DANFEVER: claim verification dataset for Danish

Jeppe Nørregaard
IT University of Copenhagen
jeno@itu.dk

Leon Derczynski
IT University of Copenhagen
leod@itu.dk

Abstract

Automatic detection of false claims is a difficult task. Existing data to support this task has largely been limited to English. We present a dataset, DANFEVER, intended for claim verification in Danish. The dataset builds upon the task framing of the FEVER fact extraction and verification challenge. DANFEVER can be used for creating models for detecting mis- & disinformation in Danish as well as for verification in multilingual settings.

1 Introduction

The internet is rife with false and misleading information. Detection of misinformation and fact checking therefore presents a considerable task, spread over many languages (Derczynski et al., 2015; Wardle and Derakhshan, 2017; Zubiaga et al., 2018). One approach to this task is to break down information content into verifiable *claims*, which can subsequently be fact-checked by automated systems.

Automated fact checking can be framed as a machine learning task, where a model is trained to verify a claim. Applying machine learning requires training and validation data that is representative of the task and is annotated for the desired behaviour. A model should then attempt to generalise over the labeled data.

One dataset supporting automatic verification is the Fact Extraction and VERification dataset (FEVER) in English (Thorne et al., 2018a), which supports the FEVER task (Thorne et al., 2018b; Thorne and Vlachos, 2019). The dataset is aimed both at claim detection and verification.

While the misinformation problem spans both geography and language, much work in the field has focused on English. There have been suggestions on strategies for alleviating the misinformation problem (Hellman and Wagnsson, 2017). It is however evident that multilingual models are essential if automation is to assist in multilingual regions like Europe. A possible approach for multilingual verification is to use translation systems for existing methods (Dementieva and Panchenko, 2020), but relevant datasets in more languages are necessary for testing multilingual models' performance within each language, and ideally also for training.

This paper presents DANFEVER, a dataset and baseline for the FEVER task in Danish, a language with shortage of resources (Kirkedal et al., 2019). While DANFEVER enables improved automatic verification for Danish, an important task (Derczynski et al., 2019), it is also, to our knowledge, the first non-English dataset on the FEVER task, and so paves the way for multilingual fact verification systems. DANFEVER is openly available at `https://figshare.com/articles/dataset/DanFEVER_claim_verification_dataset_for_Danish/14380970`

2 English FEVER

The Fact Extraction and VERification dataset and task (FEVER) is aimed at automatic claim verification in English (Thorne et al., 2018a). When comparing we will stylize the original FEVER dataset ENFEVER to avoid confusion. The dataset was created by first sampling sentences from approximately 50,000 popular English Wikipedia pages. Human annotators were asked to generate sets of claims based on these sentences. Claims focus on the same entity as the sentence, but may not be contradictory to or not verifiable by the sentence. A second round of annotators labelled these claims, producing the labels seen in Table 1, using the following guidelines:

"If I was given only the selected sentences, do

I have strong reason to believe the claim is true (`Supported`) or stronger reason to believe the claim is false (`Refuted`)."
"The label `NotEnoughInfo` label was used if the claim could not be supported or refuted by any amount of information in Wikipedia."

The ENFEVER guidelines state that claims labelled `NotEnoughInfo` could possibly be verified using other publicly available information, which was not considered in the annotation.

Label	Verifiability	#	%
Supported	Verifiable	93,367	50.3
Refuted	Verifiable	43,107	23,2
NotEnoughInfo	NotVerifiable	48,971	26,4
Total		**185,445**	-

Table 1: Annotated classes in ENFEVER.

In the FEVER task (Thorne et al., 2018b), automatic verification is commonly framed as a two-step process: given a claim, relevant evidence must first be collected, and secondly be assessed as supporting or refuting the claim, or not providing enough information. ENFEVER contains data for training models for both steps.

We tasked annotators to create claims for DANFEVER based on the same guidelines and without regulation of class-distribution. The class-distribution of DANFEVER is therefore a bit different that that of ENFEVER; there is about the same ratio of `Supported` claims, but more `Refuted` and less `NotEnoughInfo` claims in DANFEVER that in ENFEVER.

3 Method

A FEVER task instance consists of a claim, zero or more pieces of evidence, and a label. The labels take one of the following values:

`Supported` Claims that can be supported by evidence from the textual data

`Refuted` Claims that can be refuted by evidence from the textual data

`NotEnoughInfo` Claims that can neither be supported or refuted based on the textual data

The claims were created based on data from Danish Wikipedia and Den Store Danske (a privately-developed, non-profit, online encyclopedia based in Denmark and financed through foundations and universities). Both sites are generally considered high quality and trustworthy. Along with the claims, DANFEVER supplies the Wikipedia dump used for creating the claims as well as the content of the articles used from Den Store Danske. The remaining articles from Den Store Danske are not included (due to rights), and all articles should be considered to be iid.for modelling.

