to social media. Many things we read on social media may appear genuine, but they can be fake news. False information is news, stories, or hoaxes intended to deceive or misinform readers [18]. Hate speech is utterance intended to vilify, humiliate, or instigate hatred towards a group or class of people based on race, religion, skin colour, sexual identity, gender identity, handicap, or ethnicity [19]. A troll is a term used to describe someone who tries to incite disagreement, hate, or arguments in an online social network. Trolls may target platforms such as Twitter, YouTube comment sections, forums, or chat rooms [20]. Therefore, it is crucial to flag content on social media to maintain a safe online world. In this paper, we have provided Machine Learning based solution to classify harmful content on social media.

People all over the globe use social networks. In 2020, 518 million people used social media in India alone. However, due to numerous representations of low-resource Indic languages, detecting toxicity is a significant task. The lack of a specific format, grammar, or sentence structure in social media postings and comments further complicates the process of abuse detection for social media platforms that support multiple languages. Therefore, a Machine Learning model must be able to detect toxicity in as many languages as possible. Even though every well-known social media platform supports regional languages, recent models for multilingual toxic comment detection have only considered a few specific languages into account. In this paper, we have considered five Indic languages Gujarati, Hindi, English, Marathi, and Punjabi, for the data set. Then, Different Machine Learning and Deep learning models are applied to the data set. At last, we will compare the performance results of these models.

Literature review

Dignity, liberty, equality, individual, group, minority rights and freedom of expression are related to hate speech. The approaches to solving these problems are in three categories: dictionary-based, machine-learning-based, and deep learning-based. A keyword list is used in the dictionary-based technique to ensure accurate matching. The effectiveness of this strategy mostly depends on the vocabulary, which demands domain expertise and has limited generalization capacity. As a result, the approach integrated with additional text features rather than working as a stand-alone technique. Feature engineering must be done manually for traditional machine-learning models. N-gram, part-of-speech (POS), syntactic dependence and keyword significance features are in natural language processing (NLP). [6] The chosen features represent the input text sample. A learning algorithm fits the data in the training set to reduce the prediction error iteratively until convergence to train a classifier. Various text classification tasks have shown that these feature-based learning models perform satisfactorily. However, the effectiveness of feature engineering mostly depends on the model developer's subject expertise, which is typically difficult to come by. On the other hand, deep neural networks can overcome this problem by extracting text semantic information from raw text input without the need for manual feature engineering while also improving detection performance. Various text classification studies have lately used deep learning techniques to achieve this.

Researchers are extensively using transfer learning in NLP tasks. A model created for one job is used as the basis for another using the machine learning technique known as transfer learning. The model can better understand the meaning of a letter, word, or sentence within its context when pre-trained to encode contextual information embedded in the raw data. The pre-trained state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) model has recently been used to report impressive results. Many researchers have deployed Bert-based models on several NLP problems, such as sentiment classification, intent detection, and emotion classification [4]. BERT uses Masked Language Model (MLM), a ground-breaking language model that enables self-supervised training on big text corpora.

Even though the research in hate speech detection is expanding quickly, one of the current problems is that most data sets are only available in English. As a result, hate speech in other languages is not identified, which could be harmful. It is challenging for businesses like Facebook, as they can only remember hate speech in specific languages (like English and Spanish). [17] Although there are few data sets in other languages, those available are, as we have seen, generally relatively small. Therefore, flagging incorrect social media content to foster a positive online environment is crucial. This research has five Indic languages for the data set. The data is helpful for several machine learning and deep learning models. Finally, these models will be contrasted based on performance outputs.

Methodology

Methodology is divided into six parts as Data Prepossessing, Word Tokenization, Design Models, Results, Evaluation and Pseudo code.

i. Data Prepossessing

I have gathered data from various social media platforms in Indic languages like Gujarati, Hindi, Marathi, Punjabi, and English. The shape of the data is (14995, 2). It has two columns comment_text and toxic. The comment_text column contains actual text, and the toxic column contains '1' for toxic and '0' for non-toxic comments. The dataset contains 7500 non-toxic and 7495 toxic comments. The pre-processing step removes URLs, hashtags, mentions, punctuations, and extra white spaces from the comments. Plus, removed rows with empty or null values. Most pre-processing libraries support popular languages, so it is impossible to perform lower casing, stemming and lemmatization for regional languages.

