Negation Disrupts Compositionality in Language Models
The Czech Usecase

Tereza Vrabcová and Petr Sojka

Faculty of Informatics, Masaryk University, Brno, Czech Republic
xvrabcov@fi.muni.cz, sojka@fi.muni.cz

Abstract. In most Slavic languages, the negation is expressed by short "ne" tokens that do not affect discrete change in the meaning learned distributionally by language models. It manifests in many problems, such as Natural Language Inference (NLI).
We have created a new dataset from CsFEVER, the Czech factuality dataset, by extending it with negated versions of hypotheses present in the dataset. We used this new dataset to evaluate publicly available language models and study the impact of negation on the NLI problems.

We have confirmed that compositionally computed representation of negation in transformers causes misunderstanding problems in Slavic languages such as Czech: The reasoning is flawed more often when the information is expressed using negation than when it is expressed positively without it.
Our findings highlight the limitations of current transformer models in handling negation cues in Czech, emphasizing the need for further improvements to enhance language models' understanding of Slavic languages.

Keywords: negation, language models, machine learning.

"Negation is the mind's first freedom, yet a negative habit is fruitful only so long as we exert ourselves to overcome it, adapt it to our needs; once acquired it can imprison us."
Emil Cioran [1, page 207]

1 Motivation

Truthfulness matters. Handling the representation of sentences with negation has always been a challenge. [4] The compositional nature of languages causes problems with negation in the latest language models (LM) with transformer architecture. Recent studies confirm that even large LMs trained to represent the meaning expressed in sentences with or without negation are biased [10].

The impact of negation on LM tasks is most studied within the sphere of English-based natural language processing. One of the more straightforward

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 17–24, 2024.

methods to evaluate the impact is the cloze task, wherein the input sentence has one or more tokens masked, and the model predicts the missing tokens. This task is highly unconstrained, with many possible correct answers.

Multiple studies [3,6,10] inspect the model's understanding of negation by comparing the model's completions between sentence pairs with opposite polarities. Models commonly correctly complete sentences with positive polarity but generate factually incorrect continuations for the negative counterparts. Notably, the models often suggest the same prediction for both, as is illustrated in Table 1.

Table 1: Example of the LM prediction from the paper [3]. The model is insensitive to the presence of negation, which causes factually incorrect prediction.

Input Sentence	Correct Token	Predicted Token
A sparrow is a <MASK>.	bird	bird
A sparrow is not a <MASK>.	tree	bird

Although in English, this methodology showcases the model's insensitivity to negation, the expressive power of the cloze task in Slavic languages is limited. From the morphological standpoint, English belongs to the family of *analytic languages*; that is, the syntactic relations of a sentence are expressed using specific grammatical words and a rather fixed word order. Most Slavic languages, on the other hand, belong to the family of *fusional languages*, capturing the syntactic relations with affixes, allowing for looser word order. This results in the already highly unconstrained task gaining even more complexity for evaluation and the lowering of the changes for any possible meaningful value of the results.

In this paper, we evaluate the 'negation bias' in the Czech language and confirm it using several multilingual language models.

2 Methodology

Natural Language Inference (NLI) [7] is a valuable task for evaluating models' understanding of negation. In this task, the model is given a text pair. One of the texts, commonly referred to as a *premise*, contains a knowledge base that is to be taken as truthful. The model is then evaluated based on its ability to correctly determine whether or not the other text from the pair, commonly known as the *hypothesis*, is supported by the information presented in the premise. The hypothesis is then classified as one of three categories based on its textual entailment: it is either supported by the premise, refuted by it, or there is not enough information.

In our experiments, we aim to evaluate the sensitivity of the models' accuracy based on the *presence of negation* in the hypothesis.

