Perceptions and Sentiments Towards the Future of AI - Deep Learning on Internet Archive Data

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 (**?**), Chinese (**?**) and Mesopotamians (**?**). Yet, it has been in the past decades that the rise of computing power according to Moore's Law[1] 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 (**?**).

This rapid development of *intelligent machines* in everyday life and application has led to both hopes and fears among the general population. **?** 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 **??**). They further argue that such perceptions, which may not align with reality, can yet influence the development, regulation, and application of AI. The encouragement of research into AI ethics by various public policy groups and governments may be a reflection of this point (**?**).

[1] https://www.britannica.com/technology/Moores-law

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 **?**.

In our work, we seek to follow up on the analysis of **?** 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 (**?**), a collection of website snapshots which offers us data from the past 10 years. We are applying a pipeline that consists of 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.

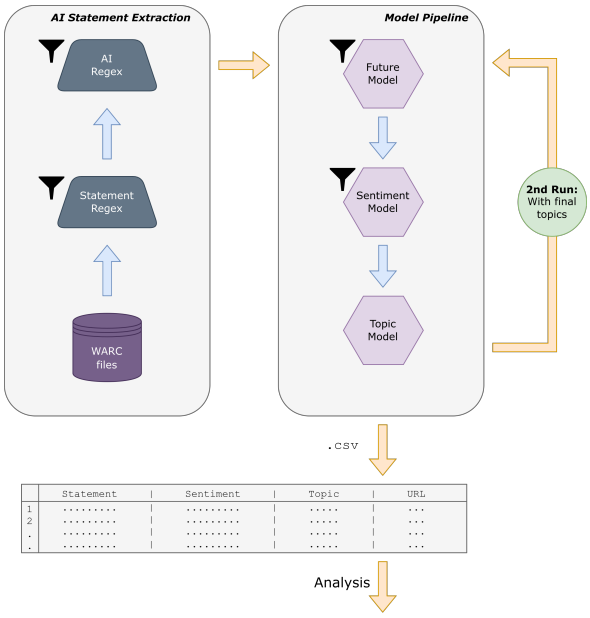

Figure 1: The statement extraction, filter and final topic assignment process. AI statements are extracted from website HTMLs located inside WARC files (left part). The extraction process involves the application of two regexes (statement regex and AI regex). Statements then enter the model pipeline (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 a sufficiently large data set with different expressions about the topic of AI.
2. Raw data transformation into a data set that matches the target schema illustrated in **??**.
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 acquisition containing AI expressions. Since a web archive with the corresponding WARC-DL data extraction pipeline (**?**) is at our disposal, we utilize both. Sentences about AI can be extracted for later processing by applying Regex on every sentence of the data.

2. Data Transformation:
For later analysis, the data must be converted to the required target schema, illustrated in **??**. 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[2] had granted us with access to 37,908 WARC files, with a total amount of XXXXXXTB of website data, and a high-performance computer cluster where we were able

[2] https://webis.de

to schedule jobs. Additionally, we decided to utilize the WARC-DL pipeline (**?**), a Python software pipeline tightly coupled with the WARC endpoint, and using the FastWARC[3] library under the hood for iterating 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 an 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 222,246 AI statements. In the next steps, starting with the model pipeline, our objective was to further refine this initial data set.

[3] 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 **??** - **??** 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 (**?**) 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 (**?**). 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 1,250 future statements and 1,250 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 **??**. We split the data set of 2,500 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 **??**.
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 **??**.

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 (**?**), which was further trained on about 40,000 tweets. It uses the BERTweet (**?**) as a base model, pre-trained on english tweets.

2.4.1 Evaluation

To evaluate the SentimentAnalyzer, we annotated 604 future statements, 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 and noticed that some of them 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 this 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%.

