The Future Tense - Paper

Dünya Baradari **Finn Bartels** **Artur Dewald** **Julia Peters**

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

Human-like artificial intelligence (AI) has been exciting and frightening humanity since the antiquity. Often intertwined with the concept of an artificial man, humanoid automata with the supposed capacity to answer questions and feel emotions have been present among all civilizations, including the ancient Egyptians and Greek (Newquist, 1994), Chinese (Cohen, 1986) and Mesopotamians (Unat, 2008). Yet, it has been in the past decades that the rise of computing power according to Moore's Law has enabled a wide-scale application of AI technologies. At the time of writing, use cases range from self-driving cars, personalization of ads in online browsing to highly complex predication tasks for protein folding (Jumper et al., 2021).
This rapid development of *intelligent machines* in everyday life and application has led to both hopes and fears among the general population. Cave and Dihal (2019) identify four dichotomy categories of excitement and fears about artificial intelligence. These are immortality and inhumanity, ease and obsolescence, gratification and alienation and dominance and uprising (Table 1). They further argue that such perceptions, which may not align with reality, can yet influence the development, regulation, and applications of AI. The encouragement of research into AI ethics by various public policy groups and governments may be a reflection of this point (Leslie, 2019).

Dichotomy	Hope	Fear
Immortality and Inhumanity	Much longer lives	Losing one's identity
Ease and Obsolescence	Life free of work	Becoming redundant
Gratification and Alienation	AI can fulfill one's desires	Humans will become redundant to each other
Dominance and Uprising	AI offers power over others	AI will turn against humans

Table 1: Categories of dichotomies of hopes and fears towards AI. Based on Cave and Dihal (2019).

In our work, we seek to follow up on Cave and Dihal (2019) analysis and examine the views of the English-speaking online community regarding the future of artificial intelligence. We discern the most common clusters of topics that are formed around AI and the average sentiment for each topic using machine learning. To that end, we employ natural language processing (NLP) to extract and analyze statements about the future of AI from the Web Archive (Deckers, 2022), a collection of website snapshots which offers us data from the past 10 years. We are applying three models on this data. The first model is our finetuned future model, which is able to recognize statements about the future. Subsequently, this data is fed into an existent sentiment classifier to add sentiments. Finally a topic is assigned to every sentence by the last model. This is followed by an analysis of the individual topic clusters. While our analysis solely concerns artificial intelligence, our pipeline and models offer a way to study online views concerning the future for any topic. By examining the prevalence and sentiment of AI topics specifically, we hope to inform social science researchers, philosophers, and policy makers about the development of artificial intelligence in the general population's perception, to direct efforts towards a better future with AI.

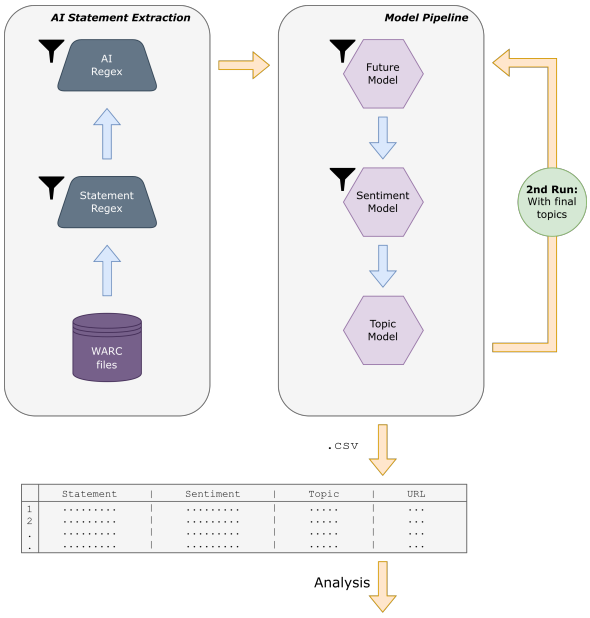

Figure 1: The statement extraction, filter and final topic assignment process. AI statements are extracted from website HTMLs located inside WARC files (left part). The extraction process involves the application of two regexes (statement regex and AI regex). Statements then enter the model pipeline (left side), where they are further filtered through the future model and the sentiment model. The model pipeline runs two times in total. The first time the topic model assigns dummy topics. After the topic selection process (green circle), where we fine-tune the topic model, the model pipeline runs a second time. Now the topic model assigns the final topics to the statements. The output is a `.csv` file with the schema `statement|sentiment|topic|url`.

