HOW DO YOU SEE THE FUTURE ?

Abstract

In today's era, research areas such as Big Data and Machine Learning are expanding and are using to solve problems like text or image classification.
Text classification, in particular, uses both fields of research to make correlations between information from news, blogs, social media platforms and other sources.
This paper focuses on methods for creating a database about future statements and finding correlations between topics and these statements.

1 Introduction

With the rise of big data and machine learning techniques, a number of large databases of different subjects have become increasingly important. They include datasets of news articles like Reuters news (Thoma, 2017) or question answering datasets like WikiQA (Yang et al., 2015).

However, to the best of our knowledge, there is no comprehensive database of statements about the future. Consequently, this article proposes the application of current methods of natural language processing to compose a database of statements about the future made online.

Furthermore the aim is to classify those statements into different categories and perform sentiment analysis on text passages referring to the future. The analysis of the sentiment of statements about the future development can be particularly relevant for e.g. capturing political sentiment. The overall questions to be answered are:

1. What do people predict for the future?
2. How do people see the future?
3. What are some common positive or negative predictions about different topics?

2 Related Work

To the best of our knowledge there is no database consisting of predictions/statements about the future.
There are rather different approaches to determine whether a sentence makes a prediction or not. For example, from (Jatowt et al., 2009), which is about extracting time-focused predictions, but with the aim of ranking for relevance to news articles. A different approach was taken from (Ozgur and Radev, 2009), who focused on scientific articles and were able to extract future-based statements using keywords representing conjectures.

However many text classification approaches have been described which can be utilized for our goal of future statement extraction from a large corpus and making our method robust (Minaee et al., 2021). Text classification can also be used for interpretation: labeling the extracted future statements and determining their topic.
Another technique to be utilized is sentiment analysis–also referred to as opinion mining. Frequent subjects of investigation are, for example, product reviews, forum discussions, blogs, Twitter or social media (Jagdale et al., 2019)(Agarwal et al., 2020) . The aforementioned subjects are particularly relevant for companies, i.e., the analyzes of opinions about products, product functionalities, or attitudes towards a company can support new product development or planning of marketing campaigns.

3 Data

In order to clarify and ultimately answer our core questions, we need various sources of data. For this purpose, the Internet Archive pipeline was used to extract the necessary data, which gives access to text data that is more than ten years old. Specifically the corpus-iwo-internet-archive-

wide00001 was used in our extraction process.

Therefore, the Webis research group provides us a large-scale high-performance compute infrastructure, totaling more than 3000 CPU cores, 10+ Petabytes of storage, and 24 high-end GPUs. This pipeline allows extracting data from WARC files on a CPU cluster and streaming it to a GPU server, where it is processed. This allows to quickly retrieve data (text or images) from the WARC files that gets classified as positive by a deep learning model. (Deckers, 2022)

In order to not only depend on this source and to get a kind of scattering of similar data, we programmed a web scraper that could provide us with further data from different sites. This is where the library "Beautiful Soup" came in, which makes it easier to extract data from HTML and XML. The scraped pages are as follows:

- 2050.earth
- futuretimeline.net

4 Pipeline

The entire process is represented by the pipeline in Figure 1 and is divided into six major steps which are explained further below.

4.1 Candidate extraction

The first step is to extract the necessary future-based statements from the unsorted data. At this, the regular expressions "in the future" or "in the future" are used to prepare all possible candidates for the further step of validation.
This step can still be optimised, as similar to the paper from (Ozgur and Radev, 2009) mentioned earlier, two keywords, especially with the keyword "future", are used to achieve our goal.

Nevertheless, this is the fastest and easiest way to get a large corpus of data that has a future context, in respect of extracting time. Due to time restrictions we were only able to extract data from around a quarter of the whole web archive corpus mentioned above. The whole line containing the regular expression was extracted.

