IT UNIVERSITY OF COPENHAGEN

Fact Extraction and Verification in Danish

Sidsel Latsch Jespersen (slje@itu.dk) - Mikkel Ekenberg Thygesen (mekt@itu.dk)

June 1, 2020

Supervised by Leon Derczynski

Contents

Abstract

In this paper we present our work in creating an NLP model for fact extraction and verification for the Danish language. Fact extraction and verification has gained interest due to the vast amount of accessible online information surrounding people on an everyday basis and the increasingly important task of navigating between correct and incorrect information, also known as information disorder. For this task, we fine-tuned a BERT model and created a data set of Danish claim and evidence entities to be used in the model. Our model achieved a weighted F1 score of 67.30 % and a weighted accuracy score of 63.18 %. With this work, we hope to create a basis for further development of fact extraction and verification models for the Danish language.

1 Introduction

Automatic detection of misleading news is a widely discussed topic, and the relevance of the field quickly becomes apparent at a swift glance over the recent events. For example, Twitter marked an official tweet by U.S. president Donald Trump as "glorifying violence", causing him to threaten its general availability.[1] Furthermore, Facebook co-founder Mark Zuckerberg recently stated that social media should not be obliged to fact-check shared information on their platforms. This was met with a multitude of irony-ridden slander articles about the co-founder, e.g. calling him a child molester[2] or pronouncing him dead[3] which supposedly was meant to demonstrate how it is possible to weaponise false information published in social media and that, consequently, fact-checking information is essential.

In the field of Natural Language Processing (NLP), there has been a growing interest in developing models that can fact-check information. Society is infused with information that seems to come from everywhere, and it is not easy to discern between correct and incorrect information. Fact extraction and verification concerns the use of neural network computer models to automatically validate if given information is correct or not. In 2018, a workshop and shared task among international researchers in NLP brought attention to the field and several models for fact extraction and verification on English source material have since been developed (Thorne et al. 2018).

Because fact checking of information seemingly continues to be relevant in an increasingly more online world, it is not only relevant in English, but also in all other languages, and especially in those countries where communication is highly digitised. We have therefore focused on developing a model for the Danish language, a model that we hope can be a foundation for further research into fact extraction and verification in other languages.

[1] https://www.theguardian.com/technology/2020/may/29/twitter-hides-donald-trump-tweet-glorifying-violence

[2] https://chaser.com.au/world/social-media-should-not-fact-check-posts-says-child-molester-mark-zuckerberg/

[3] https://www.theshovel.com.au/2020/05/28/mark-zuckerberg-dead-at-36-says-social-media-fact-check/

1.1 Research Objective

Our project started with two problems: a) How do we collect a data set of claims, evidence, and veracity labels, and b) How do we train a machine learning model to predict the veracity of a claim? These were the initial tasks at hand, but as the project progressed, numerous other questions quickly arose, such as a) How do we collect the evidence? b) How do we organise the annotation process as cleverly as possible? c) What is the state-of-the-art technology to use for the fact extraction and verification task?

Our first subproblem was to collect the data. The unique part of this challenge was that the data should be written in Danish. Currently, there exists other data sets that fit the task, but these are primarily in English. We wanted this project to contribute to the field of Danish fact extraction and verification. We have described this in further detail in Section 4, where we have also described the tool which we developed and used for this task, the annotation process, and the characteristics and properties of the final data set.

We aimed to explore what was possible for automatic claim verification. To do this, we looked towards machine learning, where research in Natural Language Processing is growing steadily. Researchers in this field have been working on training models that can recognize languages and extract their features, and, more relevantly, detect the veracity of a postulate when compared to a body of evidence. We sought to use state-of-the-art technologies to develop our model and hopefully manage to create a benchmark for future research in Danish fact extraction and verification. This task is described in Section 5, where we also address how we prepared the data, what technological components we used, and how we measured this benchmark. Results will be presented in Section 6.

2 Fact Extraction and Verification

2.1 Background

There has been a lot of debate around the term "fake news" in recent years. The advent of internet-based communication, social media, and news channels have intensified what was previously mostly associated with historical authoritarian regimes, i.e. information issued with the intentions to direct or mislead public opinion. The term *information disorder* has been used to comprise the different terms used in the field, especially *misinformation*, false content not always intended to be false or misleading; and *disinformation*, content intentionally designed to mislead or cause harm (Wardle 2019: 7-8). Information disorder such as rumours, hoaxes, or conspiracy theories that before the arrival of the internet would eventually die out in lack of a broader audience are now proliferating with possible far-reaching consequences.

Authorities, political parties, international organisations, media companies and similar agents have discovered the need for counterbalancing information disorder, especially after the 2016 US presidential election, where false information was amplified on social networks (Watts, Rothschild 2017: 1). One method is to outright ban the publishing of incorrect information with laws such as the Singaporean *Protection from Online Falsehoods and Manipulation* bill from 2019 (Thygesen, Thomsen, Jespersen 2019: 18). This step must also be viewed in a context of free speech and censorship, since such a law can effectively prevent people from issuing critical utterances and be used to punish political opponents or dissidents. Such laws can be seen as adding to existing flexibility clauses in authoritarian regimes, but also Germany has passed its "Netzwerkdurchsetzungsgesetz" to "fight hate crime, criminally punishable fake news and other unlawful content on social networks"[4], and Italy and France are pursuing similar strategies (Alemanno 2018: 2-4).

Therefore, there is a huge interest among social media networks and other parties to track and remove information disorder.

2.2 Measures to Counter Information Disorder

Media houses and other agents with an interest in countering information disorder have different tools available. There are fact checking networks and organisations such as Poynter's International Fact-Checking Network[5], with a code of principles for fact-checking that tech companies such as Google and Facebook adhere to, Bellingcat[6], an NGO which gathers information and conducts research on information disorder and events affected by it, and Snopes, a fact-checking publication[7]. Journalists can themselves conduct research and investigation using databases, encyclopedias, and other tools such as Google Maps, web scrapers to collect online information, etc.

Fact checking is an extensive and time-consuming task which would benefit from a more automatic approach. This is what fact extraction and verification in NLP seeks to offer.

[4] `https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html`
[5] `https://ifcncodeofprinciples.poynter.org/`
[6] `https://www.bellingcat.com/`
[7] `https://www.snopes.com/`

2.3 Fact Extraction and Verification Definition

Fact extraction and verification is an area in NLP that concerns the general task of using AI and neural networks to seek out content and verify or classify such content. This includes question answering, rumour verification, stance classification, and fake news detection among other things, and also smaller subtasks such as evidence search. The overall goal is to create models that can be used to generate knowledge or predictions on something - text, images, audio files. As the name says, fact extraction and verification strives towards building models that can extract and verify facts.

2.4 Natural Language Inference

Within NLP, Natural Language Inference (NLI) is an area with tasks where a premise and a hypothesis should generate a label. This could be a fact extraction and verification task, where the evidence is the premise and the hypothesis is the claim, and the label should tell if the claim is either supported or refuted by the evidence, or if there is not enough information in the evidence for neither of the two. Another NLI task is textual entailment, where entailment in this respect means that there is a directional relationship between the premise and the hypothesis. An NLI model should then be able to predict if the hypothesis entails or contradicts the premise. An example of this is The Stanford Natural Language Inference (SNLI) Corpus, which is a corpus containing a premise, i.e. "A man inspects the uniform of a figure in some East Asian country.", and a hypothesis, "The man is sleeping". A model should then return either a contradiction, entailment, or neutral label (Bowman et al. 2015).

2.5 The FEVER Shared Task

The first FEVER shared task was initiated in 2018 with the purpose of evaluating "the ability of a system to verify information using evidence from Wikipedia" [8]. This task was based on the need to identify verifiable knowledge versus false information from unreliable sources and as such a possible tool to encounter information disorder. The subtasks included extracting textual evidence from Wikipedia that could support or refute a given claim. The system should label the claims either SUPPORTED, REFUTED, or NOTENOUGHINFO. The latter was used when there was not found enough evidence to either support or refute the claim, and the evidence could consist of one or more sentences from the data extracted from Wikipedia. The FEVER task was solved in different ways by the participants whose solutions were not perfect but gave an understanding of where fact extraction and verification could lead to with further development.

We have tried to implement a model much like requested in the FEVER task, but with significant differences: 1) The FEVER participants must create a model that, in addition to evaluating claims, also searched for the right evidence to evaluate a given claim; 2) The FEVER annotation process differed from ours, since the FEVER participants were given a data set beforehand containing ready-made claims and English Wikipedia articles to be used as evidence.

[8] https://fever.ai/2018/task.html

3 Collecting and Preprocessing Data

A large part our project has been to collect and preprocess the data needed for our model. Generally, researchers in NLP invest much time and effort in creating data sets and the amount of work in this task should not be underestimated.

3.1 Data Statement

Fact extraction and verification projects usually entail an annotation team with curators, annotators, and sometimes speakers. A curator is in charge of selecting relevant data and setting up classification criteria, while annotators do "the hard work" of assigning annotations to data. In the case of our project, the curators and annotators overlap and consist of the two authors of this paper. With only two annotators, there is a higher risk of bias compared to a larger annotation team, since the annotator demographic spans shorter. According to Bender & Friedman, data will always include pre-existing bias, and it is not possible to build NLP systems immune to bias (Bender & Friedman 2018: 589). Therefore, researchers must seek to mitigate any skewness resulting from imperfect data sets, and for this purpose they propose to include a Data Statement describing the closer details of the data sets involved (Bender & Friedman 2018: 587).

Consequently, apart from having described the data set used in the model in this chapter, we have included a Data Statement in the appendix of this paper which gives a detailed description of the metadata surrounding the data set and the annotation process; see 9.1.

3.2 Data Characteristics

The data we needed for our model should have the form *Claim - Evidence - Label*. Claim would be a statement, i.e. "Copenhagen is the capital of Denmark", and Evidence would the textual evidence that proves the claim to be true. The label would indicate whether the claim was SUPPORTED, REFUTED, or NOTENOUGHINFORMATION if the evidence neither supported or refuted the claim.

Claim	Evidence	Label
Skatte-niveauet i Finland er lavt	Finske politikere har ofte efterlignet de øvrige nordiske lande og den skandinaviske velfærdsmodel. De nordiske lande har gået ind for frihandel og har taget relativt godt imod veluddannede indvandrere gennem det seneste århundrede, om end indvandring i Finland er et relativt nyt fænomen. Niveauet af handelsbeskyttelse har været relativt lille med undtagelse af landbrugsprodukter. Finland ligger i toppen på de fleste områder inden for økonomisk frihed, selv om skatteniveauet er ganske højt, og arbejdsmarkedet er ret ufleksibelt. Finland står som nummer 17 (8 i Europa) på Index of Economic Freedom for 2012.	REFUTED

Figure 1: Example of a claim entity.

3.3 Annotation Process

As opposed to the participants in the FEVER shared task we did not have an annotation team to create our data sets of claims and evidences. We quickly found out that in order to generate material of an acceptable quality for our project, we needed to create the claim entities manually. The first step was to find data sources that would qualify. A data source would have to be reliable, neutral, trustworthy, and accessible.

We decided to extract data from the Danish Wikipedia and the encyclopedia Den Store Danske. Den Store Danske is a Danish encyclopedia originally published in paper as "Den Store Danske Encyklopædi" but since 2009 online and with the possibility of user editing, although editors and specialists at the publisher, Gyldendal, verified the content[9]. Den Store Danske was, before going online and allowing user-generated content, considered an authoritative source of knowledge in Denmark.

Danish Wikipedia is considered trustworthy and has marked articles which are considered either good (plus-marked) or excellent (star-marked). We focused on these highlighted articles from Wikipedia along with some of the more thorough and in-depth articles from Den Store Danske in order to have an acceptable amount of data. Because of time constraint, we would not create the claim entities like the participants in the FEVER shared task. Instead, we would write claims and label these manually. We created a claim and evidence-generating application for us to work smarter - and faster - when generating claims. The application scrapes articles from Wikipedia and Den Store Danske and divides them into sections of each three sentences. These sections then serve as basis for writing claims. For each claim, we would then label the claim as either SUPPORTED, REFUTED, or NOTENOUGHINFORMATION. In this way, we created a data set with 3010 entries. An example of such entry can be seen below.

[9]as from May 2020 the content has moved to lex.dk

Claim	Evidence	Label
Hillary Clinton er uddannet på Yale Law School	Hillary Clinton, Hillary Rodham Clinton, f 26101947, amerikansk jurist og politiker, gm. Bill Clinton i 1975; udenrigsminister 2009-13. Efter uddannelse på bla. Yale Law School, hvor hun mødte sin mand, skabte Hillary Clinton sig en succesrig juridisk karriere.	SUPPORTED

Figure 2: Example of entry in data set with claim, evidence and label.

This annotation process differs from the FEVER workshop (see 2.5) but we considered it to be efficient and satisfactory given the project's size and time frame.

3.3.1 Annotation Reliability

In order for the annotation process to be reliable, the annotators must create similar labelling of claims in the annotation process. This means for example that a claim which one annotator labels as REFUTED must also be annotated as REFUTED by another annotator. The larger a difference between annotators' labelling there is, the more unreliable the annotation and hence the evaluation model will be. To investigate the annotation reliability, we examined a sample of the annotated claim entities to calculate the Inter Annotation Agreement (IAA), which is the percentage of claims that the annotators have classified with the same label. Our results showed an IAA of 90.48 % which is not quite satisfactory.

3.4 Data Collection

First, we needed to collect a data set of evidence, which we collected from the Danish Wikipedia and Den Store Danske, as previously described. Each piece of evidence had a three-sentence maximum and a 200 word count maximum. The reasons for these restrictions were technical. The transformer architecture, which our model is based on, can experience performance problems with large inputs. This issue is currently being looked at by researchers, and recently by Beltagy et al. (2020). It is however out of scope for this project to directly consider their techniques, so we opted for a segmentation of our data.

We collected 378 pieces of evidence from Danish Wikipedia and 109 pieces of evidence from Den Store Danske.

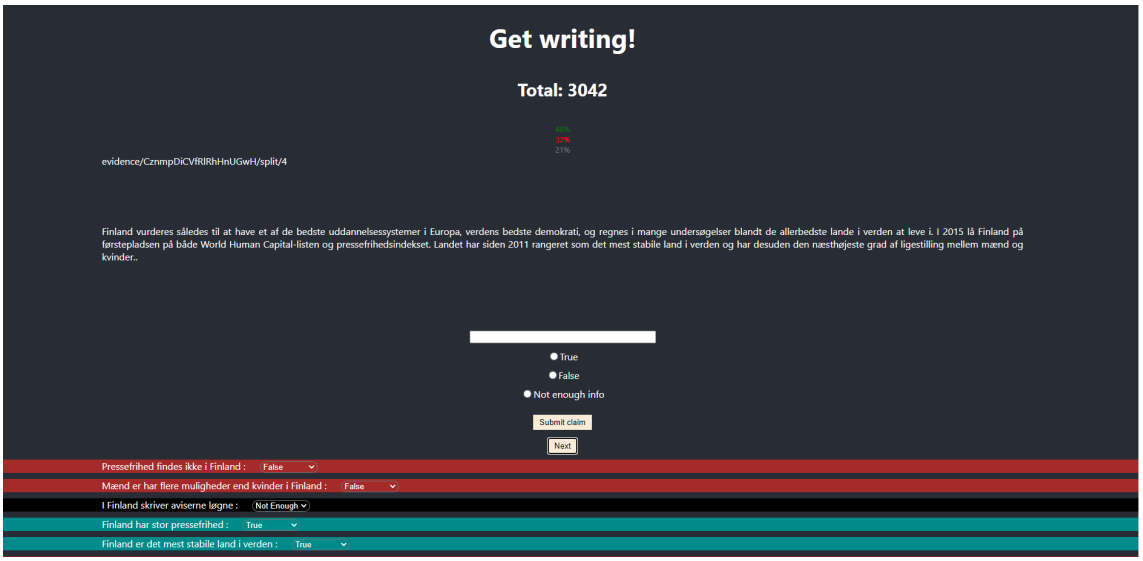
Source	Pieces of evidence	Claims	Dist. of Evidence	Dist. of Claims
Wikipedia	378	1744	77.61 %	57.95 %
DenStoreDanske	109	1266	22.38 %	42.05 %

Table 1: Data distribition by source

Table 1 shows that the majority of our evidence came from Wikipedia. This was to be expected, as Wikpedia's pages were easier to access. The distribution of claims is also favoring Wikipedia,

but it is more even with about 58 % of claims being from Wikipedia and 42 % from Den Store Danske.

