
CLIMATEXT: A Dataset for Climate Change Topic Detection

Francesco Saverio Varini
Department of Computer Science
ETH Zurich
Zurich, Switzerland
fvarini@student.ethz.ch

Jordan Boyd-Graber
CS, iSchool, LSC, and UMIACS
University of Maryland
College Park, MD, USA
jbg@umiacs.umd.edu

Massimiliano Ciaramita
Google Research
Zurich, Switzerland
massi@google.com

Markus Leippold
Department of Banking and Finance
University of Zurich
Zurich, Switzerland
markus.leippold@bf.uzh.ch

Abstract

Climate change communication in the mass media and other textual sources may affect and shape public perception. Extracting climate change information from these sources is an important task, e.g., for filtering content and e-discovery, sentiment analysis, automatic summarization, question-answering, and fact-checking. However, automating this process is a challenge, as climate change is a complex, fast-moving, and often ambiguous topic with scarce resources for popular text-based AI tasks. In this paper, we introduce CLIMATEXT, a dataset for sentence-based climate change topic detection, which we make publicly available. We explore different approaches to identify the climate change topic in various text sources. We find that popular keyword-based models are not adequate for such a complex and evolving task. Context-based algorithms like BERT [1] can detect, in addition to many trivial cases, a variety of complex and implicit topic patterns. Nevertheless, our analysis reveals a great potential for improvement in several directions, such as, e.g., capturing the discussion on indirect effects of climate change. Hence, we hope this work can serve as a good starting point for further research on this topic.

1 Introduction

The World Economic Forum [2] continues to rank climate change as one of the top global risks in the next ten years. Not surprisingly, climate change receives prominent public attention and media coverage, which makes it a fascinating object of study for natural language understanding (NLU). The first step in tasks such as sentiment analysis, fact-checking, and question-answering is the identification of the climate-change topic in text sources. This seems like an obvious task, which is commonly addressed by simple string matching from a keyword list; e.g., as in prominent financial economics literature [3]. However, consider the following statements:

1. Compliance with these laws and regulations could require significant commitments of capital toward environmental monitoring, renovation of storage facilities or transport vessels, payment of emission fees and carbon or other taxes, and application for, and holding of, permits and licenses.
2. Al Gore’s book is quite accurate, and far more accurate than contrarian books.
3. The temperature is not rising nearly as fast as the alarmist computer models predicted.
4. The parties also began discussing the post-Kyoto mechanism, on how to allocate emission reduction obligation following 2012, when the first commitment period ends.
5. The rate of Antarctica ice mass loss has tripled in the last decade.
6. Globally about 1% of coral is dying out each year.
7. Our landfill operations emit methane, which is identified as a GHG.
8. Due to concerns about the de-forestation of tropical rain forests and climate change, many countries that have been the source of these hardwoods have implemented severe restrictions on the cutting and export of these woods..
9. The 2015 conference was held at Le Bourget from 30 November to 12 December 2015.
10. Polar bear numbers are increasing.
11. In 2006, CEI arranged a public-service television commercial about carbon dioxide with the slogan “They call it pollution; we call it life”
12. Human activities emit about 29 billion tons of carbon dioxide per year, while volcanoes emit between 0.2 and 0.3 billion tons.
13. Further, these emission control regulations could result in increased capital and operating costs.

The first six sentences enumerated above cannot be detected using glossaries (Table 3) for keyword based models. These sentences are, instead, correctly identified using a machine-learned text classifier, BERT [1]. However, sentence 7 and 8 are straightforward examples. Very rarely, for some reason, these kinds of sentences are not correctly classified by BERT according to our results. On the other hand, they are always detected correctly by the simplest keyword based models. Climate change topic detection can be difficult and often patterns are implicit and ambiguous. For instance, while it is implicit but unequivocal that sentence 9 talks about a climate change conference, it is unclear whether the sentences 10 and 11 are about climate denialism, therefore about climate change. Sentence 12 talks only about a scientific fact: carbon dioxide emissions are there with or without climate change. Finally, sentence 13 is talking about regulations on emissions, however, it is omitted which specific ones: emissions of oxygen, for instance, would not relate to climate change.