ii. Word Tokenization

Tokenization involves breaking the raw text into manageable pieces. Tokenization divides the original text into tokens, which are words and sentences. These tokens help in context comprehension or model development for NLP. By examining the word order in the text, tokenization aids in comprehending the text's meaning. The comment_text is then converted into tokens.

iii. Design Models

A. Logistic Regression

Logistic Regression is a supervised learning algorithm.[25] It can model the probability of a specific class or category. It is applied when the outcome is binary, and the data may be linearly separated. That means problems involving binary classification are solved using Logistic Regression. [25]

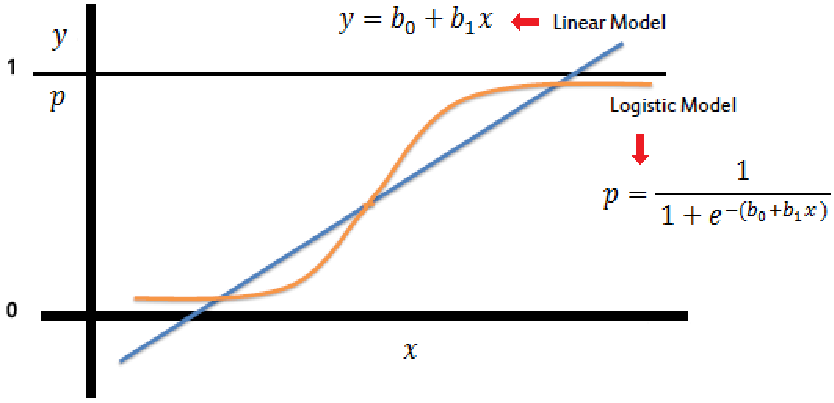

FIGURE 1 : LOGISTIC REGRESSION

This model converts the tokens into Tf-idf vectors using the SKLearn library and then trains the model using the following hyperparameters.

Iterations	500
Cross validation	6
Random State	0

B. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm. It is helpful for both classification and regression challenges. SVM categorizes data points even when they are not linearly separable by mapping the data to a high-dimensional feature space.[24] Once a separator divides the categories, the data are converted to make it possible to draw the separator as a hyperplane.

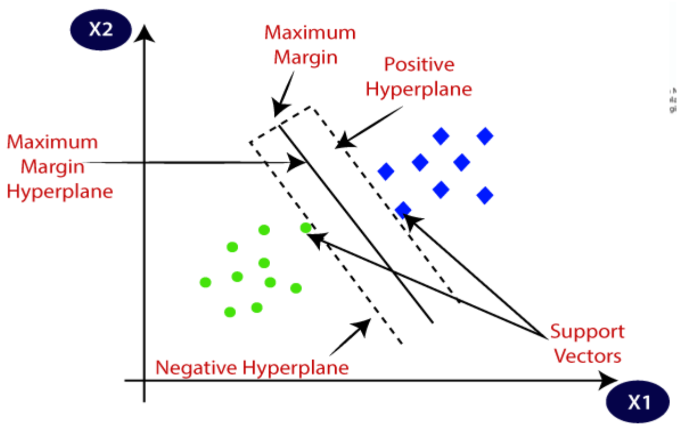

FIGURE 2 : SVM

This model converts the tokens into Tf-idf vectors using the SKLearn library and then trains the model using the following hyperparameters.

Iterations	500
Random State	0

C. Long short-term memory (LSTM)

Recurrent neural networks (RNNs) have a long-term dependency issue that LSTM networks solve. LSTMs include feedback connections.[26] With feedback connection property, LSTMs may process whole data sequences without considering each data point individually. Instead, they can process new data points by using the information from earlier data in the sequence to assist their processing. LSTMs are helpful for processing data sequences like text, audio, and general time series.

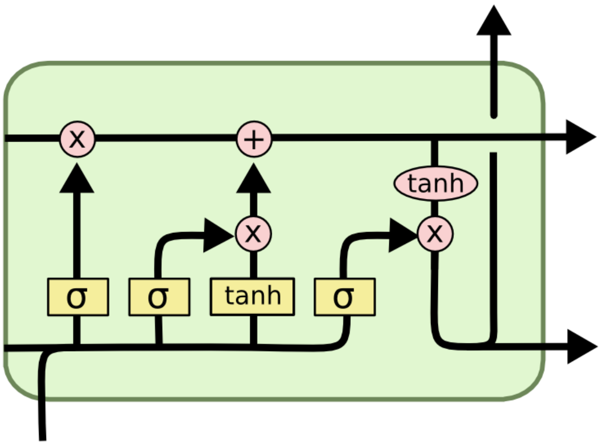

FIGURE 3 : LSTM ARCHITECTURE

This model has four layers Embedding (Input), LSTM and two Dense (Output). It is trained using the following hyperparameters.