Secondly, we uploaded the evidence to an online database created by us, and we developed a small web application tool that provided a random piece of evidence to the user. The user could then type in, label, and submit claims based on the piece of evidence. This made for an easier and less error-prone experience to generate data for our data set. To quality-ensure the data created, we added a label counter that continuously showed the percentual distribution of labels, such that the user would be able to keep track of serious skewnesses in the distribution of labels - e.g. if the user had created too many claims labelled SUPPORTED. Another feature was a drop-down label menu for all the claims generated for each piece of evidence such that the user had the possibility of changing a label if they mistakenly labelled a claim incorrectly. All the generated claims and their labels were shown on each page with the related evidence such that the user could do a quality check after finishing labelling claims for a piece of evidence.



Screenshot of the tool created by the authors to generate claims.

With the tool in place to make the claim creation process more effortless, we continued with the process of actually creating the claims.

3.4.1 3K Data Set

In total we created 3010 claims based on 487 paragraphs of evidence. The distribution of veracity labels can be seen in the figure below.

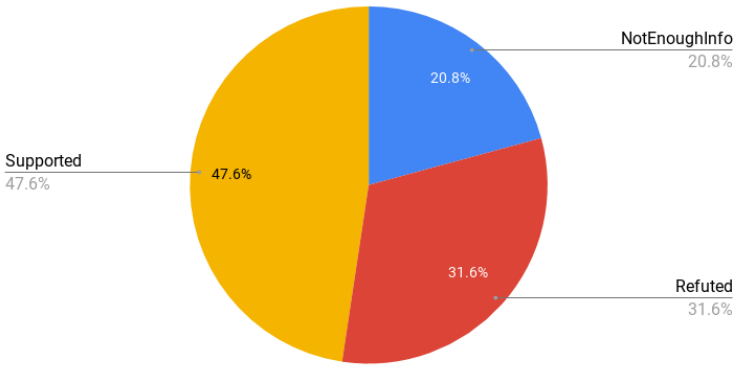


Figure 3: Distribution of labels for the 3K data set

The claims had an average word count of 6 words and every piece of evidence had an average word count of 77.

	Avg. word count	Min. word count	Max. word count
Claims	6	3	16
Evidence	77	22	191

Table 2: 3K data set properties

And lastly we can see the word count distribution across the entire data set:

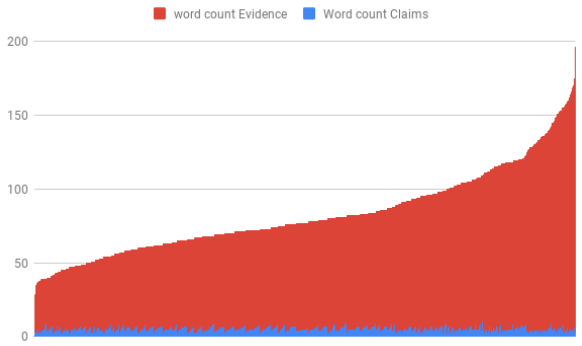


Figure 4: Word count per claim entity for the 3K data set

We concluded from the presented figures that the majority of claim entities were in the 50 to 100 word count range. The average word count of the claims was much lower than the average word count of the associated pieces of evidence. We assessed this as insignificant, since other FEVER task projects have not only dealt with a word count ration like ours, but also dealt with handling multiple pieces of evidence per claim.

In the distribution of labels there was also an overrepresentation of the SUPPORTED label, which was almost half of the claims in the data set. We have addressed this in the upcoming section.

3.4.2 4K Data Set with Randomly Generated Claim Entities

We suspected that an overrepresentation of SUPPORTED claims and an underrepresentation of NOTENOUGHINFO claims could cause problems. The primary problem being that the model could learn that, by solely predicting the overrepresented label, it could score a better performance. A technique in data analysis to address this is called over- and undersampling, wherein one adjusts the label distribution of the data set. However, this technique is usually done by either removing or adding data entries with the over- or underrepresented label. Since our data set was already relatively small (and it is generally not advisable to delete valid data), we decided to modify the distribution with artificial data.

We wished to generate and test a data set that included 1000 automatically generated claim entities all labelled NOTENOUGHINFO. This label is already applied to claim entities with no immediate correlation between the claim and the piece of evidence. A random claim and a random unrelated piece of evidence would evidently be labelled NOTENOUGHINFO. We thus theorised that we could artificially generate random pairs of claims and pieces of evidence and correctly label them with NOTENOUGHINFO. The generation was thus simple: pair a random claim with a random evidence and label it NOTENOUGHINFO. We assessed that due to statistical probability, the entries would be random enough to be correctly labelled. A quick qualitative look at the generated entries showed that they were completely random:

Golfkrigen fandt sted i 1991	Finland gik ind for på betingelserne i Kyoto-aftalen og inden for EU omkring udledning af drivhusgasser ...	NOTENOUGHINFO
Elektricitets-produktionen i Norge er baseret på vedvarende energi	Da Hitler invaderede Sovjetunionen udtrykte Winston Churchill, som ellers var en hårdnakket anti-kommunist	NOTENOUGHINFO
Der lever rødlosser i Canada	Nazisterne havde nu fået regeringsansvaret, og i de næste fem måneder ...	NOTENOUGHINFO

Table 3: Sample data from the auto-generated data set

We added the automatically generated entries to the data set with 3010 claims. The resulting data set now had 4010 entries based on 526 paragraphs of evidence. The distribution of labels was:

	Avg. word count	Min. word count	Max. word count
Claims	6	3	19
Evidence	76	5	191

Table 4: 4K data set properties

As we can see, the properties of the data did not change much. The average word count and maximum word count remained largely identical to the properties of the 3K data set. The

minimum word count for the evidence went from 22 to 5, which is a large leap. However, from our distribution graph, we could see that this was a single, random pair of a short claim and a short piece of evidence. From the new distribution graph we could see that it, too, remained visually unchanged.

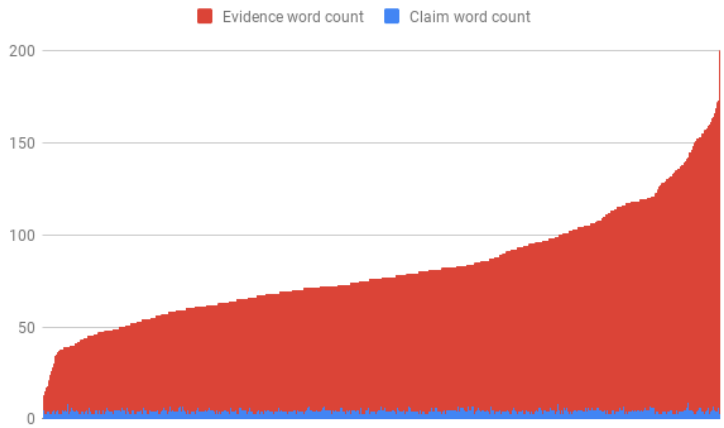


Figure 5: Diagram showing the individual word count of each pair of claim and associated piece of evidence.

While the properties of the individual entries were seemingly unchanged, our objective with the added random generated entries were to equalize the distribution of labels across the entire data set. From the figure below, we can see that this succeeded.

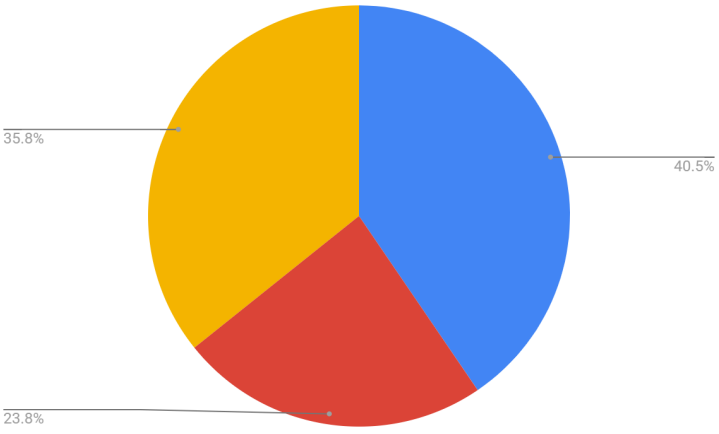


Figure 6: Diagram showing the distribution of labels of our 4K data set. Notice that the distribution is more equal now.

The data set now had a more equal distribution of at least two of the three labels of SUPPORTED, REFUTED, and NOTENOUGHINFO.

3.5 Individual Overrepresentation

While the data set itself can be overrepresented by a label which is caused by a data imbalance, so, too, can the individual pieces of evidence. Our claims were based on 487 pieces of evidence, but the distribution of labels for each piece of evidence was not necessarily equal. As such, some claims associated with the same piece of evidence may have been overrepresented.

We found a few such examples in our data set:

Evidence	SUPPORTED	REFUTED	NOTENOUGHINFO	Total claims
1	0	**10**	0	10
2	**10**	2	1	13

Table 5: Overrepresented labels per piece of evidence

As we can see, some of our pieces of evidence have an imbalance of associated labels. The evidence in row 1 have 10 claims labelled REFUTED and 0 claims labelled SUPPORTED and NOTENOUGHINFO. The latter evidence does have claims that represent each label, but the number of SUPPORTED claims outweigh the other labels.

The graph (figure 7) below shows how much the most dominant claim is represented for every piece of evidence. Ideally every piece evidence would be 33% - since there are only three labels available. However, the graph shows that the majority of our pieces of evidence have a label imbalance, and a portion of our pieces of evidence have only claims with one label.

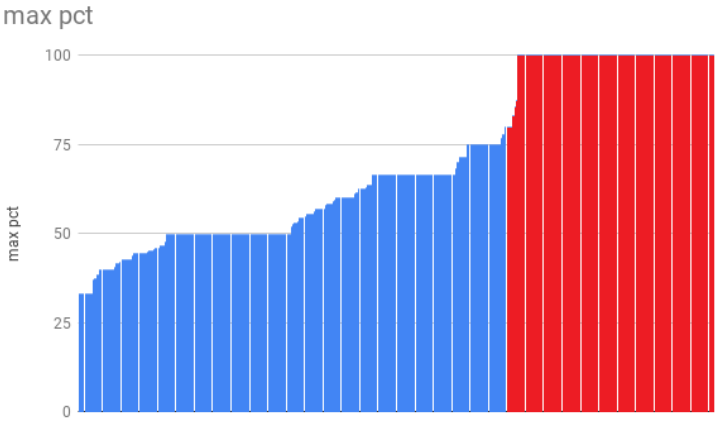


Figure 7: Distribution of dominant labels per piece of evidence. The red section of the graph marks the evidence whose claims only have a single label.

We believe that if the the piece of evidence have multiple-labelled claims, it will not affect the performance of our model, so let us then take a closer look on the upper percentile of the pieces of evidence.

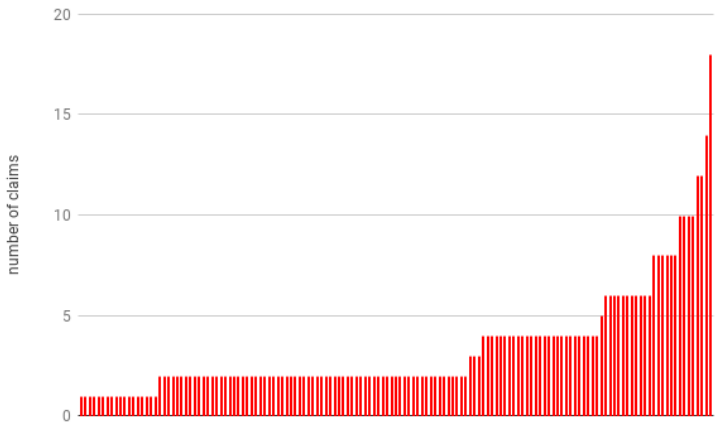


Figure 8: Number of claims for each piece of evidence with a dominant label.

The graph above shows how many claims are associated with a given piece of evidence that only have a single label. The vast majority have between one and five claims, while a minority have between six and ten. Lastly, a few outliers have 11, 12, and 18 claims with the dominant label.

3.5.1 Addressing the Bias

Having overrepresentation in a model can cause unreliable performance. Locating the overrepresentation of the entire data set is a trivial task, and specialized libraries have been developed to counter those problems. However an individual overrepresentation is a somewhat overlooked problem. Mostly because it is unsure if it actually is a problem. Measuring whether the individual overrepresentation has a severe effect on the performance is tricky to measure - especially with a data set as small as ours.

From our analysis of the problem, we found that for the selected pieces of evidence where only one label were represented, the amount of associated claims were few enough to not be a problem for the model's performance. The outliers with a high amount of claims were themselves few. This could be a problem for the performance, but we do not have the resources to test this theory. We do believe that they are few enough to not be a significant problem.

4 A Model for Danish Fact Verification

In this section we will describe the technical implementation of our model.

4.1 Data Preparation

Any machine learning model makes use of complicated math to perform its task. As such, it is important that the training data is numeric and calculable. Since our data was based on language, we needed to prepare it by converting it from words and letters into numbers. This process is a multi-part operation, and will be described in this section.

Sequencing

Currently, the data is separated into individual values in rows and columns. For each row, we concatenate a claim, a piece of evidence and its veracity label. The model requires the sentence to include specific tags to figure out when one value ends, and another begins. We mark these properties by applying a [CLS] tag to the beginning of the sentence and a [SEP] tag in between and at the end of the individual sentences:

Claim	*USA er et monarki*
Evidence	*USA, United States of America, forbundsstat i Nordamerika ...*

The two sentences become a sequence:

[CLS] USA er et monarki [SEP] USA, United States of America, forbundsstat i Nordamerika ... [SEP]

Each word in the sequence is then assigned an ID, which is used for indexing and lookup.

$$[101, 11061, 10163, 10131, 34372, 105707, \dots]$$

Next we extract the sequence's Token Type IDs, which are an array of 0s and 1s, where 0s mark all the tokens of the claim and 1s mark all the tokens of the evidence:

$$[0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, \dots]$$

With the data formatted and indexed, we can now tokenize our data.

['[CLS]', 'USA', 'er', 'et', 'mon', '##arki', '[SEP]', 'USA', ',',
'United', 'States', 'of', 'America', ',', 'for', '##bund', '##ss', '##tat',

As can be seen, our text has been applied a lot of #'s . When we tokenize a sequence, every subword will be stored and replaced with double hashtags. This is to recognize when a subword is used for different words. For example, the word "for" is not only a word by itself, it is also part of "forbund" and "forsøge".

Padding

The inputs arrays need a standardised length, but our sequences were naturally of variable length, so we needed to pad our sequences with empty data. Every sequence shorter than a specified maximum length was padded with the appropriate amount of 0s.