Climate change is a complex topic with many different facets. These facets can be textually described in different ways, potentially combining several words distributed within the text. Moreover, climate change is a fast-moving topic for which new terms and concepts are emerging, e.g., in the public debate and new legislation. Therefore, we need to catch up with the language used for it continually. We argue that NLU and machine learning are needed to solve this task, and that solutions will have broader societal value, by helping in keeping track of the topic and its ramifications. To encourage research at the intersection of climate change and natural language understanding we built and make public a dataset for climate-related text classification together with preliminary findings.¹

2 Constructing the data set

The data consists of labeled sentences. The label indicates whether a sentence talks about climate change or not. Labels are generated heuristically or via a manual process. The manual labeling rules emerged through inspection of sentences and a collaborative labeling process with four raters, for which we monitored the inter-rater reliability through the Kappa statistic [4] (Table 7). We list the rules in Appendix B. Sentences are collected from different sources: Wikipedia, the U.S. Securities and Exchange Commission (SEC) 10K-files [5], which are annual regulatory filings in which listed companies in the US are required to self-identify climate-related risks that are material to their business, and a selection of climate-change claims collected from the web [6].

For Wikipedia, we select documents through graph-based heuristics based on Wikipedia Inlinks (see Appendix C). We collect 6,885 documents, 715 relevant to climate change and 6,170 not relevant to climate change. We divide the documents between train, development, and test sets. Then, we split

¹Our data is made available on www.climatefever.ai.

the documents into sentences, and we label these as climate change related or not, heuristically, using the same label as that of the document of origin (see Table 1).

Table 1: Wikipedia document labeled data sets (positives vs negatives in parentheses).

Data	Tag	Sentences
Train split	Wiki-Doc-Train	115854 (57927 vs 57927)
Development split	Wiki-Doc-Dev	3826 (1913 vs 1913)
Test split	Wiki-Doc-Test	3826 (1913 vs 1913)

Training on the data from Table 1 does not yield good predictive models because of the assumption that all sentences in a positive document are positives. Therefore, we follow up with Active Learning (AL) [7] to manually label thousand of additional instances. For this purpose, we use DUALIST [8][9], a web-based framework performing AL and running a multinomial NB model in the loop (see Appendix D). We label sentences from **Wiki-Doc-Train**. We also run this labeling process on Item 1A of the 10-K files from 2014, as this is the relevant section in which climate risk must be reported. Table 2 on the right side provides an overview of the data set created with AL.

Table 2: Evaluation and AL train sentences (positives vs negatives in parentheses)

Evaluation sentences		Active learning train sentences		
Data	Sentences	Data	Tag	Sentences
Wikipedia (dev)	300 (79 vs 221)	Wikipedia	AL-Wiki	3000 (261 vs 2739)
Wikipedia (test)	300 (33 vs 267)	10-Ks	AL-10Ks	3000 (58 vs 2942)
10-Ks (2018, test)	300 (67 vs 233)			
Claims (test)	1000 (500 vs 500)			

For the evaluation data, we proceed as follows. First, we create a development set from **Wiki-Doc-Dev** and a test set from **Wiki-Doc-Test**. We sample 150 sentences from the positives and 150 from the negatives. The four raters then label these sentences according to the labeling rules. Each sentence is deemed negative only if all raters labeled it as negative, positive otherwise. Then, we create another test set using only the 10-K files by adopting a Wikipedia trained BERT-predictions-based sampling scheme, randomly selecting 150 examples both within the positive and negative predictions. Then, the four raters label these sentences according to the labeling rules. Lastly, we collect 500 positive and 500 negative claims from the sources used in [6]. The left side of Table 2 gives an overview of the development and test sets created.

3 Analysis

For our analysis of the dataset, we rely on three model frameworks for classification.

Keyword-based models: We use several existing climate-related keywords sets as a benchmark, see Table 3.

Table 3: Glossaries used for the keyword-based models

Keywords source	Tag	Number of Keywords
Wikipedia Glossary [10]	Wikipedia-Keywords	175
IPCC Glossary [11]	IPCC-Keywords	340
Global Change Glossary [12]	GlobalChange-Keywords	126
FS-US Glossary [13]	FS-US-Keywords	241
Small	Small-Keywords	6
All	All-Keywords	771

Näive Bayes: The Näive Bayes (NB) classifier from DUALIST [8]. NB models usually provide competitive baselines, though NB assumes independence of the features given the class.

BERT: A popular attention-based text-classifier [1]. We use the BERT_{BASE} model pre-trained on Wikipedia and fine-tune it by adding an output layer for our specific binary classification task.