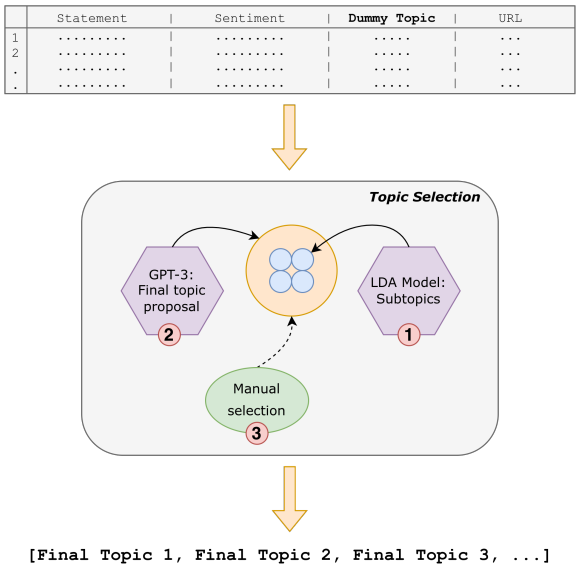

Figure 2: The three steps of topic selection. First, the LDA Model generates a set of subtopics for a topic cluster. In this example, there are four subtopics (small blue circles). Then, GPT-3 proposes a possible general topic for the subtopic set (large yellow circle). Lastly, we either pick this final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

2.5 Topic Model

For assigning topics to the future statements we employ the bart-large-mnli model from Facebook, which was pretrained on the MultiNLI (**?**) data set, which consists of 433,000 pairs of sentences annotated with textual supplementary information. The bart-large-mnli is a natural language processing model based on the technique of **?** 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 candidates: *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 (**?**). 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 (**?**)

We preprocessed our data set for LDA. 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 60% statements and less than 20 of the documents were filtered out. The remaining bigram candidates were fed 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.
Inspired by these suggestions, we created the topics that we found to be the best fit for the set of tokens of every cluster. 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 analysis of the previously created data set in **??**, we focused on the attributes topic, subtopic, network, and sentiment of each tuple. Since the future dataset generated by the Model Pipeline only captures the topic or category of a future statement and the sentiment attribute, we included further subtopics. To accomplish this, we proceeded as described in section **??**. Consequently in the resulting data set for every statement not only a topic but also a subtopic was contained. This way we intended to analyze such a future statement category in more detail. To accomplish this we extracted the original subcategory list of each topic utilizing the LDA model and combined these lists to a single one. We aimed to find a subtopic for almost every statement. For this reason we checked whether a statement contained a subtopic from the list. The first one found was selected. To obtain the most suitable assignment of subtopics to statements, we maintained the order of these subcategories within each original list. In conclusion the more meaningful statement subtopics found by the LDA model were prioritized. Sentences that did not contain any of this subtopics were marked as 'undefined'. This way, we were able to obtain the average sentiment scores for each theme and subcategory.

Furthermore, we added to every statement a number, mapped to every statement for a later sentiment score calculation. To negatively perceived sentences we assigned a `-1`, while positively labeled terms receive a `1`. Finally, every statement containing a neutral sentiment obtains a `0`.

3 Results

Examining the distribution of sentiments, a definite majority of neutral statements (69%) is clearly obvious . The proportion of positive annotated statements (21%) is about twice as large as the number of negatively annotated terms (11%). This implies a slight tendency to an overall positive attitude towards the future of AI (**??**). In figure **??** a domination of neutral statements is illustrated for each of the 9 topics. With two exceptions, Gaming and Machine Human Interface, there are visibly more positive than negative statements on each topic.

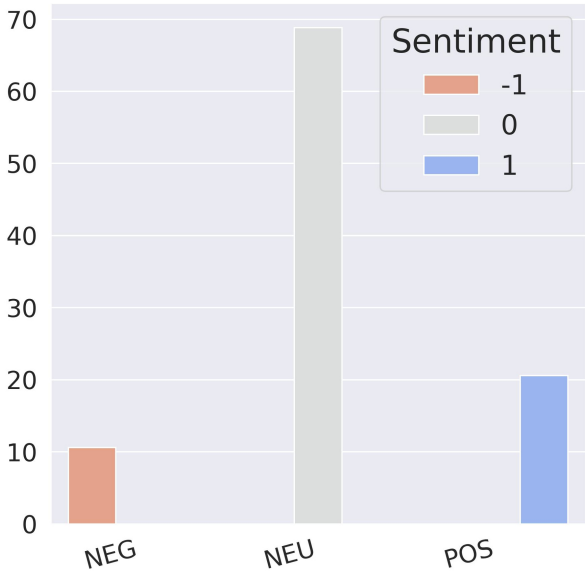

Figure 3: Dummy caption.