[	...	16081	119	13218	10163	10181	10349	34364	11773
177	12532	119	13218	10163	19891	10452	43655	117	26578
38025	98348	10156	31078	44376	10858	119	102	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	...	]

Attention masks

The new padded data now had a lot of irrelevant data. We then needed to specify which data was actually relevant to model. We generated an attention mask that maps relevant data to 1 and irrelevant data to 0.

$$[1, 1, 1, ..., 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, ..., 0, 0, 0, 0, 0, 0]$$

Batching data

We needed to prepare the data before batching it. When the model is training, it should continually test itself to validate its direction. To do this, we needed to split our data into two subsets: training and validation. We used the train test split method of the sklearn library.[10] This function picks random samples from the data set and appends them to their appropriate training set or validation set. The test size parameter indicates the ratio between the training data and the test data. We picked 0.1 for validation and 0.9 for training because we wanted to train on as much data as possible.

The smaller the batch size is, the more batches will be produced. We experienced that with larger batch sizes, our model would run out of memory and stop before finishing. However, using a smaller batch size would lengthen the training time. We experienced no performance change based on the batch size however.

4.2 Pretrained BERT model

For this project, we used a pretrained model that we could fine-tune to our task, as opposed to training our own model, like we did in our Research project (Jespersen, Thomsen, & Thygesen, 2019). The pretrained model is trained on massive amounts of language data, which will hopefully improve the performance. We were using a pretrained BERT model, as it has shown interesting performances in the NLP field.

4.2.1 Regular BERT

The Bidirectional Encoder Representations from Transformers (BERT) model was developed by Google and delivers a new approach to Natural Language Processing. The model is a training technique that has been pre-trained on a large text corpus. This means that developers using the model only need to apply fine-tuning to their models which makes BERT relatively simple to use. The Transformer feature in BERT is a mechanism that learns contextual relations between words in a text. As opposed to word embeddings in earlier NLP techniques where each word is mapped to a unique vector of numbers, in contextual relations, words are mapped to number vectors according to the textual relations they are part of. This means that a word can have a different embedding depending on how the word interacts in the sentence which it is part of. For example, the word "bank" would have different embeddings in the two sentences "the bank was robbed yesterday" and "we had lunch on the river bank". In this way, contextual relations emulate some of the language understanding that humans possess.

[10] https://scikit-learn.org/

BERT is also trained bidirectionally, which is new to model training. In the training process, BERT then uses two approaches for the pre-training: Masked Language Modelling (MLM), and Next Sentence Prediction (NSP). MLM masks 15 % of the words in a text sequence by replacing these words with tokens and the model then predicts the masked words on basis of the non-masked words in the sentence (Horev 2018). In the NSP part, the model attempts to predict if a second sentence in a sequence of two sentences is subsequent to the first. Since this pre-training has already been performed, the user only needs to apply fine-tuning to the BERT model. The fine-tuning consists of stages where the sentences in the input data are split into tokens. These tokens are separated by separator tokens and then substituted with ID numbers, as shown in the figure and described in Section 4.1.

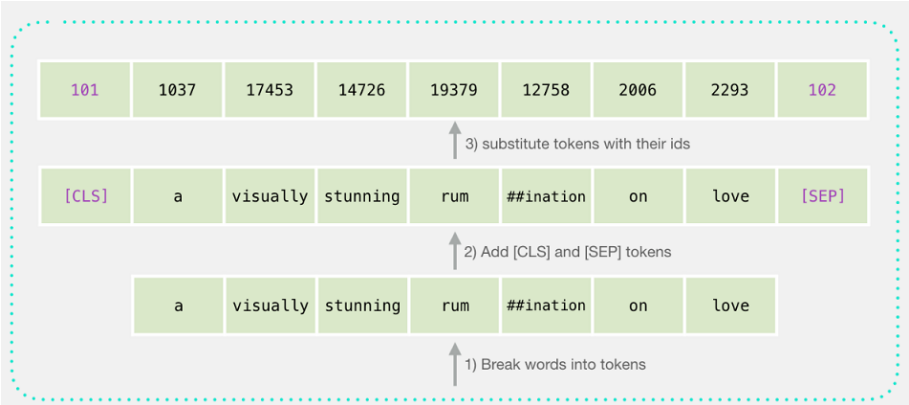


Figure 9: Tokenization in BERT model. (Alammar 2018)

A vector with the ID numbers is then passed to the BERT model which outputs output vectors for classification.

BERT contains a self-attention mechanism in its Transformer which enables encoding of concatenated text pairs such as sentence pairs in paraphrasing, hypothesis-premise pairs in entailment, or question-passage pairs in question answering. Sentence A and sentence B are processed in one sentence with separators in between. The token representations of the words and subwords are then fed to an output layer for tasks such as question answering, and the [CLS] representation is fed to an output layer for classification, such as entailment or sentiment analysis, which in our case would be the claim labelling (Devlin et al. 2019). The strength of BERT is that its fine-tuning is inexpensive, because the biggest part of the NLP job has been done in the pre-processing phase. This part is incorporated in the BERT model and therefore can be used just by applying the model in the user's code. The model can be used on both single-sentence tasks, such as sentiment analysis or text classification tasks, i.e. control of grammatical errors in text, or it can be used in two-sentence tasks, such as the FEVER task or question-answering tasks where the model seeks an answer, B, to a question, A.

4.2.2 Multilingual BERT

The BERT model is an English language model trained on English language corpora. In 2019, a Multilingual BERT model was released (T. Pires et al. 2019) which had been trained on

Wikipedia corpora in 104 languages, with a shared word piece vocabulary. One of the 104 languages is Danish, which enables us to implement this model in our project. The Multilingual BERT model does not need to be adjusted when applying the model on non-English language text - the model automatically chooses the correct language. *T. Pires et al.*, who tested the model on 16 languages, argue that Multilingual BERT is able to transfer between languages written in different scripts (i.e. alphabets), having no lexical overlap, which indicates that it captures multilingual representations very well, where English BERT is more dependent on overlap (T. Pires et al. 2019). The team investigates cross-language transfer and concludes that the training that have been conducted on single-language corpora for the multilingual model creates a shared space for all the languages involved. This not only ensures a high complexity for each language included in the model, it is also interesting in a broader perspective involving cross-language NLP.

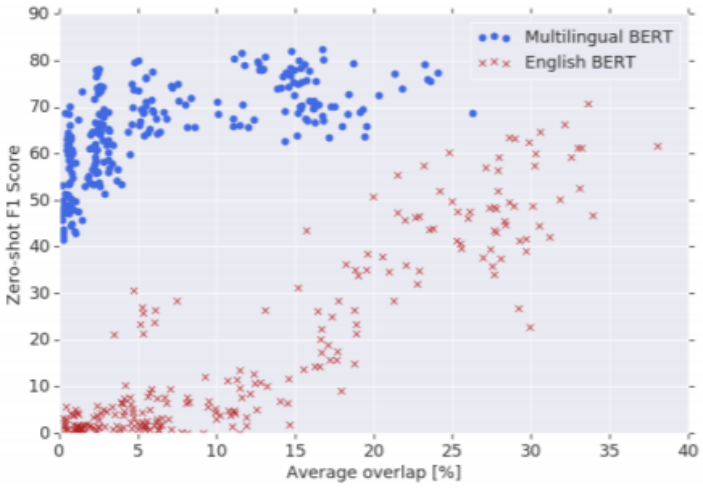


Figure 10: Named Entity Recognition F1 score versus entity word piece overlap among 16 languages (Pires et al 2019).

Figure 10 shows the F1 score for Named Entity Recognition versus entity word piece overlap among 16 languages for Multilingual BERT and English BERT respectively. This indicates that the performance of Multilingual BERT is not as dependent on overlap as English BERT.

A BERT model exclusively for Danish has been developed by the IT company BotXO (Møllerhøj 2019). However, since this model has not been tested thoroughly as of yet, we have chosen to apply the Multilingual BERT in our model.

4.3 Optimiser

With our data ready for training, we needed to configure the optimiser. The optimiser provides the functionality that calculates, evaluates, and changes the data of the internal weights of the model. There exist several different optimisers to choose from, each with their own set of parameters that can be configured to tune the model.

We tested 2 optimisers: BERTAdam and Stochastic Gradient Descent (SGD). Of the two, SGD

showed to be the most reliable for a consistent performance, where BertAdam on the other hand showed high results. We will discuss this in the Results section.

4.4 Scoring Metrics

In the original FEVER shared task, results are measured with a FEVER score. This score is based on two features, the evidence and the label for each claim, and it is calculated as the proportion of claims for which both correct evidence is returned and a correct label prediction is made. Since we have had a different approach to the annotation process, we will not be able to calculate a FEVER score as in the FEVER Task. Instead, we will use a weighted F1 score and a weighted accuracy score to compare our results. These metrics were also used in the FEVER Task.

4.5 Baseline

The original FEVER shared task used a data set of 185,445 entries based on English Wikipedia content with development and test data sets evenly distributed between SUPPORTED, REFUTED, and NOTENOUGHTINFO claims. This gave a balanced data set with a random baseline label accuracy of 33.33 % when ignoring the requirement for evidence. This corresponds to predicting NEI for every claim.

Split	SUPPORTED	REFUTED	NOTENOUGHINFO
Training	80,035	29,775	35,639
Dev	6,666	6,666	6,666
Test	6,666	6,666	6,666

Table 6: Distribution of claims in the original English FEVER task data set.

Label	SUPPORTED	REFUTED	NOTENOUGHINFO
Amount	1435	950	1625

Table 7: Distribution of claims in our Danish data set of 4010 entries.

In order to have a random baseline to compare our model against, we shuffled the labelling of the claims in our data set. This gave us a data set with 4010 entries with a completely random labelling. We then evaluated this data set in our model and got an accuracy score of 0.305 and an F1 score of 0.265. In the confusion matrix[11] of the random data set evaluation the model is evidently influenced by the distribution of SUPPORTED, REFUTED, and NOTENOUGHTINFO claims since the random baseline predicts 0 occurrences of the Refuted claims, and there are 950 Refuted claims compared to 1435 and 1625 of respectively SUPPORTED and NOTENOUGHTINFO claims.

[11] The confusion matrix is made up of three rows, representing the actual values, and three columns, representing the predicted values. A diagonal line from (1,1) through (2,2) to (3,3) shows the correctly predicted values.

	NEI	Ref.	Sup.
NotEnoughInfo	31	0	3
Refuted	47	0	13
Supported	76	0	30

Table 8: Confusion matrix for our model's random baseline.

We then used these baseline scores for comparison against the results of our model to check if the model performs better than a random baseline.

5 Results

5.1 Choosing Model Parameters

As previously mentioned, we chose to use the SGD optimiser for our model, since this optimiser should be especially useful for smaller data sets, and the BertAdam optimiser which is generally used in BERT models. Other parameters that we adjusted during the testing of our model are the number of epochs and the learning rate. Usually in BERT the number of epochs is set to 2-4 but we also experimented with increasing the number in order to get more training. When plotting the training and validation loss against the number of epochs, we can see if the increase in number of epochs influences our model. The following figure shows an example of such plotting.

Figure 11: Training and validation loss against number of epochs. The model runs 4 epochs, an SGD optimiser with a learning rate of 2e-5 and a momentum of 0.

In this model it is not useful to increase the number of epochs since the validation loss curve is rising and the training and validation loss curves intersect. When the validation loss curve exceeds the training loss curve, the model is overfitting. Plotting the training and validation loss like this enables us to check if the number of epochs is the right for our model or if it needs adjustment.

Apart from choosing optimiser and number of epochs, it is relevant to experiment with the learning rate as well since this is an important parameter. The learning rate is a hyperparameter with a value between 0.0 and 1.0. It is used in the training of the model where the learning rate adjusts the weights used in the training to help avoid getting stuck at a local minimum. If the learning rate is low, it will take longer time for the model to learn and the model will not perform optimally. On the other hand, if the learning rate is too high, the model can oscillate between inadequate solutions (Han et al. 2011).

5.2 Using Randomly Generated NotEnoughInfo-labelled Claims

Our model suffered from poor predictions of the claims labelled Not Enough Information (NotEnoughInfo). An example of such evaluation results is a NotEnoughInfo prediction of [7 25 13] from a confusion matrix, which means that only 7 NotEnoughInfo claims were correctly predicted, whereas 25 + 13 NotEnoughInfo claims were incorrectly predicted. This might be due to a smaller amount of these claims in the training data set which was initially at about 20 %. In order to improve the prediction of the NotEnoughInfo claims, we created an additional pool of 1000 randomly generated claim entities, as described in Section 3.4.2, where we shuffled claims and evidence from our data set and then labelled these claims with NotEnoughInfo. This, we hoped, would improve the prediction of the NotEnoughInfo claims.

When included, we got fairly good results on the NotEnoughInfo claims which increased the F1 score. One example is a confusion matrix with 66 correctly predicted and only 14 incorrectly predicted NotEnoughInfo claims and an F1 score of 0.718. The range of F1 scores generated with this approach was between 0.70 and 0.78. But this approach would likely teach the model that the NotEnoughInfo label only should be applied on a claim and a piece of evidence with no similarity. We therefore decided not to include the randomly generated NotEnoughInfo claims in the evaluation data set.

	NEI	Ref.	Sup.
NotEnoughInfo	66	7	5
Refuted	7	5	25
Supported	2	4	79

Table 9: Confusion matrix of data set with randomly generated claims. This evaluation mirrored the fact that the prediction data set contained randomly generated claims.

Hence, we continued running the test predictions with data sets without the randomly generated NotEnoughInfo claims, but still with these claims contained in the training set. This meant that the model was still trained in labelling claims with no similarity to its piece of evidence as NotEnoughInfo, only did we not use these claims for evaluation, and it had a significant effect on the prediction results. The distribution of correctly and incorrectly predicted NotEnough-Info claims changed to being almost randomly distributed with 10-23 correctly predicted and 3-20 incorrectly predicted, and an F1 score between 0.522 and 0.586. The result from this approach was better than before the addition of the randomly generated NotEnoughInfo claims

in the model training so we decided to keep these claims in the training of the model but avoid using them for evaluation.

5.3 Performance

5.3.1 SGD

One of our first models was with an SGD optimiser, a momentum of 0.5, a learning rate of 10e-3, and 4 epochs. With three runs, we got an average weighted F1 score of 0.521 and a weighted accuracy of 0.492. This is better than our random baseline results of an F1 score of 0.265 and an accuracy of 0.305 which indicates that our model is learning, although not perfectly. We could see from the confusion matrices that the model predicted the Supported category well, with between 2/3 and 4/5 correctly predicted values, but it also predicted a larger percentage of Supported values, 51-60 %, than there actually was, which was 44-46 %. This could be the result of a skewness in the training data set since most of the entries in this data set, 40 %, are labelled Supported compared to 23 % Refuted.

In the Refuted category in the prediction results, more claims are also incorrectly labelled Supported than Refuted. This indicates that the model does not interpret the Refuted claims well enough and assigns more claims to the category Supported than its actual statistical proportion.

We then tuned the momentum of the optimiser to 0.8. This gave an average weighted F1 score of 0.546 and an average weighted accuracy of 0.506, which means a slightly better result than with a momentum of 0.5 in the same model.

Optimiser, Momentum, Epochs	F1 avg.	F1 min.	F1 max.	Accuracy avg.	Acc. min.	Acc. max.
SGD, 0.5, 4	0.521	0.495	0.536	0.492	0.449	0.520
SGD, 0.8, 4	0.546	0.516	0.586	0.506	0.473	0.548

Table 10: Results for SGD optimiser with momentum of 0.5 and 0.8 and 4 epochs.