We create paired evaluation data – each example comes in two variants. Both variants share the same premise, and the hypotheses are negations of each other, one being supported by the premise and the other refuted by it.

```
[
    {
       "role": "system",
     "content": "You are a fact checker for queries in the Czech language.
     You will be given a premise, which you know is factually correct, and
     a hypothesis. You will return the truth value of the hypothesis, based
     on the premise. Return True if the hypothesis is correct and False if
     the hypothesis is incorrect."
    },
    {
      "role": "user",
      "content": [
        "Premise: Antigua a Barbuda. Jméno zemi dal Kryštof Kolumbus
        v roce 1493 po objevení ostrova na počest Panny Marie La Antigua
        v sevillské katedrále.",
        "Antigua a Barbuda nebyla rodištěm Kryštofa Kolumba."
      ]
    }
]
```

Fig. 1: Example of the evaluation prompt with the system message, containing the task explanation, and the user message, containing the premise and the hypothesis for evaluation.

If the model processes negations correctly, its accuracy should remain the same regardless of whether the hypothesis contains negation or not.

3 Data

Our aim was to evaluate the language model's ability to correctly determine which sentence of the provided hypothesis pair is correct given the specified premise. The first step was the location of a suitable test dataset. There is a couple of Czech NLI test datasets, for our experiments we have used a modified version of the CsFEVER dataset [11].

Each entry in the original CsFEVER dataset contains the following:

- entry id,
- label of the textual entailment category,
- hypothesis (called claim in this dataset), and
- premise (called evidence in this dataset).

First, we removed any entries with the "Not enough information" label and entries containing empty values. Second, we have created a pipeline to generate hypothesis pairs for each premise.

3.1 Sentence Negation Pipeline

As there are many ways to express negation in the Czech language, with varying levels of complexity, we have focused on the method of expressing syntactic negation using a negative morpheme (prefix "ne") to negate the verb within the hypotheses.

The pipeline consists of the following:

1. The hypothesis is first tagged using the UDPipe service [9], a pipeline for tokenization, tagging, and lemmatization. We have used the current Czech model, `czech-pdt-ud-2.12-230717`.
2. We extract the first verb in the hypothesis – a word with either the VERB or AUX tag. AUX stands for the auxiliary verb, in Czech it is the verb "to be".
3. We analyse the extracted verb using the morphological analyser Majka [8], obtaining the verb's lemma and morphological tags. We filter the results to include only verbs to avoid picking the wrong homograph.
4. We modify the tags to reverse the polarity of the verb.
5. We generate the negated version of the verb based on the lemma and the modified tags. We return the verb with the value with the value of the original polarity.
6. We create a copy of the original hypothesis, replacing the verb with its negated version.
7. Based on the textual entailment label and the value of the original polarity, we determine which hypothesis is the truthful one.

The created dataset was named CsNoFEVER and is structured as described on Figure 2.

The naïve implementation of the sentence negation pipeline necessitated manual fine-tuning of the dataset – for hypotheses containing general or strong negation, multiple words beside the verb are impacted and need to be modified. The pipeline also removes non-parsable sentences from the final dataset. The final dataset currently consists of 2,600 entries, in comparison to the original number of 10,000 entries. The final dataset CsNoFEVER is available in the GitHub repository [12].

4 Results

Using CsNoFEVER dataset specified in Section 3 and the methodology specified in Section 2, we have evaluated the NLI task on language models listed in Table 2.

Table 3 summarizes our main results. We show that the inclusion of negation in the evaluated hypotheses has a definite negative impact on the models'

```
{
  "id": 10567,
  "premise": "Google Search, běžně označovaný jako Google Web Search nebo
  jednoduše Google, je internetový vyhledávač vyvinutý společností Google.
  Patří sem synonyma, předpovědi počasí, časová pásma, burzovní kotace,
  mapy, údaje o zemětřesení, časy promítání filmů, letiště, seznamy domů
  a sportovní výsledky.",
  "positive_hypothesis": "Vyhledávač Google zobrazuje informace o domovech.",
  "negative_hypothesis": "Vyhledávač Google nezobrazuje žádné informace o
  domovech.",
  "correct_polarity": "P"
}
```


Fig. 2: Structure and contents of an entry of the CsNoFEVER dataset: `id` of the entry in the original CsFEVER dataset, premise `premise`, pair of hypotheses `positive_hypothesis` (P), `negative_hypothesis` (N), and `correct_polarity` of the hypothesis ("P" or "N").