4 Results and discussion

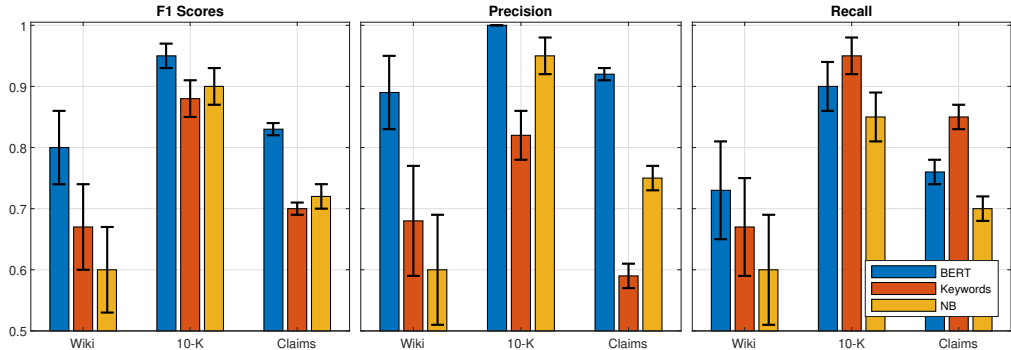


Figure 1:

The figure reports the highest F1-score (left panel) with the respective precision (middle panel) and recall (right panel) for the BERT, the keyword-based, and NB models.

We present results on the test sets, using mean accuracy, precision, recall, and F1 as performance metrics, estimating the standard error via bootstrapping.² Our results are as follows. First, as Figure 1 suggests, in terms of the F1 score, BERT outperforms the other models for all three test sets. The NB classifier, in contrast, fails to beat the keywords-based approach for the Wikipedia test set. Second, the best BERT model achieves the highest precision for all test sets, see Figure 1. Indeed, BERT’s performance in precision is the main cause for the performance in terms of F1.³ Third, while the BERT models reach a precision around or above 90% (see Figure 1), and maintain a remarkable recall higher or equal to 73%, the keyword approach has better recall on two of the three tasks.

We also find that the 10-K files are well standardized and present climate change topic patterns that are easier to detect than in Wikipedia or the collected claims. Indeed, these regulatory reports are often criticized for being boilerplate.⁴ However, we should remark that the 10K test set is sampled conditioning on the predictions of a BERT model. This means that it might be not entirely representative of the climate change topic distribution in the 10K files.

The above findings indicate that the climate change topic detection task in sentences can be challenging, even for state-of-the-art neural network models such as BERT. Keyword-based models are outperformed by BERT in terms of precision, but can be competitive in terms of recall. Can keyword-based models, being deterministic models, rise to the challenge of climate change language/patterns shifts over time? In principle, we could periodically enlarge the keywords set through automatic keywords discovery methods [15][16]. However, such a procedure may not work when the topic is as complex as climate change. We may need to discover combinations of words in the same sentence rather than single keywords/key-phrases. Therefore, one may need an oracle intervention, like a human encoding complicated patterns periodically in the deterministic search algorithm.

In the future, we plan to enlarge the data sets to eventually include a wider variety of the representative climate change topic distribution in sentences. An exciting avenue of further research is to understand the challenges for models like BERT when applied to complex topics like climate change. Finally, while we currently use sentences as the unit of analysis, a contextual understanding of whether a text is climate-related would be more nuanced and incorporate contextual information.

²BERT and NB are optimized with respect to F1 on the Wikipedia development set in Table 2.

³Tables 4 to 6 in Appendix A provide a detailed overview on the test results.

⁴However, using BERT, [14] find that 10-K filings do include important climate risk information that influence prices in financial markets.