Epochs	5
Batch size	128
Activation	Sigmoid
Loss Function	binary_crossentropy
Optimizer	Adam
Regularization Parameter	0.01

D. Convolutional neural networks

The most popular deep learning architectures for image processing and recognition are Convolutional neural networks (CNNs). Nevertheless, CNN's have recently become common in solving NLP-related issues. This technique treats each comment as an image by displaying the text in vector form and applying a CNN.

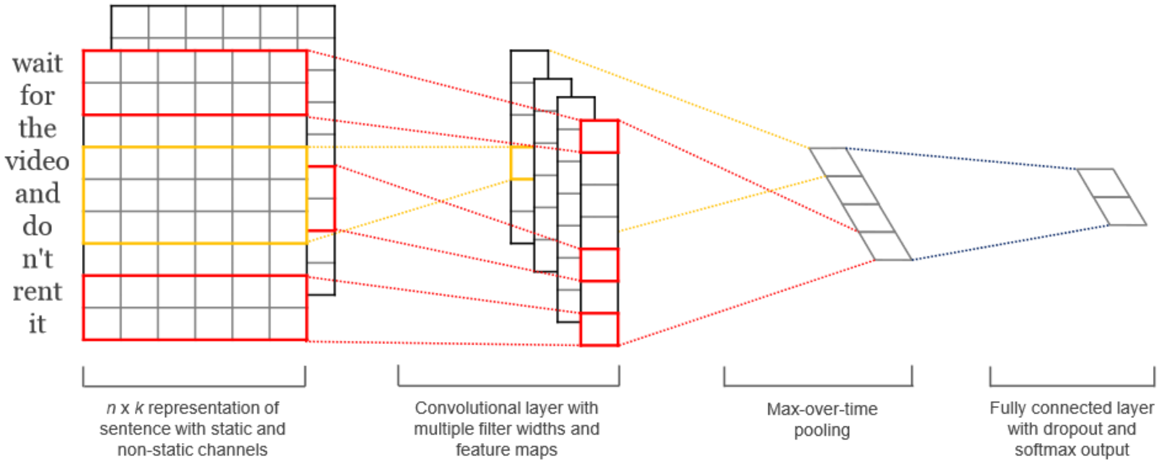
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

FIGURE 4 : CNN ARCHITECTURE

This model architecture contains one embedding, two convolutions, two max-pooling, one flattened, and two dense layers. At last, the model is trained using the following hyperparameters.

Epochs	5
Batch size	128
Activation	Sigmoid
Loss Function	binary_crossentropy
Optimizer	Adam

E. DistilBERT Embedding

Bidirectional Encoder Representations from Transformers are known as BERT. [27] Towards the end of 2018, Google created and presented a brand-new language model. BERT is a multi-layer bidirectional Transformer encoder based on fine-tuning. [28]

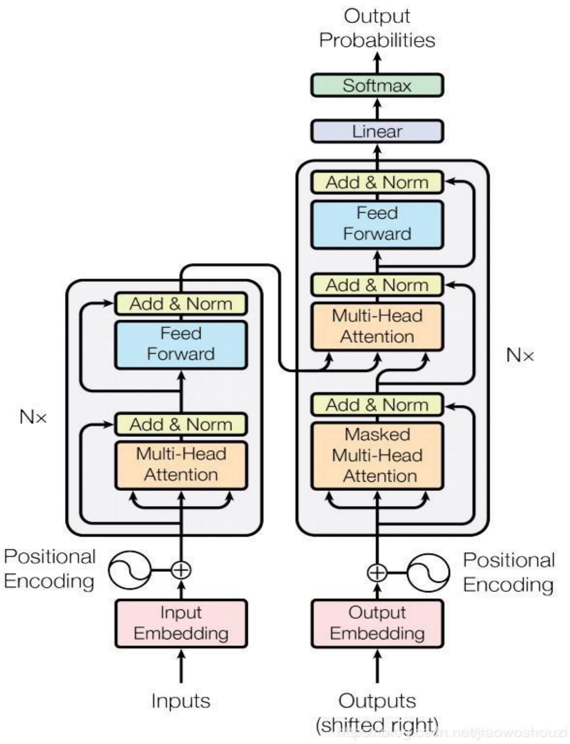

FIGURE 5 : BERT ARCHITECTURE

Many natural language processing applications, including Question Answering, Named Entity Recognition, Natural Language Inference, and Text Classification, depend on pre-trained language models like BERT.