The main difference between these two models was that with a momentum of 0.8, the prediction of Refuted claims improved significantly. From a recall[12] of 0.28-0.36 for the model with a momentum of 0.5, to a recall of 0.43-0.50 when the momentum was increased to 0.8. On the other hand, the prediction of Supported claims deteriorated, from a recall of 0.71-0.84 for the model with a 0.5 momentum, to 0.70-0.77. For the NotEnoughInfo claims, recall was more dispersed.

[12]Recall is the number of true positives divided by the number of true positives plus false negatives (i.e. all actual positives).

Optimiser, momentum	SUPPORTED	REFUTED	NOTENOUGHINFO
SGD, 0.5	0.71 - 0.84	0.28 - 0.36	0.29 - 0.44
SGD, 0.8	0.7 - 0.77	0.43-0.46	0.21-0.44

Table 11: Recall for SGD optimiser with momentum of 0.5 and 0.8. Recall shows how many of all the variables in each category were correctly predicted.

Another important parameter is the number of epochs. The authors of BERT recommends 2-4 epochs (Devlin et al. 2019:14). Initially we set the number to 4 since this would generate more training of the model, and since we have a small data set, we expected that more training would improve the model. We therefore also experimented with 6 epochs, with a momentum of 0.6 and 0.8, respectively. The results are shown here:

Optimiser, Momentum, Epochs	F1 avg.	F1 min.	F1 max.	Accuracy avg.	Acc. min.	Acc. max.
SGD, 0.6, 6	0.532	0.466	0.592	0.492	0.449	0.568
SGD, 0.8, 6	0.590	0.507	0.665	0.546	0.490	0.585

Table 12: Results for SGD optimiser with momentum of 0.6 and 0.8 and 6 epochs.

As with the model with 4 epochs, the increase of momentum improved the result significantly.

Since J. Devlin et al. recommends using 2-4 epochs, we also experimented with a lesser number of epochs. Decreasing the number of epochs in the model with the SGD optimiser with 0.5 momentum and a learning rate of 10e-3 would be a reasonable comparison. As can be seen in the figure, decreasing the number of epochs aggravates the result. We believe this to be a consequence of the small data set since the original Fever shared task data set was of a considerable size and also because other researchers have described how a larger data set decreases the need for many epochs.

Optimiser, Momentum, Epochs	F1 avg.	F1 min.	F1 max.	Accuracy avg.	Acc. min.	Acc. max.
SGD, 0.5, 4	0.521	0.495	0.536	0.492	0.449	0.520
SGD, 0.5, 3	0.502	0.448	0.561	0.469	0.398	0.514
SGD, 0.5, 2	0.442	0.420	0.462	0.445	0.435	0.454

Table 13: Results for SGD optimiser with momentum of 0.5 and epochs of 4, 3, and 2.

Since the learning rate is an important parameter, we also experimented with this. Initial small tests led us to a higher learning rate than what was mostly used in the original FEVER shared task, where most teams set a learning rate of 0.001 (10e-4). When testing with SGD, 4 epochs,

0.8 momentum, and a learning rate of 10e-4, we got a weighted F1 score of 0.335 and a weighted accuracy of 0.363, which made us abandon searching for lower learning rates, although we also tested on the learning rate found to work well with the BertAdam optimiser. For our SGD optimiser, we found the best results with a learning rate of 0.01 (10e-3).

In this final table for our SGD optimiser we compare the different models performed with an SGD optimiser.

Optimiser, Momentum, Epochs, Learning rate	F1 average	Accuracy average
SGD, 0.5, 4, 10e-3	0.521	0.492
SGD, 0.8, 4, 10e-3	0.546	0.506
SGD, 0.6, 6, 10e-3	0.532	0.492
SGD, 0.8, 6, 10e-3	0.590	0.546
SGD, 0.5, 2, 10e-3	0.442	0.445
SGD, 0.5, 3, 10e-3	0.402	0.469
SGD, 0.5, 4, 10e-4	0.327	0.386
SGD, 0.8, 4, 2e-5	0.217	0.326

Table 14: Results for SGD optimiser.

5.3.2 BertAdam

We also wished to investigate the BertAdam optimiser which has been used by other researchers in Fact Extraction and Verification. We started out with a model with the same hyperparameters as with the SGD optimiser: 4 epochs and a learning rate of 10e-3. This gave us an average weighted F1 score of 0.277 and an average weighted accuracy of 0.333 which is comparable to our random baseline and therefore a quite bad result. Hence, we increased the learning rate to 10e-4 which also did not provide a satisfactory result.

Optimiser, Epochs, Learning rate	F1 average	Accuracy avg
BertAdam, 4, 10e-3	0.277	0.333
BertAdam, 4, 10e-4	0.059	0.333

Table 15: Results for BertAdam optimiser with 4 epochs and a learning rate of 10e-3 and 10e-4 respectively.

After adjusting the learning rate to 2e-5, we reached a better result. With the learning rate set, we ran models with 2, 3, and 4 epochs respectively. The results can be seen in the following

table.

Optimiser, Epochs, Learning rate	F1 avg.	F1 min.	F1 max.	Acc. avg.	Acc. min.	Acc. max.
BertAdam, 2, 2e-5	0.387	0.073	0.569	0.445	0.333	0.509
BertAdam, 3, 2e-5	0.521	0.344	0.597	0.502	0.412	0.563
BertAdam, 4, 2e-5	0.607	0.566	0.671	0.563	0.540	0.632

Table 16: Results for BertAdam optimiser and epochs of 2, 3, and 4.

When running the model with 2 and 3 epochs, we experienced cases when the model did not predict entire categories. For example, one evaluation did not predict any REFUTED claims, and another evaluation did not predict neither Refuted nor Supported claims, which resulted in low F1 and accuracy scores. This taken into account, we continued to test the BertAdam with 4, 5, and 6 epochs.

Optimiser, Epochs, Learning rate	F1 avg.	F1 min.	F1 max.	Acc. avg.	Acc. min.	Acc. max.
BertAdam, 4, 2e-5	0.607	0.566	0.671	0.563	0.540	0.632
BertAdam, 5, 2e-5	0.606	0.574	0.673	0.554	0.510	0.624
BertAdam, 6, 2e-5	0.580	0.465	0.635	0.524	0.399	0.600

Table 17: Results for BertAdam optimiser and epochs of 4, 5, and 6

The evaluations with 4 and 5 epochs gave quite similar results with average F1 scores of 0.607 and 0.606 respectively. Also the maximum evaluation results were similar with 0.671 and 0.673 respectively. This means that when looking at the average F1 score, the model with 4 epochs is slightly better than the one with 5, but looking at single results, the model with 5 epochs obtained the highest F1 score overall. When increasing the number of epochs to 6, evaluation results deteriorated. Hence, good results must be found within the range of 4-5 epochs for this model.

When comparing the different models, the best result we have had is with the BertAdam optimiser, with 4-5 epochs, and a learning rate of 2e-5. The best average F1 and average accuracy scores were obtained with 4 epochs, while the best single F1 score, 67.30 %, was found with 5 epochs.

5.3.3 Results Comparison

As can be seen from the table below, our model reaches a maximum F1 score of 67.30 % with the BertAdam optimiser with 5 epochs, and a maximum F1 score of 66.47 % with the SGD

optimiser. The highest average F1 score and average accuracy score is reached by the BertAdam optimiser with 4 epochs with 60.69 % and 56.27 %, respectively. All values are weighted.

Since our training and evaluation data sets are never equal due to the random sampling of the evaluation data set, we have evaluated the model 4-6 times for each variant to obtain average values for both F1 and accuracy scores. This means that the maximum values that we obtained should not be interpreted as a result that will always appear when evaluating on the model. We built the model with the random sampling such that we would not by chance obtain results that were affected by special characteristics of the evaluation data set. In this way, we hoped to achieve qualitatively better results. It is therefore also our opinion that the average values in the results are more significant in regards to the model performance than the maximum (or minimum) values.

Optimiser, Epochs,	F1 avg.	F1 min.	F1 max.	Acc. avg.	Acc. min.	Acc. max.
BertAdam, 4	0.6069	0.5662	0.6710	0.5627	0.5402	0.6318
BertAdam, 5	0.6063	0.5735	0.6730	0.5537	0.5096	0.6240
SGD, 6	0.5904	0.5071	0.6647	0.5464	0.4903	0.5848

Table 18: Best results of our model with a BertAdam and an SGD optimiser.

It would have been preferable to have other results from similar fact extraction and verification models to compare our results with. The FEVER shared task listed the FEVER results as F1 scores, and not weighted F1 scores which makes a direct comparison problematic. Moreover, the FEVER task was undertaken with English data, and we cannot know how these FEVER models would react if they were trained on and evaluated Danish data with the Multilingual BERT. We have also applied a different annotation process which renders superfluous the search for evidence sentences as in the FEVER task.

Notwithstanding this, it is interesting to see that the state of the art F1 score from the FEVER task, 64.85 % (Thorne et al. 2018: 3), is somewhat similar to our model's, and the evaluation results from our model still provide information about to which extent it is feasible to fine-tune a BERT model for Danish.

We do believe that our model obtains decent results but also that its performance is hindered by the small size of the data set. Compared to the FEVER data set of 185,455 entries, our data set with 4010 entries is minimal. It could be highly interesting to see how our model would perform with a much larger data set.

5.3.4 Results Stability

The results that we get from our model diverge somewhat. Each time we evaluate a data set for evaluation, we extract a data set of 200 random entities from the main data set. This means that we never get the same results since the data set that the model trains on, as well as the evaluation data set, is different each time we run the model. There are though some

results that are more stable than others. We experienced that when the number of epochs was decreased, the model sometimes disregarded an entire category or two such that only one or two of the categories Supported, Refuted, or NEI was predicted, which gave results similar to a random baseline. This means that even though Devlin et al. recommends a number of epochs of 2-4, using a small data set like we have done, the model gives too unstable results when applying less than 4 epochs. Besides this, there seemed to be an indication that our results with the SGD optimiser were more divergent than with the BertAdam optimiser. For the best achieving models, the standard deviation of the SGD model was 0.0559, compared to 0.0318 for the BertAdam model. We ascribe the fluctuations in results to our small data set and the fact that the data sets used for training and evaluation were never the same. Possibly, the BertAdam optimiser gives less divergent results, but this would have to be tested on a larger scale, with a larger data set and more evaluation rounds.

5.4 False Prediction Analysis

In order to understand the model better, we examined the errors it made. We therefore analysed on the results from the evaluation of the model by looking at the confusion matrix from a single evaluation. Here we trained the model with the data set of 3010 sentences and evaluated with a test data set of 200 sentences. The numbers in the confusion matrix give us information on how the model predicts.

	NEI	Ref.	Sup.
NotEnoughInfo	10	7	11
Refuted	8	17	29
Supported	6	9	103

Table 19: Confusion matrix

Here, we can see the distribution of the prediction results. In the first row, 10 claims have been correctly predicted as NotEnoughEvidence, whereas 7 claims have incorrectly been predicted as Refuted while having the true value NotEnoughEvidence, and 11 claims have incorrectly been predicted as Supported while having the true value NotEnoughEvidence. In the second row, 17 claims have correctly been predicted as Refuted, while 8 (29) have incorrectly been predicted as NotEnoughEvidence (Supported) while having the true value Refuted. In the third row, 103 claims have correctly been predicted as Supported, while 6 (9) have incorrectly been predicted as NotEnoughEvidence (Refuted) while having the true value Supported.

This means that 10 out of 28 NotEnoughEvidence-claims, 17 out of 54 Refuted-claims, and 103 out of 118 Supported-claims have been predicted correctly. Our model therefore seems to have a good prediction on the Supported claims, while not on the NotEnoughEvidence and Refuted. In the following we will analyse those claims that have been incorrectly predicted.

(PV,AV)	(0,1)	(0,2)	(1,0)	(1,2)	(2,0)	(2,1)
Count	8	6	7	9	11	29

Table 20: Distribution of incorrectly predicted claims between Predicted Value (PV) and Actual Value (AV), where 0 = NotEnoughInfo*, 1 =* Refuted*, and 2 =* Supported*.*

5.4.1 Reasons for Incorrect Predictions

Model not good enough	50
Annotation error	6
Interpretation required	6
Inference/world knowledge required	6
Calculation required	2

Table 21: Count distribution of evaluated reason for incorrect prediction

Here we have divided the incorrectly predicted claims into categories describing the reasons for the incorrect predictions. The main reason for 50 claims out of the 70 is that the model has simply not been good enough. Other reasons have been errors from the annotation, where the actual value has been wrong when compared to the evidence, or evidence where labelling of the claim relies on inference or an interpretation of the evidence, or language or world knowledge, or even mathematical calculations. Examples of these claims are:

```
Claim:  "Finland var allieret med USA"
Evidence:  "Under anden verdenskrig var Finland allieret med aksemagterne,
Nazityskland, Italien og andre pro nazistiske stater som bla.  Ungarn og
Rumænien.  Finland kæmpede to gange mod Sovjetunionen:  Første gang i
vinterkrigen 1939{40, efter at Sovjetunionen havde angrebet landet, og
igen i fortsættelseskrigen i 1941-44 i kølvandet på operation Barbarossa,
i hvilken Nazityskland invaderede Sovjetunionen.  I 872 dage belejrede
tyskerne Leningrad, Sovjetunionens næststørste by Belejringen af Leningrad
resulterede i omkring en million døde af byens befolkning.  Finske tropper
kontrollerede nogle områder omkring byen, men nægtede selv at angribe
eller lade tyskerne bruge disse områder som basis for deres angreb;
det er fortsat et kontroversielt emne, om de finske tropper primært var
behjælpelige med belejringen eller var modstræbende i dette.  Efter at
have kæmpet mod en stor sovjetisk offensiv i juni-juli 1944 fik finnerne
våbenstilstand, men snart efter fulgte Laplandskrigen i 1944{45, hvor
Finland trængte tyske styrker ud af Nordfinland."
Predicted label:  NotEnoughInfo
Actual label:  Refuted
```

Here, there is no mentioning of "USA" in the evidence, and it is not mentioned that Finland was or was not allied with the USA. The correct labelling would be inferred by a human reader, since a human reader with some world knowledge would know that a country allied with Nazi

Germany would probably not also have been allied with the USA. Thus, the correct labelling of this claim requires both world knowledge (that the USA was fighting Germany during WWII) and inference (that if a country is allied with one side in a war, it is supposedly not also allied with the opposite side).

```
Claim:  "Vendsyssel ligger i Sønderjylland"
Evidence:  "Vendte man blikket mod øst i disse yderste, nordlige egne, så
man, at landskaberne ud mod Kattegat havde deres helt eget præg.  I det
østlige Vendsyssel, fra Hammer til Tolne Bakker rejste sig et voldsomt
formet, skovklædt højland.  Og længere nede ad den jyske østkyst tårnede
de ligeledes skovklædte højder fra istidens morænedannelser sig op over de
dybt indskårne fjorde, bakke op og bakke ned helt ned til de milde egne i
den vestlige ende af Østersøen."
Predicted label:  SUPPORTED
Actual label:  REFUTED
```

In this example, the correct labelling of the claim cannot directly be read from the evidence but must rely on inference. First, the reader must have enough language or world knowledge to know that "Sønderjylland" lies to the south of Denmark. Then, it must be inferred that the parts "i disse yderste, nordlige egne" and "I det østlige Vendsyssel" belong together, and that therefore, there is a contradiction between "Sønderjylland" lying to the south and "Vendsyssel" to the north.

```
Claim:  "Polen mistede områder til Preussen i 1700-tallet"
Evidence:  "Hertil kom, at Preussen erhvervede Vestpreussen ved Polens
første deling i 1772, bla.  Posen og Danzig ved Polens anden deling i 1793
og yderligere store områder, bla.  Warszawa, ved Polens tredje deling i
1795."
Predicted label:  NOTENOUGHINFO
Actual label:  SUPPORTED
```

In order to label this claim as Supported, the reader must apply some interpretation of the evidence, since the evidence does not directly support the claim. The evidence states that Preussia acquired West Preussia at Poland's first division in 1772, whereas the claim states that Poland lost territory to Preussia in the 18th century. But the evidence does not prove the claim – one could imagine that Poland had already lost the territory to a third part before Preussia obtained it, or that the territory was not lost but perhaps sold off.

```
Claim:  "Kongens Nytorv har eksisteret i 1300 år"
Evidence:  "Indtil for nylig var det ældste spor efter bymæssig bebyggelse
i Københavnsområdet inden for voldene fra omkring år 1000, hvor der er
fundet spor fra et mindre fiskerleje dér, hvor København ligger i dag.
Fiskerlejet lå lige nord for Københavns Rådhus omkring Mikkel Bryggers
Gade, der dengang lå ud til havet.  Men i forbindelse med udgravning af
Metroen har man fundet spor af bådebroer ved Gammel Strand, der daterer sig
helt tilbage til omkring år 700.  Ved udgravningen til metrostationen ved
Kongens Nytorv har man endvidere fundet spor af en gård fra vikingetiden.
Første gang forløberen til København under navnet "Havn" nævnes i kilderne,
```

er i forbindelse med et søslag mellem Svend Estridsen og den norske konge Magnus den Gode i 1043. Derefter er der tavshed om byens skæbne i de næste 120 år"
Predicted label: REFUTED
Actual label: NOTENOUGHINFO

In this example, if the claim's statement is not directly stated in the evidence, some calculation of numbers is required to label it correctly. We must know what year we have now; find the year coupled with the named entity "Kongens Nytorv"; subtract the two numbers; and compare the result with 1300.