References

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.
- [2] “World economic risks report 2020,” <https://www.weforum.org/reports/the-global-risks-report-2020>.
- [3] R. F. Engle, S. Giglio, B. Kelly, H. Lee, and J. Stroebe, “Hedging Climate Change News,” *The Review of Financial Studies*, vol. 33, no. 3, pp. 1184–1216, 02 2020. [Online]. Available: <https://doi.org/10.1093/rfs/hhz072>
- [4] M. McHugh, “Interrater reliability: The kappa statistic,” *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, 10 2012.
- [5] “Sec’s filings,” <https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm>.
- [6] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, , and M. Leippold, “CLIMATE-FEVER: Building a dataset for verification of real-world climate claims,” Tackling climate Change with Machine Learning, Workshop at NeurIPS 2020, 2020.
- [7] B. Settles, “Active learning literature survey,” https://www.researchgate.net/publication/228942691_Active_Learning_Literature_Survey, University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.
- [8] —, “Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances.” 01 2011, pp. 1467–1478.
- [9] B. Settles and X. Zhu, “Behavioral factors in interactive training of text classifiers,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT ’12. USA: Association for Computational Linguistics, 2012, p. 563–567.
- [10] “The Wikipedia glossary of climate change,” https://en.wikipedia.org/wiki/Glossary_of_climate_change, accessed: 2020-06-12.
- [11] “The ipcc glossary of climate change,” <https://www.ipcc.ch/sr15/chapter/glossary/>, accessed: 2020-06-12.
- [12] “The global change glossary of climate change,” <https://www.globalchange.gov/climate-change/glossary>, accessed: 2020-06-12.
- [13] “The U.S. Forest Service glossary of climate change,” <https://www.fs.fed.us/climatechange/documents/glossary.pdf>, accessed: 2020-06-12.
- [14] J. Koelbel, M. Leippold, J. Rillaerts, and Q. Wang, “Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure,” Working Paper, Available at SSRN 3616324, 2020.
- [15] S. Beliga, “1 keyword extraction : a review of methods and approaches,” 2014.
- [16] S. Siddiqi and A. Sharan, “Keyword and keyphrase extraction techniques: A literature review,” *International Journal of Computer Applications*, vol. 109, pp. 18–23, 01 2015.
- [17] R. Cilibrasi and P. M. B. Vitanyi, “The Google similarity distance,” 2004.
- [18] I. H. W. David Milne, “An effective, low-cost measure of semantic relatedness obtained from Wikipedia links,” 06 2010.
- [19] “Wikipedia dumps,” <https://dumps.wikimedia.org/>.

A Test Results

In the following results tables the model name is flagged with the training data used. In particular, whenever an “&” is present in a model name, it means that we train the model on successive training steps using different data. From left to right, the model training data is listed in the exact order of usage.

Table 4: Claims bootstrap test results sorted by F1 score descending. Standard deviation in parenthesis.

Model	Accuracy	F1	Precision	Recall
BERT-Wiki-Doc-Train & AL-10Ks \cup AL-Wiki	0.85 (0.01)	0.83 (0.01)	0.92 (0.01)	0.76 (0.02)
BERT-Wiki-Doc-Train & AL-10Ks & AL-Wiki	0.84 (0.01)	0.82 (0.01)	0.95 (0.01)	0.73 (0.02)
BERT-AL-Wiki	0.81 (0.01)	0.81 (0.01)	0.83 (0.02)	0.79 (0.02)
BERT-AL-10Ks & AL-Wiki	0.82 (0.01)	0.81 (0.01)	0.89 (0.01)	0.74 (0.02)
BERT-Wiki-Doc-Train & AL-10Ks	0.82 (0.01)	0.8 (0.01)	0.92 (0.01)	0.72 (0.02)
BERT-Wiki-Doc-Train & AL-Wiki	0.83 (0.01)	0.8 (0.01)	0.96 (0.01)	0.69 (0.02)
BERT-Wiki-Doc-Train & AL-Wiki & AL-10Ks	0.82 (0.01)	0.79 (0.01)	0.92 (0.01)	0.7 (0.02)
NB-AL-Wiki	0.73 (0.01)	0.72 (0.02)	0.75 (0.02)	0.7 (0.02)
BERT-Wiki-Doc-Train	0.62 (0.02)	0.72 (0.01)	0.57 (0.02)	0.98 (0.01)
BERT-AL-Wiki & AL-10Ks	0.76 (0.01)	0.72 (0.02)	0.89 (0.02)	0.6 (0.02)
All-Keywords	0.63 (0.01)	0.7 (0.01)	0.59 (0.02)	0.85 (0.02)
IPCC-Keywords	0.62 (0.02)	0.67 (0.02)	0.59 (0.02)	0.78 (0.02)
Wikipedia-Keywords	0.72 (0.01)	0.64 (0.02)	0.9 (0.02)	0.5 (0.02)
BERT-AL-10Ks	0.71 (0.01)	0.64 (0.02)	0.85 (0.02)	0.51 (0.02)
FS-US-Keywords	0.69 (0.01)	0.62 (0.02)	0.81 (0.02)	0.5 (0.02)
NB-AL-10Ks	0.69 (0.01)	0.6 (0.02)	0.82 (0.02)	0.47 (0.02)
GlobalChange-Keywords	0.62 (0.02)	0.41 (0.02)	0.88 (0.03)	0.27 (0.02)
Small-Keywords	0.6 (0.02)	0.34 (0.02)	1.0 (0.0)	0.2 (0.02)

Table 5: 10-K bootstrap test results sorted by F1 score descending. Standard deviation in parenthesis.