2 Methodology

For the realization of our concept, the following three objectives have to be accomplished:

1. Obtaining of a sufficiently large data set with different expressions about the topic of AI.
2. Raw data transformation into a data set that to the target schema illustrated in Table 2
3. Creation of a visualization from which society's perceptions on different topics of AI can be extracted.

1. WARC Data Extraction:
Initially, we outline the raw data acquiring containing AI expressions. Since the web archive provided by the Webis group with the corresponding WARC-DL data extraction pipeline (Deckers, 2022) is at our disposal, we utilize those. The data set consists of long texts. For that reason, the text must be splitted into separate sentences. Then, sentences about AI can be extracted for later processing by applying Regex.

2. Data Transformation:
For later analysis, the data must be converted to required target schema, illustrated in Table 2. Therefore we employ the model pipeline, which generates the final data set with the attributes future statement, sentiment, topic and url. Within this model pipeline three models are applied sequentially:

- Future statements model:
  This model is able to distinguish between statements about the future and all other types of terms. On the basis of its classifications, statements about the future are extracted.
- Sentiment model:
  A sentiment is assigned to every future statement by this model
- Topic model:
  Before our selected topic model enables the assignment of reasonable topics, those must be provided as label candidates to this topic model. Accordingly, the Model Pipeline runs twice. In the first execution the chosen model runs with dummy topics. Based on the output the topic selection is conducted. For each subsequent run, the model pipeline is considered complete.

The URLs of the pages are also included in the data schema for a later discussion. This pipeline generates the final data set with the attributes future statement, sentiment, topic and URL. The resulting data set can then be used for analysis and interpretation.

3. Analysis:
For the analysis we start with a graphical visualization. Therefore, we decided to group all statements according to their topics. This way for every topic a sentiment analysis can be conducted separately.

In the following sections, each step of the implementation is discussed in detail.

2.1 WARC Data Extraction

The Webis Group[1] had provided us with access to 37908 WARC files, with a total amount

[1] https://webis.de

statement	sentiment	topic	url
AI can be a risk for many workers.	NEG	finances	... ...
AI will definitely revolutionize games!	POS	gaming	... ...
...	...	...	...

Table 2: Data schema for visualization and analysis

of XXXXXXTB of website data, and a high-performance computer cluster where we were able to schedule jobs. Additionally, we decided to utilize the WARC-DL pipeline (Deckers, 2022), a Python software pipeline tightly coupled with the WARC endpoint, and using the FastWARC[2] library under the hood to iterate over WARC records. It enabled us to extract text automatically from the WARC records using several customizable filters. We made some slight modifications to the source code, in order to make it fit our needs better.

Since we were going to analyze AI statements, we did not need all of the text provided by a website html, but grammatically correct statements of the English language. It is not possible to extract all statements with an accuracy of 100% using regex only, so we narrowed it down to passages which started with a capital letter and ended with a period or exclamation mark. We built a regex pattern which extracted a list of all valid statements from a html source.

Furthermore, we were only interested in statements about AI, so we compiled a list of AI keywords beforehand and applied a second regex on the extracted statements. We compiled a blacklist which contained keywords for filtering out hostnames related to pornographic websites, and a whitelist to only accept top level domains which were most common, and where the websites most likely contained English text. This was necessary, since initial runs of the WARC-DL pipeline extracted lots of unusable content (including content in languages other than English).

We ran into some problems with the cluster and WARC-DL pipeline, so we could not extract all statements in one job run. Jobs would suddenly stop extraction because of connection errors or out-of-memory errors, and halt with an Exception. We separated hostnames in four groups depending on the initial hostname character: `a-h`, `i-p`, `q-x` and `yz0-9`. This way we could pin down the problematic websites more precisely. In the end we managed to work through all WARC files in group `i-p` and `yz0-9`, and most of the WARC files in groups `a-h` and `q-x`. The final yield of the WARC data extraction stage was a total amount of 222246 AI statements. In the next steps, starting with the model pipeline, our objective was to further refine this initial data set.