Incorrect predicted claims may be due to mislabelled data, which means that the model might not always be wrong when evaluating these claims. This is relevant in the case of the 20 out of the 70 incorrectly predicted claims, where the model might have predicted these claims correctly if the claims were better annotated. This emphasises the importance of a thorough annotation process with exact annotation guidelines to ensure that annotators will annotate correctly and consistently.

5.4.2 Unsatisfactory Model Performance

(PV,AV)	(0,1)	(0,2)	(1,0)	(1,2)	(2,0)	(2,1)
Count	6	3	6	4	6	25

Table 22: Distribution of claims where the model has not been good enough. 0 = NEI; 1 = Refuted; 2 = Supported.

Here we look at the claims where we assess that the model has not been good enough in labelling them. We wish to see if there is a pattern in the distribution, and if so, what this could mean. Out of the 50 claims where we assessed the model not to be good enough, 50 % of them was predicted as Supported, when the actual label was Refuted.

These two examples show claims that we would think that the model could predict correctly as Supported, but that in fact were labelled Refuted by the model. We can only ascribe this fact to the small size of our data set.

Claim: "der findes 32 regioner i Chile"
Evidence: "Chile er opdelt i 15 regioner (spansk región), der hver ledes af en intendant. Regionerne er opdelt i provinser, hvor en guvernør står i spidsen. Intendanter og guvernører udpeges af landets præsident. Det nederste administrative niveau er kommuner, som administreres af kommunalbestyrelser med borgmestre som øverste ledere. Kommunalbestyrelserne vælges for fire år ad gangen."
Predicted label: SUPPORTED
Actual label: REFUTED

Claim: "Blood work er en roman skrevet af Eastwood"
Evidence: "I 2002 spillede Eastwood en tidligere FBI-agent, der jager

```
en sadistisk morder, spillet af Jeff Daniels i thrilleren Blood Work, som
er løst baseret på en roman af samme navn fra 1998 af Michael Connelly.
Filmen var en kommerciel fiasko, og den indtjente blot $26,2 millioner, ca
halvdelen af dens budget.  Eastwood vandt dog Future Film Festival Digital
Award ved Venedig Film Festival for denne film."
Predicted label:  SUPPORTED
Actual label:  REFUTED
```

Other examples show that what we might see as a clear error from the model in fact shows our inclination to forget how humans infer information. In this example, the claim says that the World Cup was held in Argentina in 1962. The evidence says that Chile was host to the games at that same year. A human reader would infer that there is a contradiction because the World Cup cannot be held in two countries in the same year. But the model does not have this world knowledge. Again, this puts more importance on the annotation process where we need to ensure that the claims are labelled correctly for the model to predict correctly.

```
Claim:  "der blev afholdt VM i Argentina i 1962"
Evidence:  "Den mest populære sportsgren i Chile er fodbold.  Chile
har deltaget i otte VM-slutrunder, heriblandt VM i 1962, som landet
var vært for, og hvor landsholdet fik det hidtil bedste resultat med en
tredjeplads.  Andre gode resultater opnået af holdet er fire finalepladser
i Copa América og bronzemedaljer ved OL i 2000.  Den bedste række i
landets turneringssystem er Primera División, som i 2013 af IFFHS (Det
Internationale Forbund for Fodboldhistorie og -Statistik) blev rangeret som
19."
Predicted label:  NOTENOUGHINFO
Actual label:  REFUTED
```

6 Comparison with LSTM Model

6.1 RNN

Recurrent Neural Network (RNN) is a model architecture that uses sequential memory to train deep neural networks. As the model receives the input data, it retains the information throughout training. The training is broken into several epochs, where for each epoch it will learn, evaluate, and modify internal values weights. The order of which the model receives data greatly affects the weights. This is why RNN models are usually used in the NLP field, as "WHAT TIME IS IT" will be evaluated differently from "TIME WHAT IT IS". When a model has run an epoch, it attempts to predict labels on a validation data set. The result of this validation produces a gradient value. It will then go through steps of the model in reverse and adjust the values of the internal weights with the gradient value, which is called back propagation.

6.2 LSTM architecture

The problem with an RNN model is that as the model trains on the data, the rate at which the model learns can either become very small, called vanishing gradients, or very high, called

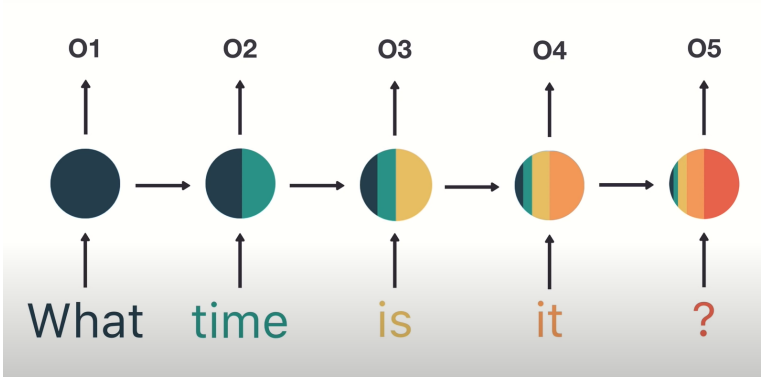


Figure 12: A graphical illustration of how the initial input becomes less important for every subsequent input. Image retrieved from: https://www.youtube.com/watch?v=LHXXI4-IEns

exploding gradients. If the gradient produced for the back propagation is very small or very high, it can cause consecutive gradients to also be very small or very high. This will have an exponential effect on the internal weights, causing them to be either almost zero, or out of control. Hochreiter & Schmidhuber (1997) introduced the Long Short Term Memory architecture that addresses these issues (Hochreiter & Schmidhuber 1997: 1735-1780). In the figure above, we can see that the model passes on information from the previous layers to the upcoming layers. However, as more information is passed through, the information from the first layers are diminishingly low. In a way this causes an RNN to suffer from short-term memory, as older entries become less relevant to the word it is trying to predict.

An LSTM model addresses this problem by selectively forgetting information that it deems irrelevant. This will cause the model to remember important information for a longer time, or long term, thus the name.

6.3 Comparison

In our Research Project (Thomsen, Thygesen, Jespersen 2019), we set up a simple LSTM-based model for the FEVER task. The model at the time performed relatively poorly, with a prediction accuracy of 0.4. During this project we decided to revisit the LSTM model with our new data set to test whether we could get an improved performance. After a few runs, we determined that the model trained poorly on our data. The best result we got were 18 correctly predicted claims out of 100. We conclude that our current model is much more fit for the FEVER task.

7 Conclusion

In this project, we have fine-tuned a BERT model for Danish using Multilingual BERT. We have created a data set for fact extraction and verification of 4010 entries which we have annotated such that the classification of the claims could be used by the model to evaluate whether claims in a different evaluation data set were either Supported, Refuted, or if there was not enough information to correctly label these. Moreover, we have created a claim-generating web application for Danish such that the annotation process became faster and smarter.

We achieved a highest weighted F1 score of 67.30 % and a highest weighted accuracy score of 63.18 %. This was achieved with a BertAdam optimiser of 5 and 4 epochs, respectively. Using an SGD optimiser provided results almost on level with the results from the BertAdam optimiser.

When evaluating some of those claim entities that the model predicted wrong, we have seen that some of the incorrectly labelled claims are due to annotation errors, while most are due to the model not performing well enough. As improvement to these shortcomings, we suggest 1) A better annotation process with focus on stricter annotation rules and guidelines; 2) A much larger data set in order for the model to have more training material. With these factors in place, we believe the performance of our model would improve considerably.

8 References

References

[1] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[2] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2018, November). Proceedings of the First Workshop on Fact Extraction and VERification (FEVER). In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER).

[3] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a Large-scale data set for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 809-819).

[4] Kirkedal, A., Plank, B., Derczynski, L., & Schluter, N. (2019). The Lacunae of Danish Natural Language Processing. In Proceedings of the 22nd Nordic Conference on Computational Linguistics (pp. 356-362).

[5] Strømberg-Derczynski, L., Baglini, R., Christiansen, M. H., Ciosici, M. R., Dalsgaard, J. A., Fusaroli, R., ... & Ladefoged, C. (2020). The Danish Gigaword Project. arXiv preprint arXiv:2005.03521.

[6] Soleimani, A., Monz, C., & Worring, M. (2020, April). BERT for Evidence Retrieval and Claim Verification. In European Conference on Information Retrieval (pp. 359-366). Springer, Cham.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT?. arXiv preprint arXiv:1906.01502.

[9] Ghelani, S. (2019). Breaking BERT Down A complete breakdown of the latest milestone in NLP [Blog post]. Retrieved from: https://towardsdatascience.com/breaking-bert-down-430461f60efb

[10] Derczynski, L. (2019). Simple Natural Language Processing Tools for Danish. arXiv preprint arXiv:1906.11608.

[11] Kendall-Taylor, A., Frantz, E., & Wright, J. (2020). The Digital Dictators: How Technology Strengthens Autocracy. Foreign Aff., 99, 103.

[12] Møllerhøj, J. (2019). BotXO has trained the most advanced Danish BERT model to date. https://www.botxo.ai/blog/Danish-bert-model/

[13] Horev, R. (2018). BERT Explained: State of the art language model for NLP. [Blog post] Retrieved from: https://towardsdatascience. com/bert-explained-state-of-the-artlanguage-model-for-nlp-f8b21a9b6270

[14] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2019, November). Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER). In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER).

[15] Sanh, V. (2019, August 28). Smaller, faster, cheaper, lighter: Introducing distilbert, a distilled version of bert [Blog post]. Retrieved from: https://medium.com/huggingface/distilbert-8cf3380435b5

[16] McCormick, C., Ryan, N. (2019, November 5). GLUE Explained: Understanding BERT Through Benchmarks [Blog post]. Retrieved from: https://mccormickml.com/2019/11/05/GLUE/

[17] Heinzerling, B. (2020). NLP's Clever Hans Moment has Arrived. Journal of Cognitive Science, 21(1), 159-167.

[18] Thomsen, D., Thygesen, M., and Jespersen, S. (2019). Fact Extraction and Verification in Danish. Research Project 2019. IT University of Copenhagen

[19] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[20] Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

[21] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. A large annotated corpus for learning natural language inference.

[22] Anoop, K., Gangan, M. P., Deepak, P., & Lajish, V. L. (2019). Leveraging heterogeneous data for fake news detection. In Linking and Mining Heterogeneous and Multi-view Data (pp. 229-264). Springer, Cham.

[23] Zulkifli, H. (2018). Understanding learning rates and how it improves performance in deep learning. Towards Data Science, 21, 23.

[24] Alammar, J. (2019, November 26). A Visual Guide to Using BERT for the First Time [Blog post]. Retrieved from: http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

[25] Foong, N. (2019, June 11). Beginner's Guide to BERT for Multi-classification Task [Blog post]. Retrieved from: https://towardsdatascience.com/beginners-guide-to-bert-for-multi-classification-task-92f5445c2d7c

[26] Tenney, I., Das, D., & Pavlick, E. (2019, July). BERT Rediscovers the Classical NLP Pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4593-4601).

[27] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., ... & Pavlick, E. (2018). What do you learn from context? Probing for sentence structure in contextualized word representations.

[28] Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6, 587-604.

[29] Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., & Koychev, I. (2017, September). Fully Automated Fact Checking Using External Sources. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017 (pp. 344-353).

[30] Schudson, M., & Zelizer, B. (2017). Fake News in context. AA. VV., Understanding and Addressing the Disinformation Ecosystem, 1-4.

[31] Wardle, C. (2019). FIRST DRAFT'S ESSENTIAL GUIDE TO Understanding Information Disorder. First Draft.

[32] Alemanno, A. (2018). How to Counter Fake News? A taxonomy of anti-fake news approaches. European Journal of Risk Regulation, 9(1), 1-5.

[33] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[34] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv, arXiv-2004.

[35] Watts, D. J., & Rothschild, D. M. (2017). The minority report on the fake news crisis:(spoiler alert: it's the real news). AA. VV., Understanding and Addressing the Disinformation Ecosystem, 23-37.

9 Appendix

9.1 Data Statement

In this Data Statement we have followed Bender & Friedman's proposed schema for academic papers introducing new data sets (Bender & Friedman 2018: 590-591).

A. CURATION RATIONALE As basis for the claim-and-evidence data set created in the annotation process, we included content from Danish Wikipedia (https://da.wikipedia.org/wiki/Forside) and the Danish online encyclopedia Den Store Danske (https://lex.dk/). From Danish Wikipedia, we chose articles that were either good (plus-marked) or excellent (star-marked). From both Wikipedia and Den Store Danske we selected articles in categories such as countries (Finland, Germany, USA, Israel), historical events (World War II), important persons (Winston Churchill), animals, and religion. Selection criteria were that the articles must contain a substantial amount of facts which would make the annotation process feasible.

B. LANGUAGE VARIETY Since the two selected text sources are encyclopedias, the language used in the articles does not vary but is kept in a formal, academic Danish.

C. SPEAKER DEMOGRAPHIC As stated in B., since the language is purely academic, it contains no further information on speaker demographics.

D. ANNOTATOR DEMOGRAPHIC The authors of this paper have been the sole annotators. Age: author 1: 27 years; author 2: 41 years Gender: author 1: Male; author 2: Female Ethnicity: author 1: Ethnic Danish; author 2: Ethnic Danish Native language: author 1: Danish; author 2: Danish Socioeconomic status: author 1: Middle class; author 2: Middle class Training in linguistics/other relevant discipline: author 1: None; author 2: Work in lexicography

E. SPEECH SITUATION Not applicable.

F. TEXT CHARACTERISTICS Genre: Encyclopedia Topics: History, Fauna, Geography

G. RECORDING QUALITY Not applicable.

H. OTHER Apart from reliability and the amount of facts in their content, articles have been selected out of preferences of the authors, such as an interest in historical events and geography.