DistilBERT is a Transformer model based on the BERT architecture. To maximize training efficiency, DistilBERT strives to keep as much performance as possible while lowering the size of the BERT and speeding up the BERT. It is 40% smaller, 60% faster, and 97% functionally equivalent to the original BERT-base model.

Pretraining with language models has two benefits. First, it just needs unlabelled text corpora, which are available even in low-resource contexts. Secondly, finetuning is cheap and can be repeated several times on the same pre-trained model once pretraining is complete.

This last model uses the Hugging faces library. This library provides different pre-trained models and word tokenizers. Moreover, this model operates on a pre-trained model called ‘distilbert-base-multilingual-cased.’ It is finetuned using our multilingual dataset. This model architecture has two inputs, dropouts and dense layers. Finally, this model employs the following hyperparameters.

Epochs	5
Batch size	5
Loss Function	binary_crossentropy
Optimizer	Adam

The Hugging faces library provides TFDistilBertForSequenceClassification model class. This model has an output classification layer on op of the pre-trained model. Therefore, this model can be directly trained on the training data using above hyperparameter.

iv. Results

After training all four models on multilingual datasets, the training and testing accuracy is as follows,

Model	Train Accuracy	Test Accuracy
Logistic Regression	99%	88%
SVM	97%	87%
LSTM	92%	85%
CNN	91%	87%
DistilBERT	78%	60%

v. Evaluation

As per the results, Logistic regression is the best-performing model on the given multilingual data set. An evaluation method is created to detect real-time toxicity. This method uses the trained model to predict toxicity. The evaluation method first loads the trained model, pass the input text to model and the model give results as output.

vi. Pseudo code

A. Data Pre-processing

```
1. Begin
2. Load the Dataset
3. Remove hashtags, mentions, URLs, and retweets from comment text
4. Remove null/empty rows
5. Save the dataset into a new file
6. End
```

B. Model Training

```
1. Begin
2. Load clean dataset
3. Convert the dataset into vectors and add padding
4. Split the dataset into train and test set
5. Define classification algorithm method/architecture
6. Train the model on training data
7. Test the model on testing data
8. Calculate the result
9. Save the model architecture, weights and hyperparameters
10. End
```

Conclusion

This paper is the first large-scale analysis of multilingual toxic comment classification. Using datasets created in five languages, we use machine learning models to develop classifiers for multilingual toxic comment classification. Moreover, we have also seen the benefits of transfer learning. Finally, we have performed many experiments under various conditions.

We can see that logistic regression and SVM with Tf-idf vectorization performed better than other deep learning models. In addition to observations, we found that on predictions, the size and ratio of the data are essential for model performance. Multilingual models also provide some other benefits. First, deploying numerous monolingual systems in a production environment might not be practical because of resource limitations. One possible approach is replacing various monolingual models with a single multilingual one.

To conclude, toxic posts or comments can be flagged or blocked using these models with social media platforms, and a safe online environment can be created.

Future Work

We can increase the dataset size for future work to cover more edge cases. In this research, a model-centric approach is followed. In future, we can shift this approach to data centric. We can focus more on data collection and pre-processing in the data-centric method. Many researchers like Andrew Ng believe that data-centric systems can significantly improve model performance.

Acknowledgements

I am writing to express my gratitude for Professor Dr. Jinan Fiaidhi's outstanding contributions. She not only pointed me in the right direction, but she also encouraged me so that I could efficiently complete the project. I appreciate her efforts and helpful suggestion for the project's improvement.