4.2 Validation and manual labeling

This step requires the manual labeling of all sentences that do or do not have a future content. Moreover we checked if it is a valid future prediction or

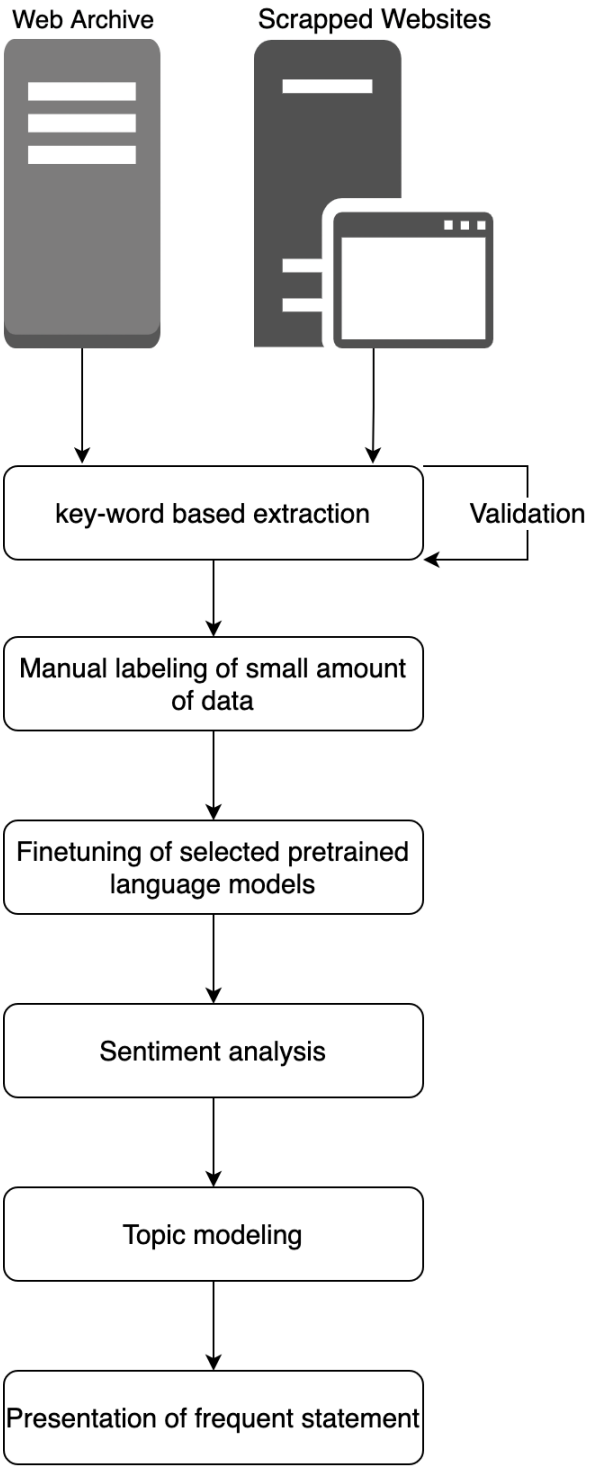

Figure 1: Pipeline

not. Checking whether it is a valid prediction at all is very important, as can be seen in Table 1. The first row shows a good example:

- "Absolutely, and you can help in the future too."

By itself, this sentence does not make a factual statement about the future, but it contains the regular expression "in the future". This is a good example of how keyword extraction reaches its limit. If a statement is taken out of context, the factual content is missing and it can no longer be considered as a future statement.

We found that further editing of the sentences, such as removing stop words or lemmatizing the sentences, did not improve the results, so as many sentences as possible were labeled. Finally only the actual sentence containing the keyword "future" was extracted from the whole line extracted in the first step as we have found accuracy to increase with that in the further steps.

Absolutely, and you can help in the future too.	0
Wind power is likely to play a large role in the future of sustainable, clean energy, but wide-scale adoption has remained elusive. Now, researchers have found wind farms' effects on local temperatures and proposed strategies for mediating those ...	1
The rest of the cast did a great job also. Eva Green was sultry and a credit to all Bond girls. Jeffrey Wright, though sorely under used did a magnificant job. Giancarlo was a revalation as Mathis, would be great to see him back in the future films in any way. Dame Judi as always was pleasure to watch, I thought the scenes she had with Daniel Craig had real chemistry. As for Mads Mikkleson, I thought he played a good part, but could have been far more threatening. As for the bit part players, they all did their jobs very admirably indeed.	0
And I intend to continue to do so in the future.	0
...	...

Table 1: Exemplary data extracted from the regex.