[2] https://resiliparse.chatnoir.eu/en/stable/man/fastwarc.html

2.2 Model Pipeline

The model pipeline follows the WARC data extraction step and is designed to prepare our final data set, which consists of the future statements and their associated sentiment and topic labels. The processing within the model pipeline is performed on batches of 30 records each from the WARC-DL output.

First the future model filters out future statements from the corresponding batch. Subsequently, the chosen sentiment model assigns a sentiment to each future statement. In this step some future terms can be sorted out. This concerns the statements to which a sentiment is classified with a probability of less than 70%. The remaining future statements receive a topic. Finally those are persisted in a csv file.

In the following sections 2.3 - 2.5 we will go into detail about each individual model. In this context, we describe how the future model was trained and justify our decision for the sentiment and the topic model selection. Furthermore we outline the choice of our topics and explain, why only those statements are kept which a sentiment with a probability above 70% can be attributed.

2.3 Future Model

Since this paper focuses on analyzing statements about the future, a system for distinguishing between future statements and other expressions is required. In this context, we decided to finetune the DistilBERT (Sanh et al., 2019) base model that accomplishes this task. Therefore, in this subsection, the collection of appropriate training data and the subsequent finetuning of the corresponding model is thematized.

2.3.1 Training Data Set

In order to provide a suitable data set to establish the future model, we adopted multiple approaches. At this point, our goal was to compose the data in such a way that we would have a balanced data

set with two classes. The first class should contain future statements and the second all other types of terms. While two of our group members manually annotated 500 observations each, the other two used an automated mechanism with subsequent verification of the collected data.

One of the automatized approaches involves a web crawler developed on the basis of the python library Beautiful Soup (Richardson, 2022). The text on a page is divided into sentences. Subsequently every sentence is examined for occurrence of certain terms, as *going to*, *will*, *won't* or *'ll*.

The second automated approach is the sentence extraction tool, which works in several aspects, similar to the web crawler. At the beginning, it searches the given directory for text files. If those exist the text is split into sentences and observed for specific expressions, as described above.

To find the phrases that are not future statements, both the web crawler and the sentence extraction tool look only at the corresponding records that do not contain the previously considered expressions. A careful manual review of all terms gathered by the automated systems was subsequently performed to remove the incorrect records.

Finally, we constructed a data set with 1250 future statements and 1250 other phrases that did not contain future statements.

2.3.2 Training

As previously described we used the DistilBERT base model and finetuned it with the data set specified in 2.3.1. We split the data set of 2500 records into a training and a test set, where the test set contains 20% of the records. From the training set we split further 20% for validation data.

After only two epochs the training ended with an accuracy over 96% as displayed in (Table 4).

Subsequently we tested the model on our test set containing 500 records never seen by the model and achieved an accuracy of 93.8%, as seen in the confusion matrix in (Table 5).

2.4 Sentiment Model

In order to assign sentiments to future statements for later analysis, we decided to select a ready-trained model. The chosen sentiment model is the SentimentAnalyzer of the open-source library pysentimiento (Pérez and Luque, 2021), which was further trained on about 40k tweets. It uses the BERTweet (Nguyen et al., 2020) as a base model, pre-trained on english tweets.

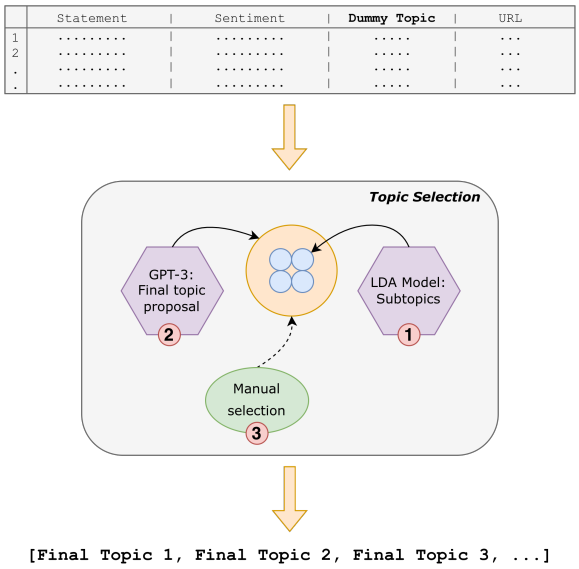

Figure 2: The three steps of topic selection. First, the LDA Model generates a set of subtopics for a topic cluster. In this example, there are four subtopics (small blue circles). Then, GPT-3 proposes a possible general topic for the subtopic set (large yellow circle). Lastly, we either pick this final topic as proposed, or replace it with a more suitable one from the same category, e.g. `level` would become `gaming`.