When analyzing the topics, we find the statements not being equally distributed among all these categories. While we divided all statements into the 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 observe a similar dominance of some subcategories. For instance, the subcategory 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%) (**??**).

As previously described we added for sentiment of every statement. Calculating the average the sentiment score is about 0.1. This shows a slightly positive tendency. The average sentiment of the most topics is majorly neutral. The categories containing the most positive rated sentiments are Transhumanism, Natural Language Technology, and Research Computing. The most negative sentiment on average were assigned to the categories 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 Results

Our results reveal a slightly positive outlook of people on future of AI. Interestingly, two thirds of our filtered statements were neutral, which, together with the nature of our top domain sources, suggests a large proportion of texts coming from academic or unbiased texts rather than strong opinions. The presence of about twice as many positively compared to negatively sentences and thereby a rather hopeful outlook into the future, may also be influenced by the optimism bias, which is when humans tend to overestimate the likelihood of positive events and underestimate that of negative (**?**). These findings also align with a large-scale study made by Google researchers on the public opinion regarding the long-term impact of AI on society (Kelley et al., 2021). This study, which surveyed people from eight English and non-English speaking countries, shows a clear dominance of neutral perceptions in English-speaking countries (USA: 53%, Australia: 57%, Canada: 56%) and a greater proportion of positive sentiments (USA: 21%, Australia: 18%, Canada: 20%) than negative (USA: 17%, Australia: 14%, Canada: 15%). Intriguingly, respondents from emerging economies such as Brazil, India and Nigeria exhibit far more optimistic opinions on the future of AI (BR: 38%, IN: 51%, NI: 37%), a finding that has been repeatedly confirmed in a recent survey by the World Economic Forum (**?**). This difference may be due to these countries' younger and generally more optimistic population, which may view AI as an essential opportunity for leapfrogging (**?**). While we filtered our statements for English-speaking websites only, the prevalence of English as the world's lingua franca must have led to the inclusion of opinions and statements from non-English speaking people. This may explain the slight skew towards positive sentiments in our data compared to sentiments from the USA, Australia and Canada from **?**. Comparing our topic clusters with the dichotomies of hopes and fears by **?**, we can only find a weak overlap between the data. Machine human interface and transhumanism well match the authors' first category of Immortality and Inhumanity. Interestingly, all subtopics within these clusters were labelled either neutral or positive, with those from transhumanism even displaying the most positive sentiments. The only subtopic receiving a negative label when compared to the

overall sentiment mean of 0.1 (Appendix XXX) is autopilot from the machine human interface cluster, suggesting the main concern to be losing conscious control by the use of such interfaces in the future. From this, we conclude a disproportionate number of opinions in favor of transhumanism and a machine-enabled future in relation to opinions highlighting the possible (existential) risks and dangers to our “human” identity. The Natural language technology cluster in turn may point towards applications of such technologies that make life easier, such as Amazon’s Alexa, thereby fitting the category of Ease and possibly Gratification. Yet, the corresponding counterparts in the dichotomies are hard to find in our clusters, where the only negatively labelled subtopics exist in the gaming cluster. This aspect can be explained by the increasing prevalence of AI-mediated systems in games, from non-playable characters (which can include enemies, a highly negative subtopic) to iterative game improvement and graphical enhancement (**?**). While the latter two examples provide advantages for the gamer, it may be that the idea of an AI as the enemy predominates since this is the primary noticeable direct touchpoint between the user and an AI. Furthermore, there are no clusters that we consider well-fitting for **?** category Dominance and Uprising, which, curiously, is a category commonly discussed by popular figures such as the late Stephen Hawking (**?**). Overall, it rather that our clusters center around applications of artificial intelligence. This is further demonstrated by the fact that several subtopics repeat within clusters (Figure XXX). For instance, “autopilot” is present within the machine human inference, natural language processing, finance, search engine, social media, and transhumanism clusters. Other repeated subtopics include data, supercomputer, computer, intelligence, recognition, and machine. Together with the slightly positive but largely neutral overall sentiment, which suggests that people speak mainly rationally about the future of AI, we reason that discussions center primarily around areas of application of AI technology instead of opinionated positions about its benefits and risks.