The strategy for manually tagging the data resulted in a simple two-column data structure, with an extracted sentence in the first column and the associated label in the second column, showed by table 1. The number "1" represents true and the number "0" represents false.

4.3 Fine-tuning of pre- trained language model

The manually labeled data was used for fine-tuning a pre trained transformer language model. We used DistilBERT (Sanh et al., 2019) for sequence classification and fit our training data to the model. We fit the model to our data for both whole lines and single sentences and found performance to increase with single sentences only.

Our training dataset consisted of 756 sentences, 474 of which were labeled 0, 130 of which were labeled with 1. They all came from manual labeling, from which a subset of data labeled with 0 was removed beforehand.
This was done because a random sample of the extracted data contained more than 10 times as many non valid statements as it did valid statements. The train, test, validation data split was chosen to be 80:10:10. Training was performed using the Adam optimizer and the sparse categorical accuracy in 5 epochs. Accuracy, precision and recall was calculated using the subset of data extracted from the training dataset.

4.4 Classification

In a next step the trained model was fit on the data extracted using regex. Sentences classified as a valid future statement were saved, the rest discarded.

4.5 Requirements

We use many different modules for our pipelines. Some of these modules were:

- Tensorflow: Tensorflow is an open source library used for machine learning in particular for the training and inference of deep neural networks. We use tensorflow to create a deep learning network that classify a statement as future predicting statement.

- Fastwarc: Fastwarc is a high performance library used to parse the WARC files. We use it to parse the internet archive data's WARC files.

- Transformers: Huggingface Transformers is an API used to download the pre-trained huggingface models. We used transformers' distilbert model to classify sentences as future predicting statements.

- Textblob: Textblob is a library for processing textual data. It is an api that can perform NLP tasks such as part of speech tagging, sentiment analysis, subjectivity analysis, classification, translation, etc. We use Textblob to get sentiment of the sentences and the subjectivity.

- Bertopic: BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. We used bertopic to find the topics discussed by the people on the internet.

- Re: Re also called Regex module is a python in-built module used to perform regular expressions on text. We use regex to obtain those texts that contained sentences containing some special phrases like "in the future" to obtain sentences predicting the future.

4.6 Sentiment analysis

In this section, we try to analyse the sentiment of the future predicting statements. Our motive is to observe how people on internet view the future to be. We aim to see if people on the internet view future as neutral, positive, negative or polarized.

This is done by plotting a histogram for the distribution of the sentiment of individual statements. To acquire the sentiment's intensity of these individual statements, we use Textblob library of python.
We obtain sentiment ranging from -1 to +1 where -1 means negative sentiment and +1 means positive sentiment. This is done by first calling the Textblob object with the sentence as the argument and then we call method called sentiment which gives us the sentiment of the sentence.

4.7 Subjectivity

In this section, we try to analyse if the people predicting the future talk more about facts or if they give more opinions. This is also done using Textblob library.
We obtain the TextBlob object of individual statement and we call this object's subjectivity method to obtain subjectivity of individual statements. After that we plot a histogram of the distribution of these individual subjectivities.
The value of subjectivity varies from 0 to 1 where 0 means more factual statements whereas 1 means more opinionated statements. To get this value, we call the Texblob object with the sentence as its argument and then call the method named subjectivity on this method which gives us the value of subjectivity.

4.8 Topic modeling

In this section, we try to find out what people on the internet are talking about. To do this, we are using a pre-trained library called BERTopic. This library uses huggingface and c-TF-IDF to create clusters that uses important words from the text to give us topics.

4.9 Presentation of frequent statements

In this section, we will look at some of the most frequently appearing sentences. Looking at them will also give us some indication about what people on internet are talking about. To do this, we grouped the sentences and counted their frequency using group by method of pandas dataframe.

5 Results

5.1 WARC Extraction

Extraction from about a quarter of the corpus used gave us roughly 325,000 lines containing "in the future|In the future". Most of these however can not be considered valid future statements.

5.2 Future Statement Classification model evaluation

Model evaluation from the validation data set yields an overall accuracy of 0.78. As seen in f1-score in 2 the accuracy for predicting a valid future statement (1) is much lower than false statement prediction.