References

1. Jan Christian Blaise Cruz, & Charibeth Cheng. (2019). Evaluating Language Model Finetuning Techniques for Low-resource Languages. ArXiv: Computation and Language. https://arxiv.org/pdf/1907.00409v1
2. Biswas, E., Karabulut, M. E., Pollock, L., & Vijay-Shanker, K. (2020). Achieving Reliable Sentiment Analysis in the Software Engineering Domain using BERT. 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME). https://doi.org/10.1109/icsme46990.2020.00025
3. Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. IEEE Access, 8, 138162–138169. https://doi.org/10.1109/access.2020.3012595
4. Gati Lother Martin, Medard Edmund Mswahili, & Young-Seob Jeong. (2021). Sentiment Classification in Swahili Language Using Multilingual BERT. ArXiv: Computation and Language. https://arxiv.org/pdf/2104.09006
5. Ostendorff, M., Calizzano, R., & Rehm, G. (n.d.). DFKI SLT at GermEval 2021: Multilingual Pre-training and Data Augmentation for the Classification of Toxicity in Social Media Comments. Journal.
6. Song, G., Huang, D., & Xiao, Z. (2021). A Study of Multilingual Toxic Text Detection Approaches under Imbalanced Sample Distribution. Information, 12(5), 205. https://doi.org/10.3390/info12050205
7. Jhaveri, Manan, Ramaiya, Devanshu, & Chadha, Harveen Singh. (2022). Toxicity Detection for Indic Multilingual Social Media Content. Cornell University - ArXiv. http://arxiv.org/abs/2201.00598
8. Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, & Partha Pratim Talukdar. (2021b). MuRIL: Multilingual Representations for Indian Languages. ArXiv: Computation and Language. http://arxiv.org/pdf/2103.10730.pdf
9. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2017). Attention is All you Need. Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

10. Resume Screening with Python. (2022, April 11). Thecleverprogrammer. https://thecleverprogrammer.com/2020/12/06/resume-screening-with-python/
11. Spooky NLP and Topic Modelling tutorial. Kaggle. https://www.kaggle.com/code/arthurtok/spooky-nlp-and-topic-modelling-tutorial/notebook
12. Mohdsanadzakirizvi@gmail.com Sanad. (2020, June 14). 3 Important NLP Libraries for Indian Languages You Should Try Out Today! Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/01/3-important-nlp-libraries-indian-languages-python/
13. Hugging Face – The AI community building the future. (n.d.). Retrieved October 31, 2022, from https://huggingface.co/
14. Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv: Computation and Language. https://arxiv.org/pdf/1810.04805v2
15. Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1493
16. Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). https://doi.org/10.1109/icacsis.2017.8355039
17. Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, & Animesh Mukherjee. (2020). Deep Learning Models for Multilingual Hate Speech Detection. ArXiv: Social and Information Networks. https://arxiv.org/pdf/2004.06465.pdf
18. McGarrigle, J. (2021, May 13). Explained: What is Fake news? | Social Media and Filter Bubbles. Webwise.Ie. https://www.webwise.ie/teachers/what-is-fake-news/
19. KPEKOLL, “Hate Speech and Hate Crime,” Advocacy, Legislation & Issues, Dec. 13, 2017. https://www.ala.org/advocacy/intfreedom/hate.
20. “The Now: What is Trolling?” GCFGlobal.org, 2019. https://edu.gcfglobal.org/en/thenow/what-is-trolling/1/.
21. Neural Networks for NLP. (2019, November 12). Devopedia. https://devopedia.org/neural-networks-for-nlp
22. CS 230 - Recurrent Neural Networks Cheatsheet. (n.d.). Retrieved November 1, 2022, from https://stanford.edu/%7Eshervine/teaching/cs-230/cheatsheet-recurrent-neural-networks
23. Shreya Ghelani, “Breaking BERT Down - Towards Data Science,” Medium, Jul. 26, 2019. https://towardsdatascience.com/breaking-bert-down-430461f60efb.
24. Support Vector Machine (SVM) Algorithm - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
25. Logistic Regression. (n.d.). https://www.saedsayad.com/logistic_regression.htm
26. Dolphin, R. (2022, February 28). LSTM Networks | A Detailed Explanation | Towards Data Science. Medium. https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9
27. Ingle, P. (2022, December 1). Top Natural Language Processing (NLP) Tools/Platforms. MarkTechPost. https://www.marktechpost.com/2022/11/30/top-natural-language-processing-nlp-tools-platforms/
28. Ghelani, S. (2021, December 11). Breaking BERT Down - Towards Data Science. Medium. https://towardsdatascience.com/breaking-bert-down-430461f60efb