Benchmarking a New Approach to Natural Language Understanding Using Non-Neural Network-Based Artificial Intelligence Technology (CRI™)

Research and Development Department, AHvos™ Corporation
Austin, Texas, USA

Abstract

We present a basic benchmark for a new approach to Natural Language Understanding (a subset of NLP) utilizing non-neural network-based Artificial Intelligence ("AI") that matches, and in some instances, outperforms state-of-the-art neural network-based NLP technology. This new procedure uses AHvos™ Corporation's proprietary CRI™ ("Contextually Responsive Intelligence") AI technology. The English language subset of the Multilingual Amazon Reviews Corpus ("MARC") is used for the benchmark. The benchmark starts with a CRI™ AI engine that has no knowledge or pre-training of any kind related to any text language and it ends with text classification of the English subset of MARC. CRI™ technology demonstrates a 91.93% text classification accuracy (accuracy benchmark using neural network-based NLP AI models report results in the 59% to 73% accuracy range). The total AHvos™ benchmark elapsed time was 1.8 seconds. The CRI™ benchmark is comprised of 2 phases: the Education Phase (called the Training Phase of neural networks' NLP AI models) and the Text Classification Phase. Comparatively, neural network-based NLPs utilize 3 phases: (1) the English-based AI LLM model generation, (2) the fine tuning of the AI model for text classification specialized in the Amazon Reviews Corpus, and (3) the text classification. It is important to note the last 2 phases of neural network-based NLP are reported to require 10 hours on average to complete. The AHvos™ benchmarking exercise was fully completed using a non-proprietary laptop with core i5 7th Gen CPU, 24 GB of RAM, and no GPU. CRI™ technology correctly identified mislabeled data in the English Amazon Reviews Corpus both during the Education and the Text Classification phases. CRI™ provides a clear linguistic rationale for each of its text classification decisions resulting in a significant improvement in the understanding of AI classification responses versus conventional neural network technology. Special attention should be given to the characteristic learning curve measured during the Education Phase showing the ability for the CRI™ engine to self-detect when it has reached knowledge equilibrium (the "KE" point). KE signals that the engine has "learned the concept" and no additional data is required for educating the engine, resulting in faster and less data intensive Education Phase completion times. The CRI™ engine consumed only 25.00% of the available training dataset provided by the English Reviews Corpus to reach KE. Further research has been completed by the Research and Development Department at AHvos™ Corporation to demonstrate that CRI™

engines are equally valid when using non-English datasets included in the Amazon Review Corpus as well as other types of text-based datasets producing both similar and improved accuracy and performance results to the ones described in this benchmark.

1 Introduction

Natural Language Processing (“NLP”) and its subset Natural Language Understanding (“NLU”) are an integral part of current efforts in AI. While Large Language Models (“LLMs”) (such as ChatGPT) have made the topic popular, NLP research has been one of the forefronts of AI research in academic and private sectors for more than 20 years. Advances in this area are the product of collaborative work between linguistics, statisticians, and computer scientists (Lewis at al., 2004) (Bowman et al., 2010).

Human language, with its multitude of permutations, dialects and grammatical rules is an excellent source of data for use when evaluating measurable “intelligence” of AI models and related AI support technologies such as transformers. One of the most common tasks studied by NLP researchers is the ability for AI methods to perform text classifications.

Over the years, and thanks to the popularity of text-based social media platforms and e-commerce sites, many language datasets formatted specifically for NLP AI research have become publicly available (Bowman et al., 2015) (Pak et al., 2010). Datasets include single and multiple languages.

Some NLP research is focused on text classification tasks within a single language, such as text summarization, text labeling, and text generation. Other AI research is aimed at developing AI algorithms that can be applied across multiple languages, such as language detection, text labeling, and sentiment analysis (Keung et al., 2020).

Textual Sentiment Analysis is a form of NLP text classification aimed at labeling text based on a summary of the “sentiment” expressed in a piece of text. While the resulting label is in some instances represented via a single word (positive, negative) or a symbol (smile icon, unhappy icon, thumbs up icon, thumbs down icon), it is most generally captured as a range of repeated symbols. The range goes from negative sentiment (single symbol such one star icon or one thumbs up icon) to positive sentiment (5 stars or thumbs up icons, for example).