Model	Accuracy	F1	Precision	Recall
BERT-Wiki-Doc-Train & AL-10Ks & AL-Wiki	0.98 (0.01)	0.95 (0.02)	1.0 (0.0)	0.9 (0.04)
BERT-AL-Wiki & AL-10Ks	0.97 (0.01)	0.93 (0.02)	0.92 (0.03)	0.93 (0.03)
BERT-Wiki-Doc-Train & AL-10Ks \cup AL-Wiki	0.97 (0.01)	0.92 (0.02)	0.95 (0.03)	0.9 (0.04)
BERT-Wiki-Doc-Train & AL-Wiki & AL-10Ks	0.97 (0.01)	0.92 (0.02)	0.95 (0.03)	0.9 (0.04)
BERT-Wiki-Doc-Train & AL-10Ks	0.97 (0.01)	0.92 (0.02)	0.95 (0.03)	0.89 (0.04)
BERT-AL-10Ks & AL-Wiki	0.96 (0.01)	0.92 (0.03)	0.91 (0.03)	0.93 (0.03)
NB-AL-10Ks	0.96 (0.01)	0.9 (0.03)	0.95 (0.03)	0.85 (0.04)
FS-US-Keywords	0.94 (0.01)	0.88 (0.03)	0.82 (0.04)	0.95 (0.03)
BERT-Wiki-Doc-Train & AL-Wiki	0.95 (0.01)	0.87 (0.03)	0.98 (0.02)	0.79 (0.05)
BERT-AL-10Ks	0.94 (0.01)	0.86 (0.03)	0.82 (0.04)	0.91 (0.03)
Small-Keywords	0.94 (0.01)	0.85 (0.04)	1.0 (0.0)	0.74 (0.05)
Wikipedia-Keywords	0.91 (0.02)	0.79 (0.04)	0.83 (0.05)	0.75 (0.05)
BERT-AL-Wiki	0.83 (0.02)	0.71 (0.04)	0.58 (0.05)	0.94 (0.03)
GlobalChange-Keywords	0.8 (0.02)	0.56 (0.05)	0.56 (0.06)	0.57 (0.06)
NB-AL-Wiki	0.59 (0.03)	0.5 (0.04)	0.34 (0.04)	0.91 (0.04)
All-Keywords	0.44 (0.03)	0.44 (0.04)	0.28 (0.03)	1.0 (0.0)
BERT-Wiki-Doc-Train	0.39 (0.03)	0.42 (0.04)	0.27 (0.03)	1.0 (0.0)
IPCC-Keywords	0.44 (0.03)	0.42 (0.04)	0.27 (0.03)	0.91 (0.04)

Table 6: Wikipedia bootstrap test results sorted by F1 score descending. Standard deviation in parenthesis.

Model	Accuracy	F1	Precision	Recall
BERT-AL-Wiki & AL-10Ks	0.96 (0.01)	0.8 (0.06)	0.89 (0.06)	0.73 (0.08)
BERT-Wiki-Doc-Train & AL-Wiki	0.95 (0.01)	0.79 (0.06)	0.79 (0.07)	0.79 (0.07)
BERT-AL-10Ks & AL-Wiki	0.95 (0.01)	0.77 (0.06)	0.76 (0.07)	0.79 (0.07)
BERT-Wiki-Doc-Train & AL-10Ks \cup AL-Wiki	0.94 (0.01)	0.75 (0.06)	0.69 (0.08)	0.82 (0.07)
BERT-Wiki-Doc-Train & AL-Wiki & AL-10Ks	0.94 (0.01)	0.75 (0.06)	0.74 (0.08)	0.76 (0.08)
BERT-Wiki-Doc-Train & AL-10Ks & AL-Wiki	0.93 (0.01)	0.71 (0.06)	0.67 (0.08)	0.76 (0.08)
BERT-AL-Wiki	0.92 (0.02)	0.69 (0.06)	0.6 (0.08)	0.82 (0.07)
BERT-AL-10Ks	0.92 (0.02)	0.68 (0.06)	0.63 (0.08)	0.73 (0.08)
Wikipedia-Keywords	0.93 (0.01)	0.67 (0.07)	0.68 (0.09)	0.67 (0.08)
BERT-Wiki-Doc-Train & AL-10Ks	0.91 (0.02)	0.66 (0.06)	0.58 (0.08)	0.79 (0.07)
NB-AL-10Ks	0.91 (0.02)	0.6 (0.07)	0.6 (0.09)	0.6 (0.09)
FS-US-Keywords	0.88 (0.02)	0.57 (0.07)	0.47 (0.07)	0.73 (0.08)
NB-AL-Wiki	0.84 (0.02)	0.55 (0.06)	0.4 (0.06)	0.88 (0.06)
Small-Keywords	0.92 (0.02)	0.46 (0.1)	1.0 (0.0)	0.3 (0.08)
BERT-Wiki-Doc-Train	0.63 (0.03)	0.37 (0.05)	0.23 (0.04)	1.0 (0.0)
GlobalChange-Keywords	0.88 (0.02)	0.34 (0.08)	0.41 (0.11)	0.3 (0.08)
All-Keywords	0.52 (0.03)	0.29 (0.04)	0.17 (0.03)	0.91 (0.05)
IPCC-Keywords	0.55 (0.03)	0.28 (0.04)	0.17 (0.03)	0.81 (0.07)