Because of the imbalance of valid and invalid statements in above mentioned data, we measured performance on a separate hand labeled validataion data set of a hundred data points. Results for that can be seen in 3

However we also noticed that repeating the fine tuning process with just a different (because ran-

	precision	recall	f1-score	support
0	0.84	0.88	0.86	58
1	0.53	0.44	0.48	18
accuracy			0.78	76

Table 2: Precision, Recall, f1- Score and Support achieved from validation data set that came from a random train, test, validation split set. These were calculated on the trained model used for future statement classification in the following process.

	precision	recall	f1-score	support
0	0.97	0.81	0.88	94
1	0.18	0.76	0.29	6
accuracy			0.80	100

Table 3: Precision, Recall, f1- Score and Support achieved from a separate hand labeled validation data set on trained model for binary future statement. These were calculated on the trained model used for future statement classification in the following process.

dom) train, test, validation split changed these numbers on the separate hand labeled validation data set. Just to reiterate that, refer to 4. In the bottom case, with accuracy 0 the model didn't extract a single future statement from a subset of 5000 sentences of extracted data. It should be mentioned that increasing the number of epochs used for training did not change anything on the calculated accuracy.

	precision	recall	f1-score	support
0	0.97	0.98	0.97	94
1	0.60	0.50	0.55	6
accuracy			0.95	100
0	0.94	0.99	0.96	94
1	0.00	0.00	0.00	6
accuracy			0.93	100

Table 4: Precision, Recall, f1- Score and Support achieved from a separate hand labeled validation data set on trained model for binary future statement. The models these were calculated on was trained on the same dataset just with a different train, test, validation split for each and compared to the model we used for further processing.

5.3 Classification results

At the end of the classification of all candidates extracted from our regular expression pipeline, we are left with 49,035 datasets classified as future statement. However even a brief look at the data shows, that the accuracy evaluated when testing the model on our split of the manually labeled data, it becomes very apparent that accuracy isn't as high as estimated. Some examples of future sentences extracted from the pipeline are:

- This means Russia will increasingly act as a superpower rival to the USA and the West in the future
- "After all, in the future all buildings will be green"
- As deterrent doping authorities are threatening to freeze athletes blood and urine samples for re-testing in the future
- Are there any plans to put in a sewer system in the city at any point in the future
- "When a multi-layer system is adopted in the future, the storage capacity can increase to 100GB* and even 200GB*"
- "Just curious, are there any current plans for a sewer system? Has it been in the past or will it be discussed in the not to distant future?"

5.4 Sentiment Analysis

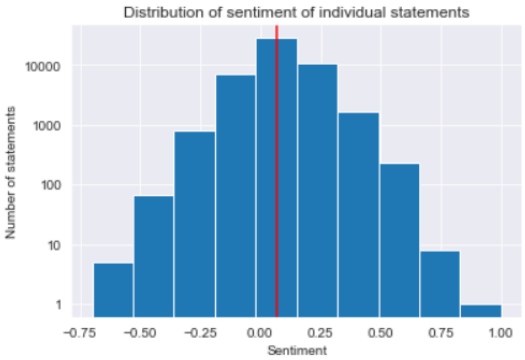

Figure 2: Histogram of sentiment of extracted statements. The total number of statements is plotted over the sentiment estimated by the model. Negative values represent a negative sentiment, while positive values represent positive sentiments. The red line indicates the average over all values.

From Figure 2, we see that the mean sentiment of the people on internet is around 0.07. This means that their views are neither positively nor negatively biased rather fairly neutral.

5.5 Subjectivity

From Figure 3, we see that the mean subjectivity turns out to be 0.33. This implies that people give

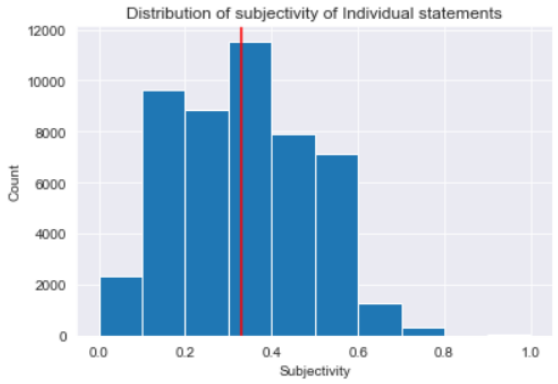

Figure 3: Histogram of subjectivity of our extracted statements. The total number of statements is depicted on the y axis. A measure for subjectivity is depicted on the x axis. A value of 1.0 represents a very subjective statement, a value of 0 a very objective statement.

more factual statements rather than giving opinions.