The format of the dataset can be found in Appendix A.1.

3.1 Dataset Goal

DANFEVER can be used for research and implementation of multi-lingual claim-detection. The dataset can be used for bench-marking models on a small language, as well as for fine-tuning when applying such models on Danish data.

3.2 Data Statement

The following is a data-statement as defined by Bender and Friedman (2018). The dataset consists of a text corpus and a set of annotated claims. The annotated part contains 6407 claims, with labels and information about what articles can be used to verify them.

Curation Rationale A dump of the Danish Wikipedia of 13 February 2020 was stored as well as the relevant articles from Den Store Danske (subset of site to adhere to rights). Two teams of two people independently sampled evidence, and created and annotated claims from these two sites (more detail in section 3.3).

Speaker Demographic Den Store Danske is written by professionals and is funded by various foundations for creating free information for the Danish public. Wikipedia is crowd-sourced and its writers are therefore difficult to specify, although the content is generally considered to be of high quality.

Annotator Demographic The annotators are native Danish speakers and masters students of IT.

Speech Situation The data is formal, written texts created with the purpose of informing a broad crowd of Danish speakers.

Language Variety and Text Characteristics The language of the texts is fairly formal Danish

Claim 3152: "Udenrigsministeriet har eksisteret siden 1848."
The Ministry of Foreign Affairs has existed since 1848.

Evidence Extract: "Dette er en liste over ministre for Udenrigsministeriet siden oprettelsen af ministeriet i 1848."
This is a list of ministers of the Ministry of Foreign Affairs since it was founded in 1848.

Evidence Entities: `wiki_93781`

Verifiable: `Verifiable`

Label: `Supported`

(a) A `Supported` claim.

Claim 1306: "Hugh Hudson er født i England i 1935."
Hugh Hudson was born in England in 1935.

Evidence Extract: "Hugh Hudson (født 25. august 1936 i London, England) er en britisk filminstruktør."
Hugh Hudson (born 25th of August 1936 in London, England) is a British film director.

Evidence Entities: `wiki_397805`

Verifiable: `Verifiable`

Label: `Refuted`

(b) A `Refuted` claim.

Claim 2767: "Lau Lauritzen har instrueret både stumfilmen Skruebrækkeren og vikingefilmen Når ræven flyver."
Lau Lauritzen directed the silent film Skruebrækkeren and the viking film Når Ræven Flyver.

Evidence Extract: ""

Evidence Entities: `wiki_833896`

Verifiable: `NotVerifiable`

Label: `NotEnoughInfo`

(c) A `NotEnoughInfo` claim.

Table 2: Examples of claims. English translations are in *italic*.

from encyclopedias. It is considered to be consistent. Any deviation from Danish language is largely due to topics on history from non-Danish regions.

3.3 Sampling and Annotation

The main text corpus was created by storing the Danish Wikipedia dump of the time as well as a subset of pages from Den Store Danske, selected from the annotation process. Two strategies were employed for gathering specific texts for claims. A selection of pages with well-known topics were selected from Wikipedia's *starred* articles and Den Store Danske (similar to the "popular articles" selection in ENFEVER). Furthermore a random selection of Wikipedia entities with abstracts were selected to ensure broad spectrum of topics. Random substrings were selected and passed to annotators, who created claims based on each substring, as in ENFEVER. The claims focus on the same entity as the substring's source document and may be supported by the text in the substring, but may also be refuted or unverifiable by the substring. It is up to the annotator to decide on what type of claim to aim for (although the final label of each claim is provided by the next annotator).

The set of claims were subsequently revisited by another annotator, who labelled the claim as `Supported`, `Refuted` or `NotEnoughInfo`, based on the original substring used to generate the claim. The majority of the claims (80%) are generated based on Wikipedia pages, while 20% were based on articles from Den Store Danske. Note that claims are independent of the source and could be verified using any text; while the FEVER format presents a list of articles where evidence is present, this list is not exhaustive, just as in the TREC and TAC challenges. The two annotating teams reported Fleiss κ-scores of 0.75 and 0.82 measured on a reduced subset. The remaining data was annotated by a single annotator.

Label	Verifiability	#	%
Supported	Verifiable	3,124	48.8
Refuted	Verifiable	2,156	33.6
NotEnoughInfo	NotVerifiable	1,127	17.6
Total		**6,407**	-

Table 3: Annotated classes in DANFEVER.

	Median	Mean	SD
Claims			
# Characters	45	50.18	22.02
# Tokens	7	8.46	3.86
# Evidence Entities	1	1.10	0.34
Evidence Extracts			
# Characters	260	305.56	257.20
# Tokens	47	53.75	44.64

Table 4: Claims and evidence extracts in dataset.