B Final Labeling Rules

The labeling rules we agreed upon are the following:

1. The sentence labeled as positive must talk about climate change.
 - (a) Just discussing nature / environment is not sufficient.
 - (b) Discussing a general scientific fact or describing an aspect of the climate is only relevant if it is a mechanism / cause / effect of (climate) change.
 - i. No: “Methane is CH4”
 - ii. No: “Monsoons can affect shipping”
 - iii. Yes: “Methane increases temperature”
 - iv. Yes: “The Monsoon season could be more volatile than the last century”
 - (c) “Change” must be an aggregate change over longer periods of time
 - (d) Just mentioning clean energy, emissions, fossil fuels, etc. is not sufficient
 - i. rather it must be connected to an environmental (CO2)
 - ii. or societal aspect (divestment, Kyoto treaty) of climate change.
 - (e) Acid rain / pollution / etc. are environmental issues but are not related to climate change.
 - (f) Acronyms or names of entities, potentially well connected to climate change, must be mentioned along with some mechanism/cause/effect of climate change
 - i. No: “EPA has adopted new regulations”
 - ii. Yes: “EPA has adopted regulations in response to findings on increased emissions of carbon dioxide”
2. The sentence can talk about climate change during any period of Earth’s history.
 - (a) Yes: Massive eruptions all over the Earth’s surface caused lower temperatures for the next few centuries.
3. There may be ambiguity because we only consider individual sentences.
 - (a) If you cannot resolve an ambiguous reference (is EPA European Pressphoto Agency or Environmental Protection Agency), then use your best judgement about how to resolve the reference.

- (b) If you don't know what a person, event, or idea is, you can expand your knowledge with a quick web search.
 - (c) If after a quick quick search you still do not understand or in all other cases, label it as not relevant.
4. In case of doubt and in all the other cases, the sentence must be labeled as negative.

Table 7: Kappa coefficients and translated agreement level [4]

Kappa's value range	Agreement level
[.0, .20]	None
[.21, .39]	Minimal
[.49, .59]	Weak
[.60, .79]	Moderate
[.80, .90]	Strong
[.90, 1]	(Almost) perfect

C Wikipedia inlinks graph

For Wikipedia, we perform a document selection through graph-based heuristics, which we describe below. The selection procedure builds on the Normalized Google Distance [17], which is given by:

$$sr(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|, |B|))} \quad (1)$$

The idea behind the Normalized Google Distance in equation (1) is to establish a similarity score between a pair of documents. The intuition is to base this similarity on how many other articles link both of the documents in the pair, as in [18]. This is signaled in equation (1) by the intersection in the numerator of the equation $\log(|A \cap B|)$. The rest of the mathematical terms are just part of the normalization. For normalization the NGD takes into account both the cardinality of the link sets to article a and b and the total size of Wikipedia. The Normalized Google Distance usually assigns a similarity score between 0 (identical) and 1 (unrelated). We can notice that the similarity score jumps to negative infinity when the articles a and b are linked by two distinct non-empty sets of articles (A intersected with B is empty). This is due to the numerator set intersection in the NGD formula.

Now that it is clear what the Normalized Google Distance is, we explain in detail how we apply it to our document selection problem. The procedure we follow consists of a lot of pre-computation to avoid being stuck with expensive calculation and, possibly, out of memory errors. Basically, the first step is to construct four dictionaries from the 01/11/2019 Wikipedia dumps [19]:

1. A "Title To Integer" dictionary mapping each article title to an integer
2. A "Integer To Title" dictionary mapping an integer to the respective article title
3. A "is Linking" dictionary mapping a certain article title to the set of this article links. Each of the link is an article title
4. A "is Linked By" dictionary mapping an article title to the title of the articles linking it.

The dictionaries in 3 and 4 contain the articles title mapped to integers by the dictionary in 1. This in order to avoid out of memory problems when loading the dictionaries on a single machine.