5.6 Topic Modeling

As we see in Figure 4, our most common topic is water followed by some interesting topics like elections, climate, genetics and trading. These topics give us an overview of what people on internet are talking about. We see that there are mainly topics that we expect from statements involving future predictions like elections, climate, genetics, trading, etc.

5.7 Presentation of frequent statements

As we see in Table 5, the most common sentences are not really talking about any particular topic. These sentences seems to be part of thread or footnote of some website. Hence, we cannot conclude anything from these sentences.

6 Discussion

6.1 Dataset evaluation

The process of manually labelling the statements proved to be a challenge. When do you tell if a statement says something about the future or not? The manual labeling requires a correct definition of a future statement. A good example of this is the following extracted statement:

- Dreaming of being an aicraft industry owner in the future? Haha!

It is indeed a statement about the future with a certain factual element. However, it does not in itself make a concrete statement about the future. Although it even contains an associated topic, in this case the aircraft industry.
Another difficult case to evaulate were conditional kind of statements. The kind that made a prediction about the future, but on a condition. For example:

- The Government needs to steer the banks towards lending policies that meet the needs of ordinary working people struggling to find their first home or we could be storing up huge problems for the UK labour market in the future

Technically a prediction about the future is made here.

Context proved to be another problem. As some of the future statements were not part of the sentence containing they keywords we extracted for. However complete lines did not necessarily improve data quality. Since a lot of the times just a small part of the whole line was a prediction. We chose to concentrate the rest of the tasks on individual sentences only because model performance seemed to be better in these cases.

To help lessen the chance of others having the issue in the future, please provide as much information as possible so the rule can be modified and adjusted	297
As such, we neither warrant the accuracy nor accept any liability or responsibility for inaccurate information other than to correct the error(s) in the future	143
A well known technical analyst has written, "If the market has shown respect in the past to a Fibonacci grid drawn on the chart, the chances are much higher that it will also respect those levels in the future market action	133
You will also pay all federal, state and local taxes, if any, levied now or in the future, that are applicable to your use or receipt of the Services	110
"Living in the past is dwelling upon what cannot be changed, living in the future is creating the milieu for fear and anxiety, living in the now is the right environment for action and change"	106