Over the last 70 years, and since its inception, the field of AI has been predominantly characterized by the utilization of perceptrons and the neural networks that they form.

While many attempts have been made to formulate that neural networks imitate or model the types of networks and interactions between neuron cells (and their dendrites) as observed in biological brains, these formulations lack the explanatory power of the complexity of behavior exhibited by their biological counterparts.

Many AI researchers argue that these critiques are the byproduct of neural networks not having reached yet the large number of perceptron nodes and neural connections typical of biological brains.

A limited, smaller group of AI researchers, both in the academic and private sectors, aim to find robust and viable alternatives to neural network-based AI methods and related technologies without sacrificing performance and/or accuracy.

AHvos™ Corporation is primarily an AI research private organization that has the ability to complete advanced AI tasks without the use of neural networks and without the utilization of the resources typically associated with complex neural networks (such as supercomputers, large amounts of labeled training datasets, and long training times -months to years).

NLP Textual Sentiment Analysis is considered an advanced AI task, and most of its AI research has been limited to the use of neural network-based AI approaches.

This benchmark exercise is focused on NLP Textual Sentiment Analysis utilizing the English subset of MARC (Keung et al., 2020) using AHvos™ Corporation's proprietary CRI™ ("Contextually Responsive Intelligence") AI technology. This technology does not relay, use, or leverage neural networks. The aim of the benchmark is to showcase a non-neural network-based AI technology accomplishing this type of NLP task.

MARC was selected because it is one of the best publicly available datasets to study and validate the behavior and accuracy of NLP Textual Sentiment Analysis AI methodologies. This dataset is characterized by a well-balanced set of product reviews in English, Japanese, German, French, Spanish and Chinese. The reviews have been curated using well established AI data standards (Keung et al., 2020).

Each review, independent of language, provides all the information required to train and validate both neural network-based AI models and non-neural network-based AI technologies. Information in the dataset includes review ID, review title, review body, review label (i.e., classification) expressed as a range from 1 star (negative) to 5 stars (positive), and product category.

Each language dataset in MARC contains 200,000 training reviews, 5,000 development reviews, and 5,000 test reviews. That means that for each classification label there are a total of 40,000, 1,000 and 1,000 training, development and testing reviews respectively.

2 Data Preparation

2.1 Data Characteristics

For purposes of this benchmark, we utilize the English subset of MARC. We focus on the following fields included in the dataset (fields not mentioned are ignored even if present in the dataset): (1) review ID, (2) review user label, (3) review title, and (4) review body. We keep the reviews in separate subgroups to be able to differentiate between training, development, and testing subsets.

The training subset is used to educate the CRI™ engine during the Education Phase. The development subset is used to determine Knowledge Equilibrium (referred to as the "KE Point"), and to measure engine logical self-consistency during the Education Phase, and to measure engine accuracy during the Text Classification Phase. The testing subset is used to measure engine accuracy during the Text Classification Phase.

The English subset of MARC shows a preferential data length bias towards user labeled negative reviews. This bias provides more verbose review bodies than the review bodies of user labeled positive reviews. This bias does have a predicted effect on CRI™ engines resulting in a potential accuracy gain of up to 5.07% when correctly classifying negative reviews in contrast to positive reviews. The bias only affects correct classification, and it is not predicted to influence an increase on incorrectly recognizing a review's label.

The English subset of MARC exhibits a significant number of grammatically incorrect sentences within the review title and/or the review body. Incorrect grammar is common in online conversations, so this finding is not surprising. Incorrect grammar does have a predicted harmful effect in correctly understanding sentences within a review for CRI™ engines. While CRI™ engines provide a significant resilience to noise in education datasets (approximately 25.00%), if the dataset shows significant noise, as is the case with the English subset of MARC, the impact is a reduction of predicted engine accuracy of up to 9.78%.

Incorrect grammar is a type of "noise" signal when performing NLP tasks. Some examples of severe incorrect grammar were found in the reviews of the English subset of MARC such as: (1) incorrect punctuation, (2) incorrect use of ellipses, and (3) incorrect use of words (including wrong verb tenses and misspelled words).