4 Dataset Details & Analysis

DANFEVER consists of 6407 claims. We have included one example from each class in Tables 2a, 2b and 2c, and shown the label distribution in Table 3.

Table 4 summarizes the lengths of claims and evidence extracts, as well as the number of entities linked to the claims.

Location	#	Person	#	Organization	#
Finland	184	Donald Trump	110	Aalborg Universitet	11
Danmark	109	Winston Churchill	73	FN	11
Preussen	89	Hillary Clinton	44	DR	10
USA	80	Mary Wollstonecraft	36	Københavns Universitet	9
Chile	79	George W. Bush	24	Electronics Art	9
København	71	Frederik Den Store	16	FC Barcelona	9
Tyskland	64	Obama	15	Apollo Rejser	8
Israel	57	Eastwood	13	Bananarama	8
Norge	54	Jens	9	EU	8
Storbritannien	49	Grant Rodiek	8	MTV	7

Table 5: Most frequent entities and number of occurrences.

4.1 Named Entities in Claims

The entities mentioned frequently in a corpus can give insight into popular themes in the data. In this case, the topic of the claims is particularly relevant. We present an automatic survey of DANFEVER's entities. Entities in claims were identified using the DaNLP NER tool (Hvingelby et al., 2020), which identifies location (LOC), person (PER), and organization (ORG) entities. Those most frequently named are shown in Table 5.[1].

5 Baseline: Recognizing Textual Entailment

The FEVER task consists of verifying claims based on a text corpus. One common strategy is to split the task into three components (as in the original work (Thorne et al., 2018a))

1. Document Retrieval: Retrieve a useful subset of documents from the corpora, based on the claim.

2. Sentence Retrieval: Retrieve a useful subset of sentences from those documents, based on the claim.

3. Recognize Textual Entailment: Classify the claims as `Supported`, `Refuted` or `NotEnoughInfo`, based on the claim and the subset of sentences.

To provide baseline performance for future research to benchmark against, we trained a baseline model on the final task; recognizing textual entailment. Since there are no evidence extracts for the `NotVerifiable` samples, we apply the random-sampling method from the original EN-FEVER paper, where evidence is randomly assigned from the data to each of these samples. We trained classifiers on the resulting 3-class problem.

[1]Interestingly the most mentioned location is Finland

The transformer based model BERT (Devlin et al., 2019) has shown promising performance for claim verification (Soleimani et al., 2020), and the team (DOMLIN) with highest FEVER-score in the FEVER2.0 competition used a BERT-based system (Thorne et al., 2019). Using the transformers repository from HuggingFace (Wolf et al., 2020) we test; `mBERT` (Feng et al., 2020) (tag: `bert-base-multilingual-cased`), `XLM-RoBERTa Small` and `XLM-RoBERTa Large` (Conneau et al., 2020; Liu et al., 2019) (tags: `xlm-roberta-base` and `xlm-roberta-large`), and the Danish `NordicBERT` (BotXO, 2019). We use BERT's sentence-pair representation for claims and evidence extracts. The classification embedding is then passed to a single-hidden-layer, fully-connected neural network for prediction. We first train the prediction layer, while freezing the weights of the language model, and consecutively fine-tune them both. We do this in a 10-fold cross-validation scheme for the 4 models.

Table 6 shows weighted-mean F1-scores, training parameters and info about the models. `XLM-RoBERTa Large` performed best, followed by `mBERT` and then `XLM-RoBERTa Small`. `NordicBERT` performed surprisingly poor. The learning curve of `NordicBERT` flattened out quickly and nothing further was learned despite the high learning rate used. `NordicBERT` was trained for Masked-Language-Modelling, but we are unsure whether it was also trained for Next-Sentence-Prediction like BERT (or even Causal-Language-Modelling like RoBERTa). If not, this may explain the poor performance on this task, even when `NordicBERT` has shown promising results for other tasks.

For comparison the multi-layer perceptron and decomposable attention models from the EN-FEVER paper (Thorne et al., 2018a) maintained

Model	F1 Train	F1 Test	Params	Time	BS	Epochs	LR	WD	DR
mBERT	94.5%	85.0%	110M	14h, 10m	32	40	10^{-5}	10^{-6}	0.3
XLM-RoBERTa Small	78.8%	78.5%	270M	11h, 40m	32	40	10^{-5}	0	0
XLM-RoBERTa Large	**98.5%**	**90.2%**	550M	18h, 20m	8	20	$5 \cdot 10^{-6}$	0	0
NordicBERT	65.5%	65.5%	110M	6h, 40m	32	20	0.001	0.0	0.1

Table 6: Model Evaluations. F1 score is weighted-mean. Params: number of parameters in model. Time: total training & evaluation time using 1 NVIDIA Tesla V100 PCIe 32 GB card; RMSProp optimizer. BS: batch size. LR: maximum learning rate in single-round, cosine schedule w/ 10% warm-up.[2]WD: weight decay. DR: dropout rate.