Table 5: Most Frequent Sentences

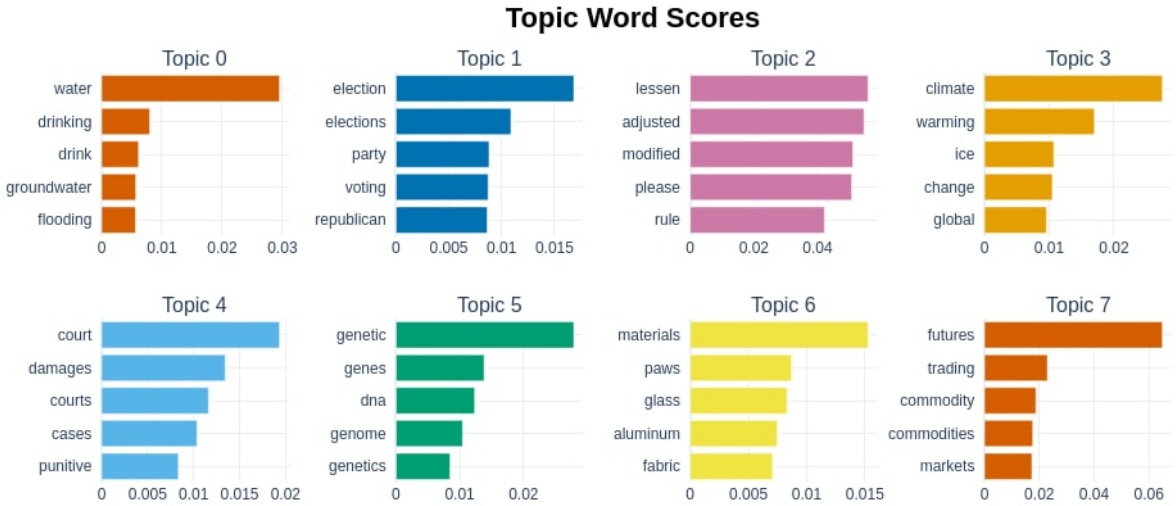

Figure 4: Most Common Topics

Manual labeling of the data was done on one of the files extracted on subsequent regex matches. This also means data that was labeled repeated a lot. Duplicate lines appearing even more than twice sometimes was among the labeled training data. It is likely that they came from similar or even identical pages. They could be part of a footer of a website or a signature of a user posting in a forum for example. In hindsight this deteriorated training data even further. A random set of data extracted from all of the extracted data should have been used for model training.
Moreover data extracted from the internet, especially from scrapped sites, are very heterogeneous, which makes it more difficult to prepare it for a natural language models. This fact results in a lot of repeated statements, not only in terms of sentence/statement structure, but in terms of content written in other words delivering the same meaning output.

The imbalance of valid and invalid sentences - judging by hand labeled data of several hundreds of lines - less than 10 % were valid future statements made training additionally difficult. A way to improve on this could be narrowing the scope of the regular expression, e.g. by requiring candidates to contain the word 'predict'. But as we already do by applying a regular expression in the first place this will cause loss of data.

Unfortunately the data scraped from both futuretimeline.net and 2050.earth did not improve results. Especially the data we were able to scrape from futuretimeline.net was much too homogeneous. All statements were short and precise and written in the same tense, much unlike data scraped from all over the internet.

6.2 Model Performance

Despite decent values in the evaluation step, the models predicted future statements are subjectively non valid predictions.
This is in part due to an imbalance in the dataset achieved by extraction. Sentences in it are predominantly non future statements.
So even if, as seen in 2 accuracy for predicting label 0 is relatively high, the model will evaluate a lot of false future statements as valid ones simply because there is a majority of false future statements present. It also reflects the models inability to predict true future statements accurately. Both precision and recall are much lower in evaluation for label 1 corresponding to valid future statements.

Another indicator for the models inability to correctly determine valid future statements is that repeating the training data just on a slight variation of training data by splitting labeled data randomly in a train, test, validation split caused the model to perform very differently on the manually labeled test set (see the difference in values of table 4 and 3). It is an indicator that not enough data was used for training. Or that data used was too similar.

Since overall accuracy for both labels is taken as a measure to train the model using raw labeled data without removing any of the 0 labeled data also

wasn't an option. Overall accuracy increases if a sentence is classified as invalid future statements in that case. Training a model on data like that, that nothing was classified as valid statement in the classification step. This can also be seen by the bottom part of 4. A recall and precision of 0 means that none of the future statements were identified correctly. However we should mention that a support of only 6 valid statements in the validation data set is not very meaningful overall with the range of possible predictions that could be made about the future. If it was, a precision of 0.18 3 would indicate that, from the extracted data in the final dataset, 18% should be a valid future statement.

7 Credits

Nikolai Kortenbruck: Conceptualization, Software, Formal analysis, Data Curation, Writing - Original Draft
Shivom Gupta: Conceptualization, Software, Formal analysis, Data Curation, Writing - Original Draft
Alexander Pavlovski: Software, Data curation,Writing - Original Draft , Formal analysis

References

Basant Agarwal, Richi Nayak, Namita Mittal, and Srikanta Patnaik. 2020. *Deep learning-based approaches for sentiment analysis*. Springer.

Niklas Deckers. 2022. web-archive-keras. https://github.com/webis-de/web-archive-keras.

Rajkumar S. Jagdale, Vishal S. Shirsat, and Sachin N. Deshmukh. 2019. Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing*, pages 639–647, Singapore. Springer Singapore.

Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama, and Katsumi Tanaka. 2009. Supporting analysis of future-related information in news archives and the web. pages 115–124.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Arzucan Ozgur and Dragomir Radev. 2009. Detecting speculations and their scopes in scientific text. pages 1398–1407.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Martin Thoma. 2017. The reuters dataset.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.