At this point, using the dictionaries in 3 and 4, we are ready to traverse the Wikipedia articles graph. We are interested in the articles related to climate change. When we start our work, the Wikipedia "Climate change" article is a redirection page to "Global warming". For this reason, though we know that climate change is a wider topic than global warming, we decide to start our traversal of the graph from "Global warming". From such a Wikipedia links graph, the goal is to find similar documents to "Global warming". The similarity score between pairs of documents is calculated with the NGD. In this regard, once again, we notice that if two articles are co-linked by no other article, then the NGD is negative infinity. This means that we do not really need to compute all possible pairs of articles

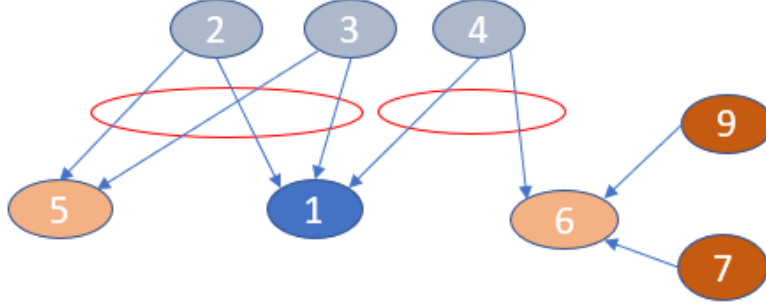


Figure 2:
A toy illustration of the links graph in Wikipedia.

similarity, but potentially only between co-linked pairs of articles. Conscious of this, we follow these steps for the graph traversal, which are represented in Figure 2:

1. We pick a set of articles S_a to start with, which is composed of only “Global warming” initially. In the Figure 2 this is represented by the node 1 in blue.
2. We collect the set of articles A_a linking the article a for each a in S_a by using the dictionary in 4. We store the results in a dictionary S_a -to- A_a . In the Figure 2 the newly collected articles in this step are represented by the nodes in gray.
3. For all the articles in each A_a stored in S_a -to- A_a , we collect the articles linked by them, excluding the ones in S_a , using the dictionary in 3. These are the yellow nodes in Figure 2. We call the retrieved set of articles S_b and we create a dictionary mapping each b to all articles a from S_a which have at least a parent article in common. We call this dictionary b -to- SS_a , where SS_a stands for a subset of the set S_a . This last dictionary is useful for us during iterations other than the initial one, where S_a is composed by more than one article.
4. Next, we gather the set of articles B_b linking article b for each b in S_b by using the dictionary in 4. We store the results in a dictionary S_b -to- B_b . In the Figure 2 the newly collected articles in this step are represented by the nodes in brown and gray.
5. Now, we calculate the Normalized Google Distance where, looking at the formula in 1, a and b comes respectively from the articles in S_a and S_b and A and B comes respectively from S_a -to- A_a and S_b -to- B_b . In our toy example from Figure 2, we will calculate the NGD between the blue article node in S_a and each of the yellow article nodes in S_b . The process can then repeat from 2, setting the new S_a as the newly retrieved S_b , which become our new blue reference nodes.

We can notice from Figure 2 that the set of articles linking to article 5, namely articles 2 and 3, is a subset of the articles linking article 1. This is signaled by the red circle on the left of Figure 2. The same situation does not happen for article 6, which has only article 4 as a common parent with article 1. Thus, we expect the NGD calculated between article 1 and 6 ($NGD(1, 6)$) to be lower, therefore better, than the one between article 1 and 5 ($NGD(1, 5)$). In fact, if we made the calculation in Formula 1, the $NGD(1, 5)$ would have a lower numerator, due to a higher cardinality of the sets intersection. At the same time, we would find that $NGD(1, 5)$ has a higher denominator than $NGD(1, 6)$, since article 5 is linked by two articles, which is lower than the three articles linking 6. Therefore $\log(|W|) - \log(\min(|A|, |B|))$ is higher in $NGD(1, 5)$ than in $NGD(1, 6)$.

We need to make still few important remarks about the procedure followed:

1. First, for each article collected during the graph traversal, we attribute a unique NGD similarity coefficient with respect to only one article. The problem is, in fact, that we can find the same article b in S_b which has common parent articles with more articles a from S_a in step 4 of the algorithm. In addition, we can still find the same article b in successive iterations. Therefore, we decide to select only the minimum score at the minimum distance level from the top “Global warming” article. We assume that the less distant linking-wise is an article to the starting article (“Global warming”), the more related to the latter it is.