CRI™ engines are logic engines, and they must not be confused with neural networks. There are no 'weights', or 'cost functions' used to minimize cost during the Education Phase. CRI™ engines are not LLMs either and they do not contain millions or billions of weights as commonly found in LLM AI models (Keung et al., 2020) (Devlin et al., 2019). Instead, logical engines are characterized by their ability to identify the rationale (i.e., 'logic') for the user labels provided in training datasets during the Education Phase. CRI™ engines do not blindly trust the labeling provided during the Education Phase and they are able to correctly identify mislabeled data in the training dataset during the Education Phase. A measure of the total mislabeled content provided during the Education Phase is linearly proportional to the engine's self-consistency measured when the KE point is reached, and it could be estimated as follows:

$$L_{incorrect} = C_k(1 - eng_{self}^{KE})$$

where $L_{incorrect}$ is the ratio of incorrectly labeled data present in the training dataset as identified during the Education Phase, C_k is a constant (usually equal to 1) that adjusts for labeling bias that may be present in the dataset, and eng^{KE}_{self} is the measured self-consistency of the CRI™ engine at the KE point.

From a linguistic and logical perspective, (user or engine generated) review labels are a summary of the concepts expressed in the review title and the review body. Review labels represent the dominant tone of their contents. Review titles and review bodies are the contents within a review. Review titles and review bodies are composed of sentences.

Sentences are also labeled. Label sentences represent the dominant tone of their content. Sentences are groupings of words and punctuation at specific locations within a sentence. The location of the words and punctuation are dictated by the grammatical rules of a specific language.

While grammatical rules tend to be logical for most of the relationships between words in a sentence, each language has special usage that directly contradicts the general grammatical rules for that language. Words are made of letters (and accentuation symbols for some languages different from English).

It is important to notice that the label of a review represents the 'summary' of the sentence labels that it contains. That means that two separate logical rules must be followed for the review to be considered correctly labeled.

(1) For reviews that contain more than one sentence in the review title and/or the review body, most of the sentence labels must match the label of the review. Conversely, this implies too that not all of the sentence labels within a review are equal to the review label.
(2) For reviews that contain only one sentence in the review title and one sentence in the review body, the labels of these single sentences must match the label of the review.

A simple way to understand these two logical rules is that if all sentences contained in a review are, let's say, labeled as negative sentences, the label for the review itself cannot be correctly labeled as positive.

In some isolated circumstances sarcasm is used in the review body and/or review title of a review. In these cases, because of the use of sarcasm, which is a human construct not defined by grammatical rules, the two logical rules are violated.

CRI™ engines are logical engines, and as such, they commonly fail to correctly identify sarcasm. Some of the reviews provided in the English subset of MARC use sarcasm so it is predicted that a fraction of these reviews would not be correctly identified by the engine, impacting its measured accuracy during the benchmark exercise.

A final source for violation of the two logical rules is user entry error when creating the review itself. A user selects the label for the review (i.e., its star rating) independently of writing the review title and the review body. Most of these user entry errors are correctly identified within $L_{incorrect}$ and it has no significant impact on the measured accuracy for the engine during the benchmark exercise.

2.2 Data Pre-processing

The English subset of MARC was downloaded and stored as master data files. We created new education files from the master data files to facilitate faster I/O processing. Additionally, we did some basic text string cleanup to simplify ingestion of data during the education phase such as removal of trailing spaces, etc. Sentences containing only non-alphanumeric characters were discarded to avoid unnecessary noise in the data.

Subjectivity in the logic applied by users when rating a review introduces a user's personal opinion, and this subjectivity is not shared between all users. This results in ratings that are not logically consistent across the range when the rating is separated by one or two stars. For example, what one user rates as a 4 star review is rated as a 5 star review by another user. Because of this we focused the benchmark in a binarized classification task, where data subjectivity is avoided.

Binarized classification focuses on the reviews at the ends of the range (i.e., the 1 star and 5 star user labeled reviews). This approach has been adopted previously by other benchmarks using neural network-based AI models (Keung et al., 2020).

Due to binarized classification, the total data utilized in the benchmark was limited to 80,000 training reviews available to the CRI™ engine to be used during the Education Phase, 2,000 evaluation reviews, and 2,000 testing reviews. The evaluation and testing reviews were used during the Text Classification Phase.