I. PROVENANCE APPENDIX Not applicable.

9.2 Incorrectly Predicted Claims

9.2.1 Predicted NEI, actual Refuted

Finland var allieret med USA

Under anden verdenskrig var Finland allieret med aksemagterne, Nazityskland, Italien og andre pro nazistiske stater som bla. Ungarn og Rumænien. Finland kæmpede to gange mod Sovjetunionen: Første gang i vinterkrigen 1939–40, efter at Sovjetunionen havde angrebet landet, og igen i fortsættelseskrigen i 1941-44 i kølvandet på operation Barbarossa, i hvilken Nazityskland invaderede Sovjetunionen. I 872 dage belejrede tyskerne Leningrad, Sovjetunionens næststørste by Belejringen af Leningrad resulterede i omkring en million døde af byens befolkning. Finske tropper kontrollerede nogle områder omkring byen, men nægtede selv at angribe eller lade tyskerne bruge disse områder som basis for deres angreb; det er fortsat et kontroversielt emne, om de finske tropper primært var behjælpelige med belejringen eller var modstræbende i dette. Efter at have kæmpet mod en stor sovjetisk offensiv i juni-juli 1944 fik finnerne våbenstilstand, men snart efter fulgte Laplandskrigen i 1944–45, hvor Finland trængte tyske styrker ud af Nordfinland. 0 1

Analysis: There is no mentioning of "USA" in the evidence, and it is not mentioned that Finland was or was not allied with the USA. This would be inferred by a human reader, since a human reader would know that a country allied with Nazi Germany would probably not also be allied with the USA. So, for the model, there is not enough evidence to refute the claim.

Reunión har altid været selvstændigt

Kysten er stejl og uden større bugter, og koralrev er kun svagt udviklet. Klimaet er tropisk med meget stor nedbør (2500-7700 mm om året); på nordvestkysten dog kun 400 mm Réunion rammes hyppigt af voldsomme tropiske cykloner; historisk har det været med til at sinke koloniseringen af øen, og i flere omgange i 1800-t. blev øens kaffeplantager ødelagt for ikke siden at blive genetableret. 0 1

Analysis: The claim can be inferred to be refuted.

Islands Brygge blev grundlagt i 1995

Der blev bygget beboelsesejendomme i stedet for det planlagte erhvervsbyggeri, da beboelsesejendomme var en bedre forretning for bygherrerne. Mange af ejendommene blev tegnet af arkitekten Thorvald Gundestrup. Med etableringen i 1905 er bydelen Islands Brygge sammen med Amagerbro de sidst grundlagte af brokvartererne. 0 1

Analysis: The model is not good enough.

mårhunden har altid levet i Finland

Polarræven, der tidligere var meget almindelig over store dele af landet, blev næsten udryddet af pelsjægere i begyndelsen af det 20. århundrede. Den røde ræv er helt op til nutiden meget udbredt, og i de seneste årtier har mårhunden bredt sig fra Rusland.. 0 1

Analysis: The model is not good enough.

der blev afholdt VM i Argentina i 1962

Den mest populære sportsgren i Chile er fodbold. Chile har deltaget i otte VM-slutrunder, heriblandt VM i 1962, som landet var vært for, og hvor landsholdet fik det hidtil bedste resultat med en tredjeplads. Andre gode resultater opnået af holdet er fire finalepladser i Copa América og bronzemedaljer ved OL i 2000. Den bedste række i landets turneringssystem er Primera División, som i 2013 af IFFHS (Det Internationale Forbund for Fodboldhistorie og -Statistik) blev rangeret som 19. 0 1

Analysis: The model is not good enough.

preusserne var tyskere

Navnet Preussen stammer fra den baltiske folkestamme preusserne, der i middelalderen beboede området mellem floden Wisła (tysk Weichsel) i Polen og floden Nemunas (tysk Memel) i Litauen og Hviderusland. Preussen opstod som hertugdømme i 1525 ved en overenskomst mellem den polske kong Sigismund 1 Stary og Den Tyske Ordenshøjmester, der derved blev hertug i en protestantisk stat under polsk lenshøjhed. På grundlag af den enevældige stat, som Frederik Vilhelm den Store af Brandenburg havde opbygget efter Trediveårskrigen, blev kurfyrstendømmet Brandenburg i 1701 af den tysk-romerske kejser ophøjet til kongedømme og tog navn efter hertugdømmet Preussen. Under hans søn Frederik Vilhelm 1, der regerede 1713-40, skabtes grundlaget for Preussens senere position som europæisk stormagt. 0 1

Analysis: The model is not good enough.

Palæstina hørte under Israel

Palæstina hørte fra 1517 under Osmannerriget, men erobredes i 1917 af Storbritannien. Samme år tilkendegav briterne i Balfourdeklarationen, at de ville arbejde for et jødisk hjemland i området. På fredskonferencen i San Remo i 1920 blev Palæstina britisk mandatområde under Folkenes Forbund; Transjordanien udskiltes i 1922. Ønsket om et jødisk hjemland i Palæstina var blevet aktualiseret af den antisemitisme og nationalisme, der herskede i Europa i slutningen af 1800-t. 0 1

Analysis: The model is not good enough.

Krimkrigen fandt sted i 1990

Da Rusland invaderede den tyrkiske del af Balkanhalvøen i 1853, fik frygten for russisk dominans af Middelhavet og Mellemøsten Storbritannien og Frankrig til at invadere Krim-halvøen for at ødelægge den russiske flåde. Den efterfølgende Krimkrig (1854–1856), hvori der blev benyttet nye teknikker fra moderne krigsførelse, var den eneste globale krig, som blev udkæmpet mellem Storbritannien og en anden imperiemagt under Pax Britannica. Denne krig gik hen og blev et dundrende nederlag for Rusland, som med 1 mio. 0 1

Analysis: The model is not good enough.

9.2.2 Predicted NEI, actual Supported

Øresundsbroen ligger i københavn

Med bygningen af Øresundsbroen i 2000 blev København centrum for en ny by-region, nemlig metropolregionen kaldet Øresundsregionen. Området umiddelbart omkring Øresund, det vestlige Skåne med Malmø og Helsingborg og det østlige Sjælland har i alt cirka 3 millioner indbyggere, mens den samlede Øresundsregion, hvortil hele Sjælland, Lolland-Falster samt hele Skåne medregnes, har omkring 4 millioner indbyggere. 0 2

Analysis: Error. The claim should be labelled 0.

Lauenburg var et hertugdømme

de slesvigske krige, to dansk-tyske krige udkæmpet 1848-51 og 1864. Baggrunden for krigene var de nationale og sproglige modsætninger mellem de danske og tyske befolkningsdele i det danske monarki, der efter Wienerkongressen (1814-15) foruden de oversøiske besiddelser bestod af kongeriget Danmark og hertugdømmerne Slesvig, Holsten og Lauenburg; modsætninger, som især forstærkedes fra 1830'erne. Krigen, som også kaldes Treårskrigen, var både en borgerkrig inden for rammerne af det danske monarki og en konflikt mellem Danmark og de tyske stater, der deltog som slesvig-holstenernes allierede. I et bredere perspektiv var den også international, idet stormagterne, især Rusland og Storbritannien, greb politisk ind. 0 2

Analysis: The model is not good enough.

Polen mistede områder til Preussen i 1700-tallet Hertil kom, at Preussen erhvervede Vestpreussen ved Polens første deling i 1772, bla. Posen og Danzig ved Polens anden deling i 1793 og yderligere store områder, bla. Warszawa, ved Polens tredje deling i 1795. 0 2

Analysis: The evidence needs interpretation to choose the right label. Given that, the model labels correctly.

Preussen blev delt op efter anden verdenskrig

Efter 1 Verdenskrig var Preussen fortsat det største tyske Land og indgik som en parlamentarisk-demokratisk fristat i Weimarrepublikken 1919-33 med en relativt stabil regering, der bortset fra korte afbrydelser blev ledet af socialdemokraten Otto Braun som preussisk ministerpræsident, indtil rigskansler Franz von Papen ved det såkaldte Preussenkup i 1932 afsatte den preussiske regering og blev indsat som rigskommissær af rigspræsident Paul von Hindenburg. Efter 2 Verdenskrig blev Preussen delt op i fire besættelseszoner, indtil den preussiske stat 2521947 blev opløst af Det Allierede Kontrolråd ud fra en opfattelse af, at den preussiske stat og militarisme var det egentlige arnested for to verdenskrige og nazismen. Fortolkningen af den preussiske historie har dels omfattet 1800-t. 0 2

Analysis: The model is not good enough.

DFDS ejer både

En anden kilde angiver, at der aldrig har været faste afgange til Island fra Islands Brygge, og at skibe til Island afgik fra kajpladser på Christianshavn. Handel med Island foregik fra Grønlandske handelsplads sammen med handelen med Færøerne, Finmarken og Grønland. Denne kilde mener at navnet på gaden bare er en del af navngivningsmønstret for de første veje på Islands Brygge som fik nordatlantiske navne.Det er sikkert, at de tidlige veje i området

blev opkaldt efter byer og sagaer fra Island samt en enkelt færøsk by Det menes, at traditionen med de nordatlantiske navne blev skabt af Axel Heide, som afløste CF. Tietgen som direktør for Privatbanken og var en af drivkræfterne bag opfyldningen af området. Axel Heide havde tætte bånd til DFDS, som havde skibe opkaldt efter islandske byer. 0 2

Analysis: You can argue that "både" is not a synonym for "skibe", and as such, the claim should be labelled with 0.

Preussen blev udvidet med områder tidligere ejet af kirken

fortsatte den brandenburg-preussiske tradition med at åbne grænserne for driftige religiøse flygtninge fra det øvrige Europa. Under Napoleonskrigene kunne Preussen i første omgang udvide sit territorium ved den franskdikterede rationalisering af det tyske territoriale kludetæppe, dvs. fra 1803 til nedlæggelsen af Det Tysk-romerske Rige i 1806, hvor flere gejstlige områder og førhen frie byer tilfaldt Preussen. 0 2

Analysis: The model is not good enough.

9.2.3 Predicted Refuted, actual NEI

Finlands mindste sø hedder Uusi

Finland er de tusinde søers land. Mere præcist har landet 187.888 søer (større end 500 m) og 98050 øer. Den største sø Saimaa er den fjerdestørste i Europa. Størstedelen af øerne finder man i skærgården i det sydvestlige af landet omkring Turku. Ålandsøerne, der hører til Finland, ligger et stykke fra skærgården med blot 38km til Sverige. 1 0

Analysis: "Uusi" is not present in the evidence, nor is "mindste", why the model should have predicted NEI.

Kongens Nytorv har eksisteret i 1300 år

Indtil for nylig var det ældste spor efter bymæssig bebyggelse i Københavnsområdet inden for voldene fra omkring år 1000, hvor der er fundet spor fra et mindre fiskerleje dér, hvor København ligger i dag. Fiskerlejet lå lige nord for Københavns Rådhus omkring Mikkel Bryggers Gade, der dengang lå ud til havet. Men i forbindelse med udgravning af Metroen har man fundet spor af bådebroer ved Gammel Strand, der daterer sig helt tilbage til omkring år 700. Ved udgravningen til metrostationen ved Kongens Nytorv har man endvidere fundet spor af en gård fra vikingetiden.Første gang forløberen til København under navnet "Havn" nævnes i kilderne, er i forbindelse med et søslag mellem Svend Estridsen og den norske konge Magnus den Gode i 1043. Derefter er der tavshed om byens skæbne i de næste 120 år. 1 0

Analysis: This claim demands that the model knows what year we have now, can find the year coupled with the named entity "Kongens Nytorv", and can subtract the two numbers. Since none of the words in "har eksisteret i 1300 år" is present in the evidence, this claim should have been categorized as NEI.

Potsdamkonferencen blev afholdt i Tyskland

Potsdamkonferencen, konference mellem USA, Storbritannien og Sovjetunionen, afholdt på slottet Cecilienhof i byen Potsdam 177. -28. 1945; i spidsen for de tre delegationer var USA's præsident Harry S Truman, den britiske premierminister Winston S Churchill, senere afløst af Clement Attlee, og den sovjetiske leder Josef Stalin. Deltagerne forhandlede om fredsordningen i Europa efter 2 Verdenskrigs afslutning, men overdrog udformningen af de konkrete fredsaftaler med de besejrede aksemagter til et udenrigsministerråd med deltagelse af USA, Storbritannien, Sovjetunionen, Frankrig og Kina. 1 0

Analysis: The text says that the conference of Potsdam was held in the city of Potsdam, but it doesn't say which country Potsdam is located in. Therefore, the prediction should be NEI. The model could have rejected the claim because it set "Potsdam" as opposite to "Tyskland".

Israel og Egypten var fjender

I 1947 foreslog FN Palæstina delt i en arabisk og en jødisk stat. I november 1977 fløj Egyptens præsident Anwar al-Sadat til Israel for at fortsætte de drøftelser, der i hemmelighed havde fundet sted vedrørende en fredsslutning, og i september 1978 underskrev Sadat og Begin Camp David-aftalen. Den blev i marts 1979 fulgt op af underskrivelsen af en egentlig israelsk-egyptisk fredsaftale. Derimod brød forhandlingerne mellem Israel og Egypten om de erobrede palæstinensiske områders fremtid sammen, ikke mindst som følge af Begins fastholdelse af det jødiske folks ret til hele det bibelske Palæstina og dermed også til de besatte områder. 1 0

Analysis: The model not good enough.

Danmark fik sine egne frimærker i 1800-tallet

Landet fik i høj grad lov til at passe sig selv, hvilket – i tråd med de nationalistiske strømninger over hele Europa – gav plads til en finsk national forståelse, især fra omkring 1860, hvor det finske sprog efterhånden blev det dominerende. Udgivelsen af det finske nationale epos Kalevala (1835) medvirkede til dette, og i løbet af århundredet blev svensk afløst af finsk også i administrationen, selv om det først var i 1892, at finsk blev ligestillet med svensk ved domstolene. Landet fik også egne frimærker og møntfod, den finske mark (fra 1860). 1 0

Analysis: The model not good enough.

Månen fjerner sig lidt fra Solen hvert år

På grund af tidevandsvirkningerne fjerner Månen sig fra Jorden med ca 38 mm om året. Over millioner af år giver denne lille virkning – og den samtidige forøgelse af Jordens døgn med omkring 23 mikrosekunder om året – anledning til en betydelig forskel. Som eksempel var året i den geologiske Devon-periode for omkring 410 millioner år siden på 400 døgn, der hver varede 21,8 timer.Månen kan på dramatisk vis have påvirket livets udvikling ved at ændre Jordens klima. Palæontologiske fund og computersimulationer viser, at Jordens aksehældning stabiliseres af tidevandsinteraktionen med Månen. 1 0

Analysis: The model is not good enough.

Ragnarock betyder dommedag

feks. kunne en hærfører symbolsk ofre fjendens mænd til Odin ved at slynge et spyd mod dem

med råbet Odin eje jer alle; det kan have været Odins egenskab af dødsgud, der lå til grund for denne praksis. Valhalla var krigerens dødsrige, hvor Odin samlede helte og konger, som gennem kamp blev viet til ham, så de kunne hjælpe ham i Ragnarok. 1 0

Analysis: The model is not good enough.