Table 2: Attributes of language models evaluated during the experimentation. All models (*Mistral-Nemo*, *Qwen2.5*, and *Llama-3.1*) are open-source and downloadable from the HuggingFace website [5].

Model Name	Size	Layers	Languages	Developer(s)
`mistralai/Mistral-Nemo-Instruct-2407`	12.20 B	40	11+	Mistral AI
`Qwen/Qwen2.5-7B-Instruct`	7.62 B	28	29+	Alibaba Cloud
`meta-llama/Llama-3.1-8B-Instruct`	8.06 B	32	8+	Meta

accuracy, with different degrees of severity. It confirms that the study of Truong et al. [10] also holds for Slavic languages such as Czech, where it exemplifies even more than in the languages with the fixed word order.

In the paper's Appendix A, we list several examples of model reasoning and errors they make. It is yet to be investigated to which extent the reasons for errors are primarily due to architectural constraints of transformer architecture such as compositionality of token processing, model size, number of layers, balancing of languages used for training, the models' training parameters, or just because of suboptimal prompting.

Table 3: Model accuracy (P = positive hypotheses, N = negative hypotheses).

Model Name	P Accuracy	N Accuracy
Mistral-Nemo	79.81% (2075)	38.73% (1007)
Qwen2.5	87.38% (2272)	76.15% (1980)
Llama-3.1	71.35% (1855)	51.35% (1855)

"Knowledge is two-fold, and consists not only in an affirmation of what is true, but in the negation of that which is false." Charles Caleb Colton [2]

5 Conclusion and Future Work

In this paper, we have investigated the difficulties language models have in correctly carrying out the NLI task on texts that include negation. We have created a new dataset for our evaluation, CsNoFEVER, modifying the CsFEVER test dataset and expanding it to contain more instances of negative hypotheses. Evaluating a number of language models, we have shown that even the latest language models, within the tested size range, have trouble correctly classifying texts with negation.

Due to the simplicity of the sentence negation pipeline introduced in this paper, the CsNoFEVER currently does not contain complex expressions of negation. We plan to further develop the negation pipeline, allowing it to parse and generate more complex cases of negation for future iterations of the dataset CsNoFEVER.

References

1. Cioran, E.: The Temptation to Exist. Quadrangle Books (1956), https://dl1.cuni.cz/pluginfile.php/1176322/mod_resource/content/1/Cioran_Temptation.pdf
2. Colton, C.C.: Lacon, or Many Things in Few Words. London: Longman, Hurst, Rees, Orme and Brown (1820)
3. Ettinger, A.: What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. Transactions of the Association for Computational Linguistics **8**, 34–48 (Jan 2020). https://doi.org/10.1162/tacl_a_00298
4. Hermann, K.M., Grefenstette, E., Blunsom, P.: "Not not bad" is not "bad": A distributional account of negation. In: Allauzen, A., Larochelle, H., Manning, C., Socher, R. (eds.) Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality. pp. 74–82. ACL, Sofia, Bulgaria (Aug 2013), https://aclanthology.org/W13-3209
5. Hugging Face – The AI community building the future, https://huggingface.co/
6. Kassner, N., Schütze, H.: Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In: Proceedings of the 58th Annual Meeting of the ACL. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.698
7. Putra, I.M.S., Siahaan, D., Saikhu, A.: Recognizing textual entailment: A review of resources, approaches, applications, and challenges. ICT Express **10**(1), 132–155 (2024). https://doi.org/https://doi.org/10.1016/j.icte.2023.08.012
8. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 13–16 (2009), https://nlp.fi.muni.cz/raslan/2009/papers/13.pdf
9. Straka, M., Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Hajič, J., Zeman, D. (eds.) Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/K17-3009