2.4.1 Evaluation

To evaluate the SentimentAnalyzer, we annotated 604 future statements, which were previously used for training the future model, as negative, positive or neutral and received an accuracy of about 65%. We then analyzed all misclassified statements. We noticed that some of the statements could not be assigned impartially to one of the three categories. An Example is *"AI will reinvent how we think about education"*. In the case of the sentence, we disagreed on whether we should value the sentence as neutral or positive and decided to use the neutral label. Subsequently, this statement was given a positive rating by the model. On closer examination of the statements that were labeled differently by us and by the model, we found over 90% of the labels given by the model to be valid, if these annotations were assigned with probability over 70%. For this reason, we decided to keep only statements about the future if the sentiment model assigned an annotation with a confidence above 70%.

2.5 Topic Model

For assigning topics to the future statements we employ the bart-large-mnli model from Facebook, which was pretrained on the MultiNli (Williams

et al., 2018) data set, which consists of 433000 pairs of sentences annotated with textual supplementary information. The bart-large-mnli is a natural language processing model based on the technique of Yin et al. (2019) utilizing pre-trained NLI models as ready-to-use zero-shot sequence classifiers. The approach involves specifying the sequence for classification as an NLI prerequisite and then constructing a hypothesis of every possible label candidate. Afterwards probabilities of agreement and contradictions are transformed into annotation probabilities. Before we are able to use the bart-large-mnli it is necessary to define the topics, which can be assigned to every statement by this model. For instance, to verify if a sentence is a political or a technological statement, we can provide the model with the label candidats politics and technology. Then the model will apply one of the labels to the sentence. In the following section our topic selection approach is described in detail.

2.5.1 Topic Discovery with LDA

For analyzing the overarching topics within our future statements, we used Latent Dirichlet Allocation (LDA). LDA is a subtype of the Dirichlet Process Mixture Models (DPMMs), a set of non-parametric, “fully-Bayesian” unsupervised clustering models which are commonly used for topic cluster analysis. DPMMs use a stochastic process to generalize the Dirichlet distribution (the conjugate prior for a categorical or multinomial distribution) for infinitely many categories (Li et al., 2019). Applied to NLP, a Latent Dirichlet Allocation model clusters observations into unobserved groups of related data. It has the advantage of following a generative process that is immune to overfitting with increasing size of the data corpus and can be scaled to a data cluster in machine learning (Pritchard et al., 2000)

We prepared our data set for LDA by removing punctuation, words with less than three letters and stopwords. The cleaned data was then converted into lowercase and fed into a tokenizer. From there on, we created bigrams (sets of two words) from the tokens and employed a Word2Vec model to select only the most occuring ones. Finally, bigrams that occurred in more than XX% and less than XX of the documents were filtered out. The remaining bigrams were inputted into an LDA model, returning clusters of related topics. To label each cluster with a matching headline or cluster name, we used OpenAI’s GPT-3 (text-davinci-002) to turn suggestions of cluster names, from which we selected the best-fitting ones if possible. For some sets of words, the suggested titles could provide only inspiration and had to be adapted.

In the following set, the most frequently occurring words for one of our statement clusters are provided:

"autopilot, translation, machine, money, business, site, search, article, engine, traffic, page, list, marketing, internet, everything, income, link, cash, help, button"

The headings generated by GPT3 for this set of keywords were

Social Media Marketing, *Search Engine Optimization*, *Internet marketing*. We found that these keywords are covering two topics. Therefore we chose the categories *search engine* and*finance*

A further group of statements was represented by the keywords:

"intelligence, year, human, machine, technology, mind, development, robotics, position, form, reason, action, release, campaign, world, nanotechnology, singularity, view, look, life"

GPT3 offered *transhumanism* as a suitable headline. Correspondingly, we have incorporated this as a category for our topic model.