3 Benchmark Results

The benchmark focuses on four areas:

(1) Accuracy of the CRI™ engine to correctly classify the reviews in the validation and testing datasets.
(2) Total time it takes to complete the benchmark exercise (i.e., the time to complete both the Education Phase and the Text Classification Phase).
(3) The amount of training data needed to educate the engine.
(4) The hardware needed to perform the benchmark exercise.

Accuracy is defined as the percentage of correctly labeled reviews by the CRI™ engine using both the development and test datasets. Because CRI™ engines do not need a 'fine-tuning' phase, there is really no need to have separate development and test datasets, so both datasets are combined into one dataset and are used during the Text Classification Phase.

CRI™ engines detect when new concepts that the engine was not exposed to previously are presented during the Text Classification Phase. The detection is captured under the term Newness, which is defined as the ratio of new conceptual information perceived by the engine over the total information received by the engine:

$$Newness = \frac{\sum I_n}{\sum I_r}$$

where I_n is the detected new information perceived by the engine and I_r is the information received by the engine.

Newness, by definition, is equal to zero during the Education Phase since during this phase all information received by the engine is considered new information that the engine has not acquired yet.

Newness is predicted to impact accuracy. Because CRI™ engines are factual engines (i.e., their responses are based on acquired knowledge, not on extrapolation or guessing), the higher the newness, the lower the confidence in the responses from the engine and the lower the accuracy of these responses.

For example, the relationship between accuracy and newness is conceptually parallel to asking an expert in US history multiple questions about general history and the expert correctly identifying when a given question is not specific to US history and his responses to these non-US history questions being less accurate than his responses to US history questions.

Self-consistency is the ability for a CRI™ engine to test itself during the Education Phase on the concepts it has learned. Self-consistency is used to signal if logical inconsistencies in labeling are present in the training dataset (i.e., detecting incorrectly user labeled data). Based on the self-consistency detected, the engine calculates the maximum potential accuracy for its responses given the training data provided. Self-consistency can also be used by human operators to correct the user mislabeled data.

While self-consistency could be easily confused with the cost function of neural network-based AI, it is important to notice that they are not equal. Cost functions are used to recalibrate the weights where the perceptron interconnects to other perceptrons to reduce the measured cost during the Training Phase.

There is no recalibration of any weights or any other parameters in the CRI™ engine based on the results of self-consistency. CRI™ engines do not have perceptrons or weights and their learning behavior is not based on any minimization/maximization of cost-like functions. Self-consistency is used only to calculate the predicted maximum potential accuracy for an engine given the training data provided.

CRI™ engines use internal knowledge domains to calculate their responses. Knowledge domains are labeled for easy understanding by the human operator. For this benchmark exercise, the CRI™ engine used its *p*, *m*, *ps* and *ms* domains. The *ps* domain is given preference first, followed by the *ms* domain, the *m* domain next, and finishing with the *p* domain.

Knowledge equilibrium (known as the KE point) signals when the engine has reached its calculated maximum potential accuracy, indicating that the ‘concept’ has been learned and that no additional training data is needed to further educate the engine on the ‘concept’. While this is factually demonstrated for CRI™ engines both mathematically and experimentally, the demonstration is outside the scope of this benchmark exercise.

The key differentiator of non-neural network-based AI approaches when compared to neural network-based AI models during the Education Phase (the Training Phase for neural network-based AI models) is highlighted by the KE point. Neural networks operate under the principle that the more data that is fed into the AI model, the more robust and optimal their weights get resulting in better AI model accuracy. Non-neural network-based AI approaches that leverage KE can stop their need to be trained with additional data when the KE point is reached because feeding more data into the engine would not result in an increase of its accuracy.

3.1 Education Phase

During the Education Phase the CRI™ engine was educated on the following concepts:

(1) Reviews contain sentences.
(2) Reviews have ratings expressed as review labels.
(3) Sentences contain words.
(4) The order of words in sentences is determined by grammatical and punctuation rules.
(5) Sentences have ‘sentiment’ expressed as sentence labels.
(6) The review label is the ‘summary’ of the sentence labels for the sentences contained within a review.
(7) Words contain alphanumeric characters.