9.2.4 Predicted Refuted, actual Supported

Tyskland brugte giftgas i første verdenskrig

I 1915 anvendte tyskerne klorgas for første gang Senere i 1915 anvendte tyskerne fosgengas der var usynlig og kun kunne opdages når soldaterne faldt om I 1917 anvendtes for første gang sennepsgas der også optages gennem huden ved ætsning 1 2

Analysis: In the claim, the word "giftgas" is used as a common word for the words "klorgas", "fosgengas", and "sennepsgas". The named entity "første verdenskrig" from the claim is not mentioned in the evidence, only the years "1915" and "1917", which are years where the first world war took place. Hence, some language and world knowledge is necessary to label this claim.

Israel ville ikke give slip på de besatte palæstinensiske områder

I 1947 foreslog FN Palæstina delt i en arabisk og en jødisk stat. I november 1977 fløj Egyptens præsident Anwar al-Sadat til Israel for at fortsætte de drøftelser, der i hemmelighed havde fundet sted vedrørende en fredsslutning, og i september 1978 underskrev Sadat og Begin Camp David-aftalen. Den blev i marts 1979 fulgt op af underskrivelsen af en egentlig israelsk-egyptisk fredsaftale. Derimod brød forhandlingerne mellem Israel og Egypten om de erobrede palæstinensiske områders fremtid sammen, ikke mindst som følge af Begins fastholdelse af det jødiske folks ret til hele det bibelske Palæstina og dermed også til de besatte områder. 1 2

Analysis: The reader must apply some interpretation of the evidence to assess the claim.

Amager ligger i nærheden af hav

Der er gjort en del fund fra forhistorisk tid i Københavnsområdet. Ved bygningen af Amager Strandpark fandt man feks. levn af en kystnær boplads fra yngre stenalder. Gravhøje i forstæderne tyder på menneskelig aktivitet i forhistorisk tid, og mange af bynavnene i nærheden af København vidner derudover om grundlæggelse af byer i det storkøbenhavnske område i vikingetiden. 1 2

Analysis: The reader must apply some interpretation, e.g. that "Amager Strandpark" is in Amager, and that "kystnær" means close to the sea.

Finland kæmpede først sammen med Tyskland

Under anden verdenskrig var Finland allieret med aksemagterne, Nazityskland, Italien og andre pro nazistiske stater som bla. Ungarn og Rumænien. Finland kæmpede to gange mod Sovjetunionen: Første gang i vinterkrigen 1939–40, efter at Sovjetunionen havde angrebet landet, og igen i fortsættelseskrigen i 1941-44 i kølvandet på operation Barbarossa, i hvilken Nazityskland invaderede Sovjetunionen. I 872 dage belejrede tyskerne Leningrad, Sovjetunionens næststørste

by Belejringen af Leningrad resulterede i omkring en million døde af byens befolkning. Finske tropper kontrollerede nogle områder omkring byen, men nægtede selv at angribe eller lade tyskerne bruge disse områder som basis for deres angreb; det er fortsat et kontroversielt emne, om de finske tropper primært var behjælpelige med belejringen eller var modstræbende i dette. Efter at have kæmpet mod en stor sovjetisk offensiv i juni-juli 1944 fik finnerne våbenstilstand, men snart efter fulgte Laplandskrigen i 1944–45, hvor Finland trængte tyske styrker ud af Nordfinland. 1 2

Analysis: Difficult for the model to assess.

Hillary fik færre valgmandsstemmer end Trump

På valgdagen førte Hillary Clinton med gennemsnitligt 3% i meningsmålingerne, men det blev alligevel Trump, der vandt i, hvad der må betegnes som den største overraskelse i amerikansk politisk historie. Hillary Clinton fik ganske vist 3,1 mio. stemmer flere end Trump, men med overraskende sejre i flere stater i Midtvesten vandt han 304 valgmandsstemmer mod hendes 232 og stod således til at blive USA's 45 præsident. 1 2

Analysis: There is some calculation to be done to the numbers in the evidence.

Blekingegadebanden havde en café

Desuden er der planer om at etablere en café på pladsen ved Gemini Residence. Af restauranter ligger der Ceco i Islands Brygge nord samt Madeleines Madhus i Islands Brygge syd. Den nu genåbnede Hatoba, som er en sushirestaurant, hvor Allan Nielsen er medejer, ligger på hjørnet af gaden Islands Brygge og Vestmannagadegade.Den første café på Islands Brygge, Café Liberation, blev etableret i 1988 af en gruppe personer, der senere blev kendt som medlemmer af Blekingegadebanden. 1 2

Analysis: There is some inference to get the true label.

nazisterne brugte terror

Nazisterne havde nu fået regeringsansvaret, og i de næste fem måneder udnyttede de behændigt de muligheder, der opstod, til i en kombination af statslige indgreb og revolutionær SA-terror at sikre sig hele magten og frigøre sig fra al politisk og retlig kontrol. 1 2

Analysis: The model is not good enough.

den Hellige Alliance var imod revolution

I 1815 indgik Preussen i det løse statsforbund Det Tyske Forbund, der var præget både af stormagternes reaktionære Hellige Alliance, som var vendt mod revolutionære opstande i hele Europa, og frem til 1866 også i stigende grad af modsætningen mellem de to største tyske stater, Preussen og Østrig. Preussens ledende rolle på det økonomiske felt blev tydelig i 1834 med etableringen af det økonomisk liberalistisk inspirerede Tyske Toldforbund (se Zollverein) under preussisk ledelse og med udelukkelse af Østrig, foruden ved at centrale industrielle regioner, fx Ruhr og Øvreschlesien, i 1850'erne lå inden for det preussiske statsområde. Politisk optrådte Preussen ikke liberalt, men undertrykte i Den Hellige Alliances ånd både tysk-nationale og liberale bevægelser. 1 2

Analysis: The model is not good enough.

Trump har mange hvide mænd som vælgere

Mod de fleste politiske iagttageres forventninger syntes det imidlertid ikke at skade hans popularitet blandt en betragtelig del af de republikanske primærvalgsvælgere. Især blandt hvide mandlige vælgere uden en videregående uddannelse fik Trump stor opbakning til sine løfter om at bremse globaliseringen og bringe industrijobs tilbage til USA. Mange så Trumps afvisning af "politisk korrekthed" som et bevis på, at han netop ikke var en typisk politiker, hvilket de satte pris på. 1 2

Analysis: The claim is indirectly supported in the evidence. The model is not good enough.

9.2.5 Predicted Supported, actual Refuted

Balfourdeklarationen gav Palæstina til jøderne

Palæstina hørte fra 1517 under Osmannerriget, men erobredes i 1917 af Storbritannien. Samme år tilkendegav briterne i Balfourdeklarationen, at de ville arbejde for et jødisk hjemland i området. På fredskonferencen i San Remo i 1920 blev Palæstina britisk mandatområde under Folkenes Forbund; Transjordanien udskiltes i 1922. Ønsket om et jødisk hjemland i Palæstina var blevet aktualiseret af den antisemitisme og nationalisme, der herskede i Europa i slutningen af 1800-t. 2 0

Analysis: It is suggested in the evidence that the claim could be true, but it is not directly supported.

Seksdageskrigen fandt sted i Israel

Israel, republik i Mellemøsten, oprettet i maj 1948 på baggrund af en vedtagelse i FN Under efterfølgende krige med sine naboer har Israel erobret store landområder. Nogle er efter forhandlinger givet tilbage (Sinai), mens andre er annekteret til Israel, således Østjerusalem i 1967 og Golanhøjderne i 1981. Endvidere har Vestbredden været besat siden Seksdageskrigen i 1967. Israel er en jødisk indvandrerstat; 80 % af indbyggerne er jøder, og langt de fleste er indvandret siden staten Israels oprettelse. 2 0

Analysis: It seems the model infers the claim to be supported, but the claim is not directly supported in the evidence.

Der er moræneler under København

Geologisk set hviler København ligesom det meste af Danmark på et istidspræget grundmorænelandskab, der igen hviler på en hårdere undergrund af kalksten. Visse steder i området er der blot ti meter ned til kalklaget, der under bygningen af metroen voldte betydelige problemer. 2 0

Analysis: It seems the model infers the claim to be supported, but the claim is not directly supported in the evidence.

Vendsyssel ligger mod øst

Vendte man blikket mod øst i disse yderste, nordlige egne, så man, at landskaberne ud mod Kattegat havde deres helt eget præg. I det østlige Vendsyssel, fra Hammer til Tolne Bakker rejste sig et voldsomt formet, skovklædt højland. Og længere nede ad den jyske østkyst tårnede de ligeledes skovklædte højder fra istidens morænedannelser sig op over de dybt indskårne fjorde, bakke op og bakke ned helt ned til de milde egne i den vestlige ende af Østersøen. 2 0

Analysis: Difficult for the model to assess correctly.

der er meget turisme i Finland

I 2005 gav turismen i Finland anledning til en omsætning på 6,7 milliarder, hvilket var fem procent højere end året forinden. Denne omsætning dækker over mere end fire millioner besøgende. Turister i Finland kan først og fremmest nyde landskabet i landet med de talløse fyrre- og grantræer, det kuperede terræn og en labyrint af søer og vige. Store dele af landskabet finder man i uberørt tilstand, og der er i alt 35 nationalparker fra sydlige bredder af den Finske Bugt til Laplands høje fjelde. Der findes også bymæssige områder med et rigt udbud af kulturelle aktiviteter. Rejser med cruisefærger mellem de største havnebyer i det østlige af Østersøen, heriblandt Helsinki, Turku, Stockholm, Tallinn og Travemünde er en vigtig faktor i den finske turistbranche. 2 0

Analysis: It's difficult to assess how much is "meget". The model's assessment might be right.

Frederik den Store var inspireret af det oplyste enevælde

Indenrigspolitisk gennemførte Frederik den Store reformer i den oplyste enevældes ånd med fortsat merkantilistisk politik og landindvinding langs Oder, men også ved retslige reformer, der i slutningen af 1700-t. gjorde Preussen til en egentlig retsstat med tale- og trosfrihed, med adskillelse af justitsvæsen og forvaltning og med en ikke længere enevældig monark, men en "statens første tjener". Hertil kom, at de oplyste herskere i 1700-t. 2 0

Analysis: The evidence does hold information that supports the claim, although not explicitly. You could argue that the model was right in its assessment.

Finland underskrev Kyotoaftalen

Finland gik ind for på betingelserne i Kyoto-aftalen og inden for EU omkring udledning af drivhusgasser. Dette kan komme til at give dyrere energipriser, hvilket forstærkes af nogle ældre produktionsanlæg, der snart står til nedlæggelse. Energiproducenterne vil snart komme til at øge produktionen af atomkraftbaseret energi yderligere, idet det finske parlament i juli 2010 gav grønt lys for etableringen af yderligere to reaktorer. 2 0

Analysis: The evidence does not directly support the claim. The model is not sensitive enough, although there is a semantic error that could blur the assessment.

bronzefund er sjældne i Norge

Fra omkring 1500 fKr. kom bronzen efterhånden til landet, men brugen af sten til redskaber fortsatte; der var kun få varer i Norge, der kunne byttes for bronzen, der derfor var sjælden. De få bronzefund er ofte rigt udsmykkede våben og brocher, som kun høvdinge havde råd til.

Store dysser med grave blev anlagt nær havet så langt nordpå som ved Harstad samt inde i landet i det sydlige Norge; de er karakteristiske for perioden. 2 0

Analysis: It can be inferred from the evidence that the claim is true, so in this respect the model is right.

Wienerkongressen gav Preussen sine territorier tilbage

I næste omgang efter det totale nederlag til Napoleon ved Jena og Auerstedt i 1806 måtte Preussen ved Freden i Tilsit i 1807 dog afstå halvdelen af alle sine territorier, bla. alle områder vest for Elben. Ved Wienerkongressen i 1814-15 generhvervede Preussen sin stormagtsstatus og kunne hertil føje territoriale gevinster, der bla. 2 0

Analysis: The model is not good enough.

Nikolaj Kirke ligger i København

Byen med de skønne tårne: turistslogan skabt af brygger Carl Jacobsen i 1910. Det hentyder til de mange tårne og spir, der dengang som nu sætter sit præg på navnlig Indre By, og hvortil den generøse brygger i øvrigt også selv bidrog i form af spiret på Nikolaj Kirke. 2 0

Analysis: The claim is actually true, but the reader must have some world knowledge to know that "Indre By" is in "København".

USA har oversøiske militærbaser

USA, United States of America, forbundsstat i Nordamerika, som i kraft af sin politiske, militære og økonomiske førerstilling har været den mest magtfulde og indflydelsesrige nation efter 2 verdenskrig. Mens de nuværende grænser og oversøiske besiddelser i store træk har ligget fast fra 1900, er der via militærbaser, alliancer og samarbejdsaftaler sket en eksplosiv udvidelse af landets interessesfærer, samtidig med at "the American way of life" har bredt sig til næsten alle afkroge af verden jf amerikanisering. USA (Landefakta)Dansk navnUSALokalt navnUnited States of AmericaEngelsk navnUnited States of AmericaUafhængighed1783Areal9670000 km2Indbyggertal (2010)308700000HovedstadWashington, DC. Sprogengelsk (off. 2 0

Analysis: The model is not good enough.

9.2.6 Predicted Supported, actual Refuted

Vendsyssel ligger i Sønderjylland

Vendte man blikket mod øst i disse yderste, nordlige egne, så man, at landskaberne ud mod Kattegat havde deres helt eget præg. I det østlige Vendsyssel, fra Hammer til Tolne Bakker rejste sig et voldsomt formet, skovklædt højland. Og længere nede ad den jyske østkyst tårnede de ligeledes skovklædte højder fra istidens morænedannelser sig op over de dybt indskårne fjorde, bakke op og bakke ned helt ned til de milde egne i den vestlige ende af Østersøen. 2 1

Analysis: The label must be found via inference: First, in the claim that "Vendsyssel ligger i Sønderjylland" the model must infer that "Sønderjylland" means "sydlige Jylland" or "Sydjyl-

land". Then, the model must infer from the evidence "i disse yderste, nordlige egne" and the named entity "Vendsyssel", that "Vendsyssel" must lie to the north. Then, the model must infer that there is a contradiction between "Vendsyssel" lying to the north and in the south of Jylland. Hence, the claim must be labelled REFUTED.

USA tilsluttede sig Versaillestraktaten

Regeringsførelsen på føderalt plan var i det væsentlige reduceret til at bistå de store virksomheder, korporationer, der opbyggede netværk, som i omfang og effektivitet kom til at overskygge postvæsenet, den eneste landsdækkende offentlige organisation. Deltagelsen i 1 Verdenskrig gav en forsmag på USA's internationale gennemslagskraft. Senatet forkastede imidlertid Versaillestraktaten, i særdeleshed præsident Woodrow Wilsons forslag om et Folkenes Forbund, som ville forpligte USA til at garantere fredsslutningens grænser. 2 1

Analysis: This claim demands that the word "senatet" be interpreted as "USA".

den norske olieproduktion giver underskud

I 1969 fandt Phillips Petroleum Company omfattende mængder af olie i Ekofisk-feltet i havbunden vest for Norge. I 1973 oprettede den norske regering det statslige olieselskab Statoil, der gik ind i olieudvindingen. På grund af de store investeringer kom olieproduktionen først til at give overskud i begyndelsen af 1980'erne, men har siden udviklet sig til at være en af de største indkomstområder for landet. 2 1

Analysis: The claim is supported in the evidence but not word-to-word.

der findes 32 regioner i Chile

Chile er opdelt i 15 regioner (spansk región), der hver ledes af en intendant. Regionerne er opdelt i provinser, hvor en guvernør står i spidsen. Intendanter og guvernører udpeges af landets præsident. Det nederste administrative niveau er kommuner, som administreres af kommunalbestyrelser med borgmestre som øverste ledere. Kommunalbestyrelserne vælges for fire år ad gangen. 2 1

Analysis: This is simply the model not being good enough.