In conclusion, with this approach we received the following headings for our Topic Model: search engine, finance, transhumanism, machine human interface, social media, search engine natural language technologies

2.6 Analysis

For the purpose of analyzing the previously created data set, we oriented on the attributes topic, subtopic, network and sentiment of each tuple. While the future data set only captures the topic and sentiment attributes, we integrated the subtopics from Topic Model, that lead to the topics, into the analysis. This means that, besides the topic, we added a subtopic to each statement to be able to have deeper insights on the sentiments by subcategories of topics. This was implemented by extracting and combining the original subtopic list of each topic from the LDA model into one list of subtopics to find a subtopic for almost all statements. Here it was checked if a statement contains a subtopic from the list. The first subtopic found, was chosen. To get the most

meaningful allocation of subtopics to statements, we maintained the order of subtopics within each original subtopics list. This means, that the more meaningful subtopics found by the LDA model, were prioritized. Statements including none of the subtopics were labeled as 'undefined'. With these steps taken, we were able to generate the average sentiment scores for each topic and subtopic.

To complement the analysis of subtopics, we added a network column to the data set. This contains a list of all subtopics within the corresponding tuple of a statement. For the implementation of a network of subtopics, each node represents a subtopic. When implementing edges between these nodes, a tuple needs at least two subtopics in its network attribute to create an edge between them. Using this method, we were able to implement a network with up to 182 nodes.

3 Results

Looking on the distribution of sentiments, it appears that the majority of statements were commented as neutral statements (69%). The proportion of positive annotated statements (21%) is about twice that of negative annotated statements (11%). This shows that overall there is a slight tendency toward a positive attitude on the future of AI (**??**). In figure **??** it can be seen that neutral statements dominate each of the 9 topics. With two exceptions, Gaming and Machine Human Interface, there are visibly more positive than negative statements on each topic.

When analyzing the topics, we find that the statements are not equally distributed among all topics. While we divided all statements into 9 topics, Machine Human Interface describes about half of all statements (48%). Gaming as well as Natural Language Technology account for about 15% of all statements (**??**).

In the distribution of subtopics we can see a dominance of some subtopics too. The subtopic Data is associated with 21% of all statements, as well as Autopilot. Other dominant subtopics are Intelligence (19%), Recognition (12%), Computer (8%) and Supercomputer (7%) (**??**).

The average sentiment of all statements is at 0.1. This means a slight tendency to positive sentiment. The average sentiment of most of the 9 topics is majorly neutral. The topics of Transhumanism, Natural Language Technology, and Research Computing have the most positive sentiment on average. The most negative sentiment on average can be seen at Gaming and Search Engine. 3 of the 5 most common subtopics of Gaming have a sentiment score of less than 0 (**??**).

4 Discussion

4.1 Website Examination

The final data set, produced by the model pipeline, contains the url for every AI future statement. Table 3 presents the domains that are mostly occurring in this final data set. When examining the main domains, these appear relatively diversified. A website dealing with phylosophical questions on the topic of AI is included (lesswrong.com). Following this, there are three sites from the field of gaming (acceleratingfuture.com, mugenguild.com, slightlymagic.net). Also, the blog of the department of defense is contained among these domains dealing with the research of defence and military needs (dodsbir.net). A store with speech recognition devices is also available (knowbrainer.com). Nevertheless a number of scientific blogs on AI-related topics are also include, which are lead by researchers or from the tech industry. Latter are mostly data scientists. Considering the other domains, many scientific websites as well as websites about gaming are also very abundant. Thus, rather the future statements were expressed by people from AI related fields. This could mean that this topic has a lower role in the general population and thus it is dealt very little with AI-specific topics in the public society. New discoveries could be made by observing domains containing statements from the last few months were used. More people might feel affected by the latest developments in this area. Consequently, there could be more blogs with people from other sectors who would exchange opinions about these developments.