4.2 Topic Assignment

As described in section **??** the most statements are assigned with the "machine human interface" topic. This is not surprising, since technical sources clearly dominate our websites. Technological systems often require human operation and contain a corresponding interface as an implicit feature. Also search engines often need the handling from the outside. For this reason, many of these search engine related phrases end up in this category instead of in the search engine topic. Since search engines also correspond to natural language systems, they are often assigned to the category "natural language technology". To avoid this, more distinct generic terms for topics should have been chosen. This issue also arises in the field of "human machine interface". Statements that describe a friendly AI are grouped in this category . We would expect this kind of sentences in "transhumanism". The reason for this could be the categpry name, since it combines the terms human and machine.
"Computer vision", "natural language technology", "social media", "transhumanism" and "human machine interface" contain many statements that do not fit the corresponding topic. Whereas this appears quite random for "natural language technology" and "human machine interface", since we did not define suitable topics for these sentences. In "computer vision" many sentences end up containing words like scan, look, see. But these do not actually refer to the topic. Similarly "social media" includes statements with the word fans, even if this sentence is about attending a soccer match. But also in this category are plenty of randomly assigned statements. Nevertheless, the categories "computer vision" and "natural language technology" also contain appropriate phrases. The quality of these subjects is mixed. "human machine interface" consists of far too many terms that do not fit to this topic. A possible solution for such random assignments could be to add the generic category "others". Furthermore, it should be considered whether the category human machine interface is actually useful. It does not seem specific enough to narrow down a particular topic.
The Topic Model seems to assign the "research computing" category to statements more reliably. Nevertheless, it is noticeable that educational topics that cannot be associated with research and hardware or software are also grouped under this theme. Sentences regarding technical topics, which have no connection with research, also often fall into this category. However, in the end it is a matter of interpretation whether these statements do not fall under "research computing".
Finally, there are also topics that seem to be cor-

rectly assigned to their statements for the majority of the time. The corresponding categories are finance and gaming. This could be an indication that the model is not suitable for too specific technical terms. Providing more common topics contained in the everyday english language use, could result in a more reliable topic annotation.

4.3 Website Examination

The final data set, produced by the model pipeline, contains the URL for every AI future statement. Table **??** 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.

4.4 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 finetune this 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.

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 2: Top Domains

Investing more time in topic selection would also be beneficial. Therefore, it might be reasonable to manually evaluate our statements with the corresponding topics with a subsequent performing of a another topic selection. Furthermore, better results can be achieved by finetuning the topic model.

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

A Data Set Card: Future Statements

Data Set Description

This Data Set Card originates from https://huggingface.co/datasets/fidsinn/future-statements The english language data set contains 2,500 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 labels were set manually or programmatically (including corresponding manual examination of the labels).

Dat Set Motivation

The sole purpose of this dataset was to fine tune the distilbert-base-uncased https://huggingface.co/distilbert-base-uncased model into our distilbert-base-future https://huggingface.co/fidsinn/distilbert-base-future model. The dataset was created by students from the University of Leipzig in the Big Data and Language Technologies Module of the Webis Group https://huggingface.co/webis.

Data Set 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

- The dataset consists of 2,500 statements in total, 50% of which relate to future events and 50% of which relate to non-future events.

Data Set Annotation

- 0: No future statement
- 1: future statement

Noise, Biases and Redundancies

The main goal of the data collection process was to find future statements and general statements in equal amount. The thematic content within the statements can be redundant and some topics can be much more present. The dataset was not created to work with the thematic content while only fine-tune an already existing model into a model which is sensible for future and non-future statements. Data in the 'statement'-column is publicly available and does not contain confidential information. It was collected in the months 06/2022-07/2022 but the content of the dataset is independent of the data collection period and can be from earlier periods.

Data Set Collection Process

The data is directly observable on the websites mentioned in upper section. It was collected manually and programmatically (using Pythons NLTK library for automatic sentence-extraction and Regex-filtering) from graduate students D. Baradari https://huggingface.co/Dunya, F. Bartels https://huggingface.co/fidsinn, A. Dewald, J. Peters https://huggingface.co/jpeters92 as part of a data science module of the University of Leipzig. The data was obtained in the months 06/2022-07/2022 but the content of the dataset is independent of the data collection period and can be from earlier periods.

Dataset Maintenance

Curators of the dataset can be contacted via the Huggingface community tab https://huggingface.co/datasets/fidsinn/future-statements/discussions.