The CRI™ engine processed the training dataset at intervals of 1,000 reviews. After each interval, the CRI™ engine performed self-consistency measurements. Self-consistency reported approximately 3.92% mislabeled data in the training dataset's reviews and a maximum potential accuracy of 92.86%. Correcting mislabeled data is outside the scope of this benchmark exercise.

The KE point was reached at interval 20 as per the knowledge curve in Figure 1. At interval 20, only 20,000 training reviews were used to reach the KE point, which corresponds to 25% of the total 80,000 training reviews available to the engine for this phase.

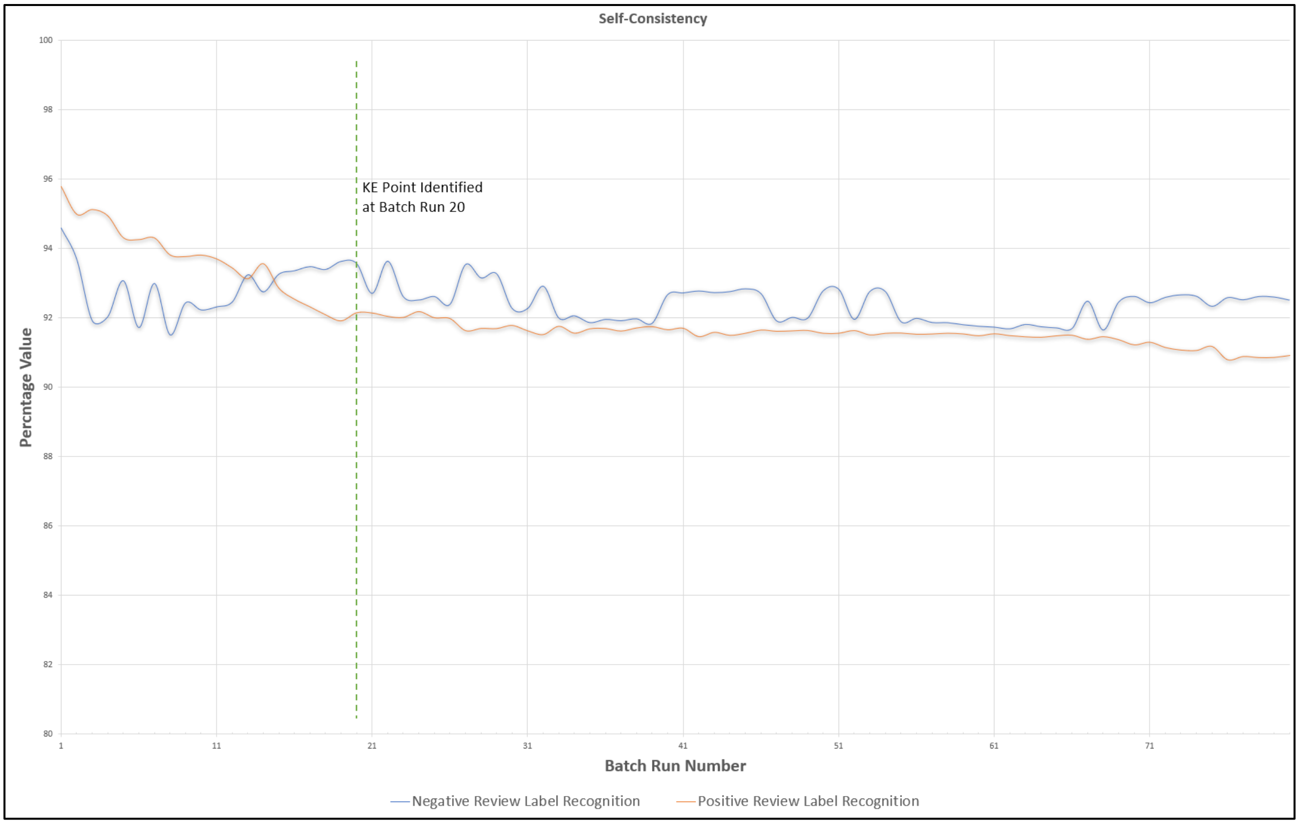

Figure 1 - Self-consistency graph showing the KE point identified at batch run 20 for the benchmark exercise.

The Education Phase was run on a non-proprietary laptop with core i5 7th Gen CPU, 24 GB of RAM, and no GPU.