Blood work er en roman skrevet af Eastwood

I 2002 spillede Eastwood en tidligere FBI-agent, der jager en sadistisk morder, spillet af Jeff Daniels i thrilleren Blood Work, som er løst baseret på en roman af samme navn fra 1998 af Michael Connelly. Filmen var en kommerciel fiasko, og den indtjente blot $26,2 millioner, ca halvdelen af dens budget. Eastwood vandt dog Future Film Festival Digital Award ved Venedig Film Festival for denne film. 2 1

Analysis: The model not good enough.

Churchill indgik en våbenhvile med Hitler

Churchill havde været blandt de første som erkendte den voksende trussel fra Hitler længe inden 2 verdenskrig brød ud, men hans advarsler blev stort set ikke fulgt. Selv om der var grupper i Storbritannien som ønskede en forhandlingsfred med det Tyskland, som tydeligvis var ved

at blive stærkere, heriblandt udenrigsministeren Lord Halifax, afviste Churchill at overveje en våbenhvile med Hitlers Tyskland. Hans retorik styrkede folkeopinionen mod en fredsløsning og forberedte briterne på en lang krig. I sin tale til Underhuset den 18 juni 1940 udtrykte han det således: "Jeg forventer at slaget om England snart begynder." Ved at afvise en våbenstilstand med Tyskland hold Churchill modstanden i live i Det britiske imperium og skabte dermed grundlaget for De Allieredes modoffensiver i 1942–45, hvor Storbritannien var base for forsyning af Sovjetunionen og befrielsen af Vesteuropa. 2 1

Analysis: The model not good enough.

finske marker krævede ikke kultivering

De fleste marker har tidligere været enten skov eller sump, og jorden har som regel krævet meget kalkning og års kultivering for at neutralisere rester af surbund og udvikle frugtbarheden. Vanding har normalt ikke været nødvendig, men til gengæld har der flere steder været brug for dræning for at lede vand væk fra markerne. Det finske landbrug er normalt regnet for at være ret effektivt sammenlignet med de fleste andre europæiske lande. 2 1

Analysis: The model not good enough.

Nemunas ligger i Polen

Navnet Preussen stammer fra den baltiske folkestamme preusserne, der i middelalderen beboede området mellem floden Wisła (tysk Weichsel) i Polen og floden Nemunas (tysk Memel) i Litauen og Hviderusland. Preussen opstod som hertugdømme i 1525 ved en overenskomst mellem den polske kong Sigismund 1 Stary og Den Tyske Ordenshøjmester, der derved blev hertug i en protestantisk stat under polsk lenshøjhed. På grundlag af den enevældige stat, som Frederik Vilhelm den Store af Brandenburg havde opbygget efter Trediveårskrigen, blev kurfyrstendømmet Brandenburg i 1701 af den tysk-romerske kejser ophøjet til kongedømme og tog navn efter hertugdømmet Preussen. Under hans søn Frederik Vilhelm 1, der regerede 1713-40, skabtes grundlaget for Preussens senere position som europæisk stormagt. 2 1

Analysis: The model not good enough.

krigen mellem Iran og Irak sluttede i 1990

Golfkrigen, (efter Den Persiske Golf), krig i januar-februar 1991 mellem Irak og en international allieret styrke. Årsagen var Iraks besættelse af Kuwait den 281990. Irak begrundede besættelsen med, at Kuwait under krigen mellem Irak og Iran 1980-88 ulovligt havde pumpet olie op fra et irakisk-kuwaitisk oliefelt og solgt olien med stor fortjeneste. Desuden havde den kuwaitiske regering vægret sig ved at efterkomme irakiske ønsker om en begrænsning af den samlede arabiske olieproduktion. 2 1

Analysis: The model not good enough.

Det Tyske kejserrige og Østrig-Ungarn var på samme side i første verdenskrig

Første Verdenskrig blev i begyndelsen hovedsageligt ført mellem Centralmagterne det Tyske kejserrige og Østrig-Ungarn på den ene side og Ententemagterne Frankrig Rusland Serbien og Storbritannien på den anden side Også styrker fra det britiske imperium bla Canada deltog 2 1

Analysis: This is an error. The model has predicted correctly.

Første verdenskrig blev ikke udløst af mordet på Franz Ferdinand

Første verdenskrig blev udløst af mordet i Sarajevo den 28 juni 1914 på den østrig-ungarske tronfølger ærkehertug Franz Ferdinand 2 1

Analysis: The model is not good enough. Clear error.

Elpriserne er højest i Finland

Enhver kan gå ind på det frie og overvejende privatfinansierede og fysiske nordiske energimarked, som handles på Nord Pool, siden sommeren 2007 en del af NASDAQ OMX Group. Dette har givet mere konkurrencedygtige priser på energi sammenlignet med det øvrige EU I 2011 havde Finland de laveste el-priser i EU for firmaer (sammen med Estland og Bulgarien).Det samlede energiforbrug i Finland i 2010 var på 1,46 millioner TJ, hvilket var en stigning på 10 % i forhold til året før. El-forbruget var på 87,7 TWh, hvilket var en stigning på 8 % fra året forinden. Finland har ud over tørv og træ ikke naturlige kulbrintebaserede råstoffer. Omkring 10-15 % af elektriciteten i landet produceres som vandkraft, en noget mindre andel end de mere bjergrige nabolande Sverige og Norge. I 2008 udgjorde vedvarende energi (heraf primært vandkraft og forskellige træbaserede energiformer) 30,5 % af det samlede forbrug, hvilket er højt sammenlignet med gennemsnittet i EU på 10,3 %.. 2 1

Analysis: The model is not good enough.

Trump fik flere stemmer end Hillary Clinton

På valgdagen førte Hillary Clinton med gennemsnitligt 3% i meningsmålingerne, men det blev alligevel Trump, der vandt i, hvad der må betegnes som den største overraskelse i amerikansk politisk historie. Hillary Clinton fik ganske vist 3,1 mio. stemmer flere end Trump, men med overraskende sejre i flere stater i Midtvesten vandt han 304 valgmandsstemmer mod hendes 232 og stod således til at blive USA's 45 præsident. 2 1

Analysis: The model is not good enough.

Hillary Clinton vandt i virkeligheden over Obama

I 2000 vandt Hillary Clinton som kandidat for Det Demokratiske Parti en plads i Senatet for staten New York, og i januar 2007 meddelte hun, at hun opstillede til det amerikanske præsidentvalg i 2008. På trods af høje stemmetal lykkedes det dog ikke for Hillary Clinton at blive Det Demokratiske Partis præsidentkandidat, og hun måtte se sig slået af Barack Obama efter en langvarig valgkamp. På trods af den hårde kampagne mod Obama udnævntes Hillary Clinton ikke desto mindre til udenrigsminister efter Obamas indsættelse i præsidentembedet i januar 2009. 2 1

Analysis: Difficult example where the claim uses different words than the evidence.

Frankrig og Østrig indgik en alliance i 1755

Premierministeren, William Pitt dæ. , havde hele tiden ønsket franskmændene engageret på kontinentet for at skaffe ro til at besejre dem i Amerika. Dette forehavende lykkedes med den

fransk-østrigske alliance 1756. 2 1

Analysis: The year is wrong; the model should have spotted that.

Franz Ferdinand blev myrdet af Gavrilo Princip

1 verdenskrig blev udløst af mordet i Sarajevo den 28 juni 1914 på den østrig-ungarske tronfølger ærkehertug Franz Ferdinand som blev myrdet af den serbiske nationalist Gavrilo Princip 2 1

Analysis: Error. The model predicts correctly.

vætter var mennesker

De ritualer, der nævnes i lovteksterne, var dem, som hovedsageligt var relateret til kollektivguder som diser, alfer og vætter. Af lovene fremgår det, at det primært var kvinder, som blev mistænkt for at udføre dem. På den baggrund har den norske religionsforsker Else Mundal foreslået, at de kvindelige hedenske kultledere kan have videreført disse kultformer, selv om de var blevet kriminaliseret, fordi deres funktion ikke kunne kanaliseres ind i de kristne ritualer. I det førkristne samfund havde kvindelige kultledere haft stor betydning, og denne status mistede de i det kristne samfund. Samtidig var der ikke nogen åbenlys erstatning for de funktioner, de varetog: Morten Warmind har peget på, at det kan være forklaringen på, at det hovedsageligt var kvinder, der videreførte de gamle ritualer. 2 1

Analysis: The model is not good enough.

John Locke var børnehaveklasseleder

De fleste af Wollstonecrafts tidligste skrifter drejede sig om pædagogik. Hun samlede en antologi af litterære uddrag "til forbedringen af unge kvinder" kaldet The Female Reader (Den kvindelige læser), og hun oversatte to børnebøger: Maria Geertruida van de Werken de Cambons Young Grandison (Det Unge barnebarn) og Christian Gotthilf Salzmanns Elements of Morality (Moralens elementer). Hendes egne skrifter handlede også om det. I begge hendes bøger om opførsel, Thoughts on the Education of Daughters og børnebogen Original Stories from Real Life (Originale fortællinger fra virkeligheden) er hun fortaler for, at alle børn bør uddannes i den fremvoksende middelklassens etik: selvdisciplin, ærlighed, nøjsomhed og social tilfredshed.Bøgerne fremhævede også vigtigheden af at lære barn fornuft, at ræsonnere og drage slutninger, noget som viser hendes intellektuelle gæld til den betydningsfulde pædagog og filosof John Locke. 2 1

Analysis: The model is not good enough.

der har været stor indvandring til Chile

Konventionen har siden været anvendt i forskellige tilfælde til fordel for oprindelige folk i landet, fx retten til vand for aymara-folk. Immigration til Chile har ikke været så omfattende som til andre lande, fx Argentina, Canada og USA. Det skyldes blandt andet landets afsides beliggenhed bag Andesbjergene samt afstanden til Europa. 2 1

Analysis: The model is not good enough.

Sovjetunionen er en vigtig handelspartner for Finland

I 1981 måtte Kekkonen trække sig tilbage af helbredsgrunde efter 25 år på præsidentposten. I de følgende år oplevede Finland en dyb recession som følge af blandt andet fejlberegnede makroøkonomiske beslutninger, en bankkrise, kollapset af Sovjetunionen, der var den hidtidigt vigtigste handelspartner, samt en global økonomisk nedgang. Nedgangen nåede bunden i 1993, hvorefter Finland i de følgende år oplevede en solid økonomiske vækst. 2 1

Analysis: Error. Should either have had the label 0 or 2.

Trump var kandidat for Reformpartiet

Her blev han gennem 13 sæsoner for alvor landskendt som chefen, der sendte deltagere ud af konkurrencen med ordene "You're fired!" ('du er fyret!'). Gennem flere årtier luftede Trump med mellemrum idéen om at indlede en politisk karriere. I 1999 undersøgte han således mulighederne for at blive Reformpartiets præsidentkandidat, men besluttede sig alligevel for ikke at stille op Senere sluttede han sig til Det Republikanske Parti, og i 2011 annoncerede han, at han overvejede at forsøge at blive partiets kandidat ved præsidentvalget det følgende år. 2 1

Analysis: The model is not good enough.

Storbritannien og Frankrig var allierede i Syvårskrigen

Syvårskrigen, Den Preussiske Syvårskrig, traditionel betegnelse for krigen i Centraleuropa 1756-63 mellem Østrig, Frankrig, Rusland og Sverige på den ene side og Preussen støttet af Storbritannien på den anden. Sideløbende dermed udkæmpedes imidlertid kolonikrigen mellem Storbritannien og Frankrig. Denne krig udspillede sig på fire kontinenter, foruden Europa i Amerika, Afrika og Asien. Betegnelsen Syvårskrigen benyttes her om dette samlede verdensomspændende opgør, der i virkeligheden kan betegnes som en verdenskrig. 2 1

Analysis: The model is not good enough.

Obama reddede over to millioner amerikanske arbejdspladser

Efter sin tiltrædelse fik præsident Obama vedtaget lovgivning, der yderligere skulle stimulere den betrængte amerikanske økonomi og redde arbejdspladser, bla. i den hårdt trængte bilindustri. Alene her lykkedes det med offentlige lån (som senere alle blev betalt tilbage) at redde omkring 1,5 millioner arbejdspladser. 2 1

Analysis: The model is not good enough.

Camp David-aftalen blev aldrig underskrevet

I 1947 foreslog FN Palæstina delt i en arabisk og en jødisk stat. I november 1977 fløj Egyptens præsident Anwar al-Sadat til Israel for at fortsætte de drøftelser, der i hemmelighed havde fundet sted vedrørende en fredsslutning, og i september 1978 underskrev Sadat og Begin Camp David-aftalen. Den blev i marts 1979 fulgt op af underskrivelsen af en egentlig israelsk-egyptisk fredsaftale. Derimod brød forhandlingerne mellem Israel og Egypten om de erobrede palæstinensiske områders fremtid sammen, ikke mindst som følge af Begins fastholdelse af det jødiske folks ret til hele det bibelske Palæstina og dermed også til de besatte områder. 2 1

Analysis: The model is not good enough.

Trump har før stillet op som præsidentkandidat

Måden, han gjorde det på, var bemærkelsesværdig: Han stillede sig i spidsen for den såkaldte "birther"-bevægelse, der hævdede, at præsident Obama ikke var en retmæssig præsident, fordi han i virkeligheden var født i Kenya, og hans amerikanske fødselsattest var falsk eller ikkeeksisterende. Kort efter, at præsident Obama offentliggjorde sin fødselsattest for at standse konspirationsteorierne, meddelte Trump, at han alligevel ikke ønskede at stille op som kandidat. Fire år senere gjorde Trump imidlertid alvor af at stille op I juni 2015 skød han sin kampagne i gang med et frontalt angreb på illegale immigranter fra Mexico, hvoraf han beskrev hovedparten som voldtægtsforbrydere, narkosmuglere og "folk med masser af problemer". 2 1

Analysis: The model is not good enough.

Blood Work bogen udkom i 2002

I 2002 spillede Eastwood en tidligere FBI-agent, der jager en sadistisk morder, spillet af Jeff Daniels i thrilleren Blood Work, som er løst baseret på en roman af samme navn fra 1998 af Michael Connelly. Filmen var en kommerciel fiasko, og den indtjente blot $26,2 millioner, ca halvdelen af dens budget. Eastwood vandt dog Future Film Festival Digital Award ved Venedig Film Festival for denne film. 2 1

Analysis: The model is not good enough.

Roskilde hører til Region Hovedstaden

Der tales endvidere om Region Hovedstaden, der indbefatter dele af Nordøstsjælland og Bornholm, men ikke Roskilde- og Køge-området. Regionen havde 1,65 millioner indbyggere, og må som begreb ikke forveksles med Hovedstadsregionen.. 2 1

Analysis: The model is not good enough.

Clint Eastwood har været præsident

Eastwood havde stor succes, første gang han aktivt gik ind i politik. Han blev valgt som borgmester i april 1986 i Carmel-by-the-Sea, Californien (befolkningstal ca 4000), en lille, men velhavende by med fortrinsvis berømtheder som indbyggere. Byen ligger ved Montereyhalvøen. Den 30 januar 1986 havde Eastwood afleveret 30 underskrifter (10 mere end minimumskravet for opstilling) med støtte til hans kandidatur. 2 1

Analysis: The model is not good enough.

Christian II stod for byggeriet af Rundetårn

Kongens København: gennem århundrederne har skiftende konger sat deres præg på hovedstaden. Det gælder navnlig Christian IV, der foruden at udvide området inden for byvoldene til det tredobbelte bidrog med bygninger som Rosenborg, Rundetårn og Børsen. 2 1

Analysis: The model is not good enough.