		Predicted		
		NEI	R	S
True Class	NEI	1118	7	2
	R	6	1643	507
	S	4	441	2679

Table 7: Test-set confusion matrix of `xlm-roberta-large` classifier.

an F1 score of respectively 73% and 88% on the verification subtask. The comparable performance indicates that pretrained, multilingual, language models are useful for the task, especially considering that DANFEVER is small relative to EN-FEVER. We show the collective test-set confusion matrix of `xlm-roberta-large` in table 7 and note that it is much easier to disregard the randomized evidence (classify `NotEnoughInfo` (`NEI`)), than it is to refute or support claims, which is to be expected.

6 Conclusion

We have presented a human-annotated dataset, DANFEVER, for claim verification in a new language; Danish. DANFEVER can be used for building Danish claim verification systems and for researching & building multilingual claim verification systems. To our knowledge DANFEVER is the first non-English FEVER dataset, and it is openly accessible[3]. Baseline results are presented over four models for the textual-entailment part of the FEVER-task.

[2]Available in Huggingface's library: `https://huggingface.co/transformers/main_classes/optimizer_schedules.html#transformers.get_cosine_schedule_with_warmup`

[3]`https://figshare.com/articles/dataset/DanFEVER_claim_verification_dataset_for_Danish/14380970`

7 Acknowledgments

This research was supported by the Independent Danish Research Fund through the Verif-AI project grant. We are grateful to our annotators (Jespersen and Thygesen, 2020; Schulte and Binau, 2020).

References

Emily M. Bender and Batya Friedman. 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.

BotXO. 2019. NordicBERT. `https://github.com/botxo/nordic_bert`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. XLM-R.

D. Dementieva and A. Panchenko. 2020. Fake News Detection using Multilingual Evidence. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 775–776.

Leon Derczynski, Torben Oskar Albert-Lindqvist, Marius Venø Bendsen, Nanna Inie, Jens Egholm Pedersen, and Viktor Due Pedersen. 2019. Misinformation on Twitter during the Danish national election: A case study. In *Proceedings of the conference for Truth and Trust Online*.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Naacl Hlt 2019 - 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:4171–4186. ISBN: 9781950737130 Publisher: Association for Computational Linguistics (ACL).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *arXiv:2007.01852 [cs]*. ArXiv: 2007.01852.

Maria Hellman and Charlotte Wagnsson. 2017. How can European states respond to Russian information warfare? An analytical framework. *European Security*, 26(2):153–170.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A Named Entity Resource for Danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4597–4604.

Sidsel Latsch Jespersen and Mikkel Ekenberg Thygesen. 2020. Fact Extraction and Verification in Danish. Master's thesis, IT University of Copenhagen.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The Lacunae of Danish Natural Language Processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. RoBERTa.

Henri Schulte and Julie Christine Binau. 2020. Danish Fact Verification: An End-to-End Machine Learning System for Automatic Fact-Checking of Danish Textual Claims. Master's thesis, IT University of Copenhagen.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

The SQLite Consortium. 2000. SQLite. `www.sqlite.org`.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against Fact Extraction and VERification. *arXiv:1903.05543 [cs]*. ArXiv: 1903.05543.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 809–819.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The Second Fact Extraction and VERification (FEVER2.0) Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

A Appendices

A.1 Format

DANFEVER contains three `sqlite` databases (SQLite Consortium, 2000); `da_fever.db`, `da_wikipedia.db` and `den_store_danske.db`.

The databases `da_wikipedia.db` and `den_store_danske.db` contain article data from Danish Wikipedia and Den Store Danske respectively. They contain an `id`-field, which is a numerical ID of the article (the `curid` for Wikipedia and a simple enumeration for Den Store Danske). They also contain the `text` and `title` of each article, as well as the `url` to that article.

The `da_fever.db` database contain the annotated claims. Each row in the database contain a `claim` and a unique `id`. With each claims comes the labels `verifiable` (`Verifiable` and `NotVerifiable`) and `label` (`Supported`, `Refuted` and `NotEnoughInfo`). The `evidence` column contain information about what articles were used to create and annotate the claim, and is composed by a comma-separated string, with IDs referring to the articles. The ID-format is `Y_X` where `Y` is either `wiki` or `dsd` to indicate whether the article comes from Danish Wikipedia or Den Store Danske, and `X` is the numerical id from that data-source. Finally the claims that were `Verifiable` contains an `evidence_extract` which is the text-snippet used to create and annotate the claim. Note that there may be some character-level incongruence between the original articles and the `evidence_extract`, due to formatting and scraping.

All three databases are also provided in TSV-format.

The data is publicly available at `https://figshare.com/articles/dataset/DanFEVER_claim_verification_dataset_for_Danish/14380970`