It is not planned to update the dataset for further work or investigations.

B Data Set Card: AI Future Statements

Data Set Description

The corresponding data set to this Data Set Card originates from https://github.com/atr2384/the-future-tense/blob/develop/stage_2_2_model_pipeline/output/future_statements.csv. This english language data set contains 15,541 tuples. Tuples consist of four attributes (statement (about the future), sentiment, topic, url). Any information in this data set is a results of combined data collection pipeline processes. This includes data extraction, filtering, preprocessing and application of several NLP models.

Dat Set Motivation

The overall purpose of this dataset is to provide information on statements about the future that can be analyzed in further steps. The dataset was created by students from the University of Leipzig in the Big Data and Language Technologies Module of the Webis Group https://huggingface.co/webis.

Data Set Composition and Annotation

The instances are represented by single-sentence statements which are tempted to be about the future of AI. Sentiments attribute represents the general sentiment in the corresponding statement. This can be NEG (NEGATIVE), NEU (NEUTRAL), POS (POSITIVE). A corresponding topic can be of a list of 9 topics including: machine human interface, finance, social media, search engine, computer vision, natural language technologiy, gaming, transhumanism, research computing. URL contains the URL from which the final AI future statement was extracted. The dataset consists of 15,541 tuples in total.

Noise, Biases and Redundancies

The main goal of the data collection process was to find future statements about AI topics and labeling it with a sentiment and topic. The thematic content within the statements can be redundant and some topics are be much more present than others. Through the fact that the data was collected from WARC-files, actuality and public availability of the url of a statement cannot be guaranteed. Data in the 'statement'-column is publicly available and does not contain confidential information. The statements within the dataset do not correspond to the opitions of the project team and is not associated with the authors of this project since the statements were extracted from WARC-files.

Data Set Collection Process

The instances are represented by single-/ or multi-sentence statements which were collected with a WARC-DL data extraction pipeline. A version of the WARC-DL pipeline which was fitted for the needs of the performed tasks can be found here: https://github.com/atr2384/WARC-DL. Since the AI statements are extracted from website HTMLs located inside WARC files, we added the url of each statement in the tuples. Processed data of the statement attribute is partially observable on the url of each tuple. Sentiments for the statements were generated by using the SentimentAnalyzer of the open-source library pysentimiento, which was further trained on about 40,000 tweets. It uses the BERTweet as a base model, pre-trained on english tweets. Topics were created in a topic model. This topic model consists of a topic discovery step. This includes a clustering of subcategories through an LDA model and a topic proposition step using GPT-3. This step also includes manual proposition. The subsequent assignments of topics to statements was performed by the bart-large-mnli model from Facebook. The complete workflow for the composition of the data set can be found at https://github.com/atr2384/the-future-tense The data in this data set was collected programmatically. Participants of the extraction process asre graduate students D. Baradari

https://huggingface.co/Dunya, F. Bartels https://huggingface.co/fidsinn, A. Dewald, J. Peters https://huggingface.co/jpeters92 as part of a data science module of the University of Leipzig. The data was obtained in 08/2022 but the content of the dataset is independent of the data collection period and can be from earlier periods.

Dataset Maintenance

It is not planned to update the dataset for further work or investigations.

C Model Card: Distilbert-Base-Future

This model is a finetuned on 2500 expressions, which contained 1250 future statements. distilbert-base-uncased serves as a base model. The corresponding Model Card can be found here: https://huggingface.co/fidsinn/distilbert-base-future. It includes an example for general use.

Model Description Contributors are D. Baradari, F. Bartels, A. Dewald, J. Peters. Questions and comments can be send via the Hugging Face community tab https://huggingface.co/fidsinn/distilbert-base-future/discussions

- 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
It achieves the following results on the evaluation set: - Train Loss: 0.1142 - Train Sparse Categorical Accuracy: 0.9613 - Validation Loss: 0.1272 - Validation Sparse Categorical Accuracy: 0.9625 - Epoch: 1

Training and evaluation data

The Distilbert-base-future model was trained and evaluated on the Future Statements data set https://huggingface.co/datasets/fidsinn/future-statements We collected 2,500 statements, 50% of which relate to future events and 50% of which relate to non-future events. For finetuning, we have used 80% of records from our self-annotated future-tatements dataset. This corresponds to 2,000 records. The remaining 500 were used to test the final distilbert-base-future model. The sole purpose of the dataset was the fine-tuning process of this model.