2. Second, as the computation complexity grows exponentially in the number of collected articles N at each completed iteration, we decide to threshold the NGD similarity coefficients obtained. We decide to keep only the articles in S_b whose pair similarity is below this threshold. We collect these articles in the new S_a for the new iteration.

We run this algorithm to select the positive documents of the data set. Then, we sample the negative documents at random from the entire Wikipedia articles collection.

D Active Learning

The main motivation behind choosing an active learning [7] algorithm is to try to achieve greater accuracy with fewer samples. This is accomplished by letting the model choose which instance/feature to label.

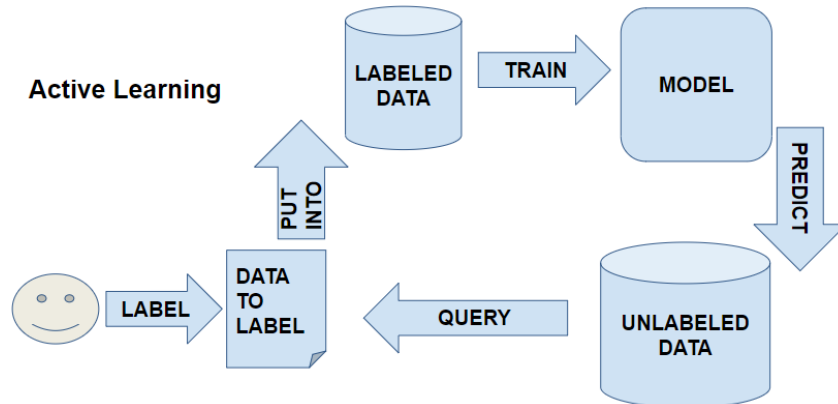


Figure 3:
Active Learning in one picture.

The high level procedure of active learning is shown in Figure 3 and it is described as follows:

1. Initialize a model or make use of an already trained model. Fit the model on a (new) set of labeled data, if any is currently available.
2. Predict with the model on the unlabeled data.
3. Use the model predictions to pose queries on the unlabeled data.
4. Show the data queried to an oracle (e.g. human) to be labeled
5. Once the queried data is labeled, put it into the labeled data, while removing it from the unlabeled set. Repeat the process from 1.

E DUALIST active queries explained

$$H_{\theta}(Y | x) = - \sum_j P_{\theta}(y_j | x) * \log P_{\theta}(y_j | x) \quad (2)$$

DUALIST pose queries on instances and features. The features are unigram and bi-gram from the sentences. Equation (2) represent the entropy based uncertainty sampling which DUALIST uses to query the instances to label by the oracle. Given the model predictions, we can attribute an entropy score to each unlabeled sentence from the unlabeled set. Then we can rank the sentences according to their entropy score. The higher the score, the more confused the model is about their classification. This usually happens to be around the model decision boundary as displayed in a toy two-dimensional space in Figure 4. However, there is no guarantee that these samples are the most informative for the task at hand. In fact, they could also be outliers, meaning samples which deviates significantly

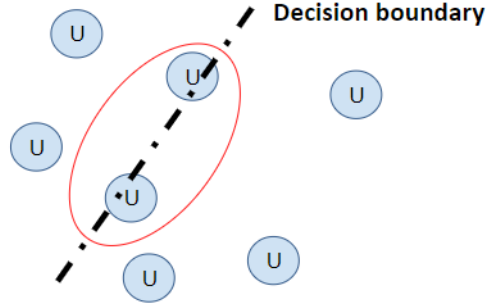


Figure 4:

A two dimensional representation of the model decision boundary and the location of different samples. The red circle indicates the closest samples to the decision boundary

from the rest of them. Outliers confuse the model and make the training, together with the labeling, inefficient.

$$IG(f_k) = \sum_{I_k} \sum_j P(I_k, y_j) * \log \frac{P(I_k, y_j)}{P(I_k) * P(y_j)} \quad (3)$$

As we previously mentioned, DUALIST queries also the features to label by the oracle. Equation (3) represents the information gain mathematical expression which DUALIST uses to score the features. To calculate the score, we need to know the label of our sentences. For this purpose, DUALIST makes use of the data labeled that far, if any, and the probabilistically-labeled instances by the model predictions on the unlabeled data. In equation (3), it is taken the ratio between the joint probability of a feature \underline{k} to occur in a sample with label j , with the chance the feature and the label occur independently of each other. This is shown in the logarithmic part on the right hand side of the equation. We can also observe that the IG scores are per class j , which means that each class has its own feature importance ranking.