The total time to complete the Education Phase was 1.2 seconds. CPU utilization was on average in the 60% to 63% range and memory utilization peaked at 143 MB.

3.2 Text Classification Phase

During the Text Classification Phase, the CRI™ engine used a total of 4,000 reviews to measure its accuracy. The measured accuracy was 91.93%.

The measured accuracy was 0.93% less than the maximum predicted accuracy calculated with self-consistency. The rationale for the difference is outside the scope of this benchmark exercise.

The Text Classification Phase was run on a non-proprietary laptop with core i5 7th Gen CPU, 24 GB of RAM, and no GPU.

The total time to complete the Text Classification Phase was 0.6 seconds. CPU utilization was on average in the 30% to 32% range and memory utilization peaked at 32 MB.

As expected, there is a difference in the accuracy of correctly labeled negative reviews and positive reviews. The difference is shown in Table 2. The difference is explained by the dataset bias towards negative reviews which contain more sentences within their review bodies as pointed out in subsection 2.1.

Label Generated by CRI Engine	Accuracy (%)
Positive	91.88
Negative	91.99

Table 2 – Difference in accuracy of CRI™ engine when generating review labels for positive and negative reviews.

Newness was identified in some of the reviews contained in the dataset used during the Text Classification Phase. As expected, the accuracy dropped for reviews with newness as compared to reviews without it. Table 3 shows the accuracy for reviews with newness and without it. The impact on accuracy due to newness is characterized by a reduction in accuracy up to 20.00% when compared to the accuracy without it. While optimizations in the generation of training data are possible to reduce the impact due to newness, optimization of the impact due to newness is outside the scope of this benchmark exercise.

Label Generated by CRI Engine	Accuracy (%)
No Newness (N=0)	91.93
With Newness (N>0)	74.28

Table 3 – Difference in accuracy of CRI™ engine when generating review labels due to newness.

4 Conclusion

The benchmark exercise took 1.8 seconds to complete, reaching a measured accuracy of 91.93%.

The benchmark used the English subset of the MARC dataset. The benchmark exercise had two distinct phases: the Education Phase and the Text Classification Phase. It took 1.2 seconds to complete the Education Phase and 0.6 seconds to complete the Text Classification Phase.

The Education Phase used only 25% of the training data to educate the CRI™ engine, resulting in a reduction of 75% less data needed to educate the engine as provided in the dataset.

The benchmark was run in full on a non-proprietary laptop with core i5 7th Gen CPU, 24 GB of RAM, and no GPU.

The benchmark exercise results demonstrate the non-neural network-based AI approaches to NLP tasks produce results that are better when compared to the results obtained via neural network-based AI models in the four focus areas selected. Table 4 summarizes the comparison.

Focus Area	Result CRI™	Result Neural Network-based AI
Accuracy	91.93%	52% to 73% (Keung et al., 2020)
Total Benchmark Elapsed Time	1.8 seconds	mBert training is unknown plus 10 hours for fine-tuning and evaluation (Keung et al., 2020)
Hardware	Regular consumer laptop	mBert: unknown, rest of work using AWS p3.8xlarge instance (Keung et al., 2020)
Data Required for Education/Training Phase	25%	100% (Keung et al., 2020)

Table 4 – Difference on focus area results between CRI™ AI technology and neural network-based AI technology.

The CRI™ engine is a non-neural network-based AI approach and the results captured in this benchmark exercise demonstrate that it is very well suited for NLP tasks such as text classification.

Since the MARC dataset also contained reviews in languages different than English, the same CRI™ engine was used to evaluate its accuracy with some of these languages. Table 5 shows the accuracy measured for English, Spanish, German and French, providing further evidence for the fitness of CRI™ engines to operate across multiple languages.

Language	Accuracy (%)
English	91.93
Spanish	90.64
French	90.10
German	90.78

Table 5 – CRI™ engine accuracy when generating review label measured for different languages included in the full MARC dataset.

5 Acknowlegments

We thank the creators of the MARC dataset for making their data publicly available for research purposes and the many Amazon user reviewers for creating the content.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics.

Philiph Keung, Yichao Lu, Gyorgy Szarvas and Noah A. Smith. 2020. The Multilingual Amazon Reviews Corpus. ArXiv, arXiv:2010.02573v1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326.