Training results

Framework versions

- Transformers 4.18.0
- Tensorflow 2.8.0
- Tokenizers 0.12.1

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 3: Training Results

D Remaining Figures

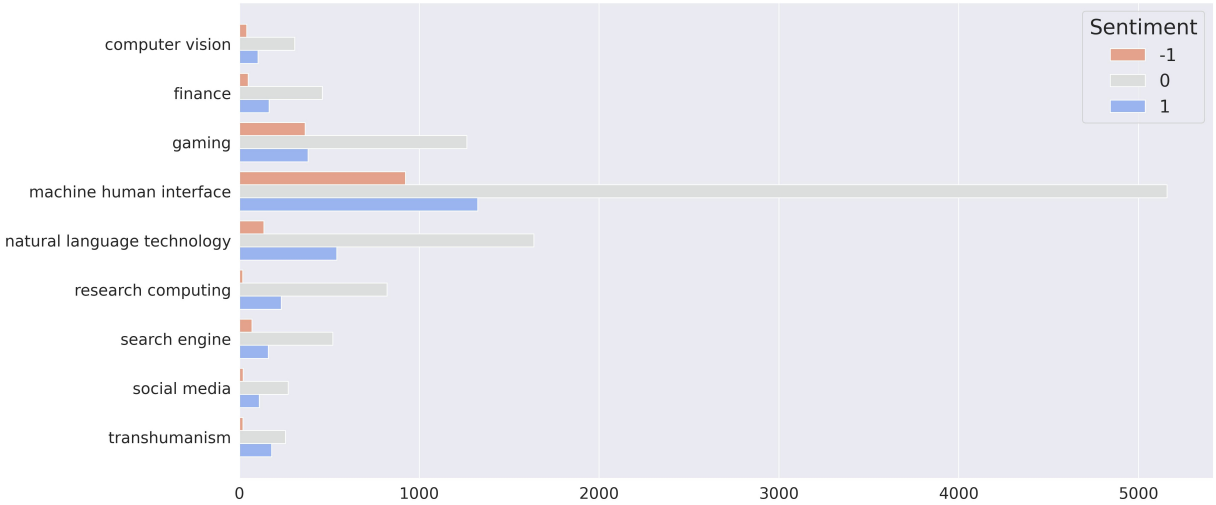

Figure 4: Distribution of statements among the 9 topic categories divided into the assigned sentiment labels (-1:NEGATIVE, 0:NEUTRAL, 1:POSITIVE). A domination of neutral statements can be observed for each of the 9 topics. There are visibly more positive than negative statements except for Gaming and Machine Human Interface.

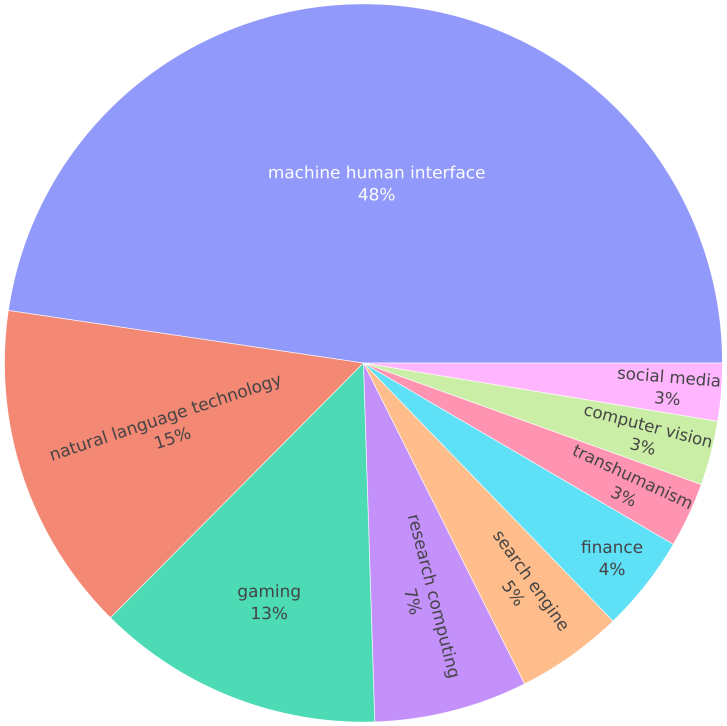

Figure 5: Distribution of statements among the 9 topic categories (in %). Statements are not equally distributed. Machine Human Interface describes about half of all statements (48%). Gaming as well as Natural Language Technology account for about 15% of all statements.