10. Truong, T.H., Baldwin, T., Verspoor, K., Cohn, T.: Language models are not naysayers: an analysis of language models on negation benchmarks. In: Palmer, A., Camacho-collados, J. (eds.) Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). pp. 101–114. ACL, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.starsem-1.10
11. Ullrich, H., Drchal, J., Rýpar, M., Vincourová, H., Moravec, V.: Csfever and ctkfacts: acquiring czech data for fact verification. Lang. Resour. Eval. **57**(4), 1571–1605 (May 2023). https://doi.org/10.1007/s10579-023-09654-3
12. Vrabcová, T.: CsNoFEVER: Negated CsFEVER dataset (2024), `https://github.com/xvrabcov/cs-no-fever`

A Examples of Models' Reasoning

This appendix contains examples showing the output and reasoning of the tested models.

A.1 Example: Scandinavian Mountain Range

Premise: Pohoří Rincon (o'odhamsky: Cew Do'ag) je významné pohoří východně od Tucsonu v okrese Pima v Arizoně ve Spojených státech. Národní park Skalisté hory (anglicky Rocky Mountain National Park) je národní park ve Spojených státech amerických, který se nachází přibližně 76 km severozápadně od mezinárodního letiště v Denveru v severní části centrálního Colorada, v pohoří Front Range of the Rocky Mountains. Mezi hlavní rysy parku patří hory, vysokohorská jezera a široká škála divoké zvěře v různých klimatických podmínkách a prostředích, od zalesněných lesů po horskou tundru. Skandinávie. Velká část skandinávských hor má klima alpínské tundry.
Positive Hypothesis (“P”): Skandinávie zahrnuje pohoří.
Negative Hypothesis (“N”): Skandinávie nezahrnuje pohoří.
Correct Polarity: “P”
Model Predictions:
Mistral-Nemo: “P”. Pohoří Rincon je významné pohoří východně od Tucsonu v okrese Pima v Arizoně ve Spojených státech, a pohoří Skandinávie je v pohoří Front Range of the Rocky Mountains, a je to významné pohoří.
Qwen2.5: “N”. Skandinávie se jedná o oblast s vlastními horami, jako jsou fjords a skandinávské Alpy, a tedy nezahrnuje pohoří.
Llama-3.1: “N”. Skandinávie je geografický region, který zahrnuje země jako Norsko, Švédsko a Finsko, a podle textu se v něm nachází alpínská tundra, nikoli pohoří.

A.2 Example: Canadian Rapper

Premise: One Dance je píseň kanadského rappera Drakea z jeho čtvrtého studiového alba Views (2016).
Negative Hypothesis (“P”): Autorem skladby One Dance je Američan.

Negative Hypothesis (“N”): Autorem skladby One Dance není Američan.
Correct Polarity: “N”
Model Predictions:
Mistral-Nemo: “P”. Píseň ’One Dance’ nenapsal Drake, vzhledem k tomu, že skladbu napsal kanadský rapper Drake.
Qwen2.5: “N”. Autor skladby One Dance, Drake, je kanadský rapper, ne americký.
Llama-3.1: “P”. Hypotéza P je správná, protože Drake je kanadský rapper a kanadský občan.

A.3 Example: Area of Yemen

Premise: Jemen je druhou největší zemí na poloostrově, zabírá 527 970 km2 (203 850 km²).
Negative Hypothesis (“P”): Jemen má rozlohu 250 750 km².
Negative Hypothesis (“N”): Jemen nemá rozlohu 250 750 km².
Correct Polarity: “N”
Model Predictions:
Mistral-Nemo: “P”. Jmenovitý údaj uvádí, že rozloha Jemenu je 203 850 km².
Qwen2.5: “N”. Podle vstupního předvolbu má Jemen 527 970 km2 rozlohy, nikoliv 203 850 km2.
Llama-3.1: “N”. Hypotéza P je nesprávná, protože uvádí rozlohu Jemenu 250 750 km², zatímco v předloze je uvedena rozloha 527 970 km².