4.2 Project Limitations

Since we had a limited time for this project, there are some aspects where we would have liked to continue our work. From a technical point of view, we would have preferred to spend additional time on labelling more data for the sentiment model. Thus, it could have been possible to fine-tune this

AI statements	Website	Description
210	lesswrong.com	Philosophical blog about AI developments
198	arcengames.com	Page of an indie game developer
182	acceleratingfuture.com	Blog about perspectives and emerging technologies
156	heatonresearch.com	Blog of a data scientist
106	dodsbir.net	Research blog of the department of defense
76	kdnuggets.com	Blog of data scientists for analytics and machine learning
71	knowbrainer.com	Shop containing speech recognition devices
58	mugenguild.com	2D fighting game
52	aidreams.co.uk	Robotics and AI blog
51	slightlymagic.net	Rules Engine for the game "Magic: the Gathering"

Table 3: Top Domains

model as well. With our current approach, we only keep the AI future predictions if the sentiment model makes a prediction with a certainty of more than 70%. This results in the loss of a few additional statements that we would have available for analysis.

Unfortunately, the location containing the corresponding date on the website does not contain the corresponding date is not consistent. Accordingly, we would have needed more time for the date extraction. Providing a year for each statement could illustrate how the perception of a certain topic in the field of AI has changed over time. Having insights about such trends, allows monitoring the developments in cultural perceptions over time periods.

5 Conclusion

References

S. Cave and K. Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2):74–78.

M. Cohen. 1986. Joseph needham, science and civilisation in china. *Arts Asiatiques*, 41(1):133–134.

N. Deckers. 2022. Warc-dl.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

D. Leslie. 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

Y. Li, E. Schofield, and M. Gönen. 2019. A tutorial on dirichlet process mixture modeling. *Journal of mathematical psychology*, 91:128–144.

H. Newquist. 1994. The brain makers, genius, ego, and greed in the quest for machines that think.

D. Nguyen, T. Vu, and A. Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

J. Pritchard, M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

J. Pérez, J anf Giudici and F. Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

L. Richardson. 2022. Beautiful soup.

V. Sanh, L. Debut, J. Chaumond, and W. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Y. Unat. 2008. Overview on al-jazari and his mechanical devices. *Published on: 25th February*.

A. Williams, N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

W. Yin, J. Hay, and D. Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

A Data Set Card: Future Statements

The english language data set contains 2500 statements. 50% of the relate to future events and 50% of which relate to non-future events. The statements were collected manually and programmatically from several websites and datasets. The sole purpose of this data set was to fine tune the distilbert-base-uncased model into our distilbert-base-future model. The data set is available on huggingface (https://huggingface.co/datasets/fidsinn/future-statements).

Composition

The instances are represented by single-/ or multi-sentence statements from following sources (unequally distributed):

- http://www.kaggle.com/unitednations/un-general-debates
- http://data.world/ian/united-nations-general-debate-corpus
- http://gadebate.un.org/
- http://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y
- http://www.wsj.com/
- http://www.vox.com/
- http://futechblog.com/
- http://www.weforum.org/
- http://wired.com/
- http://openai.com/blog/
- http://techcrunch.com/
- http://futurism.com

Annotation

- 0: No future statement
- 1: future statement

A Model Card: Future Statement Model

This model is a finetuned on 2500 expressions, which contained 1250 future statements. distilbert-base-uncased serves as a base model

Model Description

- Huggingface name: distilbert-base-future
- Creation Date: 11/08/22
- Version: 1.0
- model type: text classification

Intended Use & Limitations

- The primary intended use is the classification of input into a future or non-future sentence/statement.
- The model is primarily intended to be used by researchers to filter or label a large number of sentences according to the grammatical tense of the input.

Hyperparameters

The following hyperparameters were used during training

- optimizer: name: Adam, learning_rate: 5e-05, decay: 0.0, beta_1: 0.9, beta_2: 0.999, epsilon: 1e-07, amsgrad: False
- training_precision: float32

Training Results

For finetuning, we have 80% of of records from our self-annotated future-tatements dataset. This corresponds to 2000 records. The remaining 500 will be used to test the final distilbert-base-future model

Epoch	Train Loss	Train Accuracy	Val. Loss	Val. Accuracy
0	0.3816	0.8594	0.1547	0.9475
1	0.1142	0.9613	0.1272	0.9625

Table 4: Training Results

Framework versions

- Transformers 4.18.0
- Tensorflow 2.8.0
- Tokenizers 0.12.1

Test Set Results

		Classified as	
		f	**nf**
True	**f**	253	4
	nf	4	239

Table 5: This confusion matrix displays the number of true future (f) and non-future (nf) statements in contrast to statements which were classified as such.