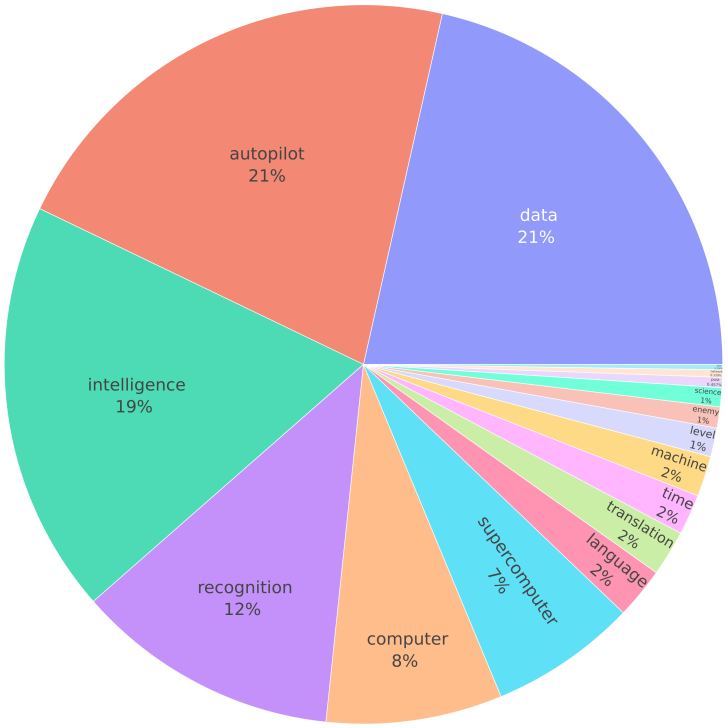

Figure 6: Distribution of statements among the most frequently occuring subtopic categories (in %). Data (21%), Autopilot (21%) and Intelligence (19%) are dominant subcategories.

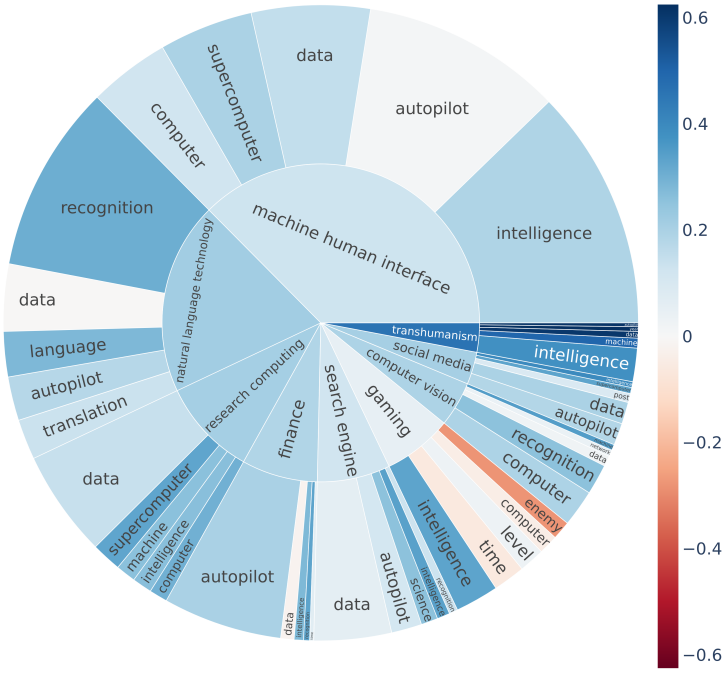

Figure 7: Distribution of statements among the 9 topics (inner circle) and the most frequently occuring subtopic categories (outer circle). The Colorpalette (on the right) represents the average sentiment score of topics and subcategories (Red: NEGATIVE, White: NEUTRAL, Blue: POSITIVE). The majority of the average sentiment in both topics and subcategories is neutral.