Mapping the Landscape of Natural Language Inference: A Survey of NLI Downstream Tasks and Universality

Luca Zhou
Sapienza University of Rome
zhou.2135393@studenti.unitoma1.it

Abstract

Natural Language Inference (*NLI*) has emerged as a crucial task in natural language processing, enabling machines to understand and reason about textual data. *NLI* is considered a universal task due to its applicability across a wide range of downstream tasks in NLP and beyond. This survey provides a review of the recent literature on how *NLI* is utilized in various downstream tasks in natural language processing. Additionally, given the universality of the *text classification* task, further evaluation of the *NLI-based universal text classifier* will be conducted on two additional and particularly challenging datasets [2][3]. Empirical results show that the *NLI-based universal text classifier* performs reasonably well in these challenging zero-shot scenarios, as measured by the *balanced accuracy*.

Experiments are conducted in the *Kaggle* notebook **here**.

1 Introduction

Unlike a usual survey which digs into one specific downstream task, this survey centers around the wide applicability of natural language inference (*NLI*) and illustrates how a variety of NLP and domain-specific downstream tasks can be cast as *NLI* problems. Consequently, a well-trained general-purpose *NLI* model can be applied to solve all these tasks off the shelf.

Specifically, this work will cover the application of *NLI* in solving:

- *Universal text classification*;
- *Hallucination detection and mitigation*;
- *Hate speech detection*;
- *Symptom status recognition*

Additionally, given the universality of the task of *text classification*, we will design and conduct additional benchmarks on the universal NLI-based text classifier proposed in [8], on two additional challenging datasets [2][3]. Experimental results corroborate that the *NLI-based universal text classifier* is a versatile off-the-shelf tool with reasonable performance in zero-shot scenarios.

As of May 2024, all works involved in this survey have been released since 2022, and preference is given to the most recent when similar works are found.

2 Natural Language Inference

Natural language inference, also known as textual entailment, is the task of determining the relationship between two pieces of text: a premise and a hypothesis. In NLI, the premise is typically a piece of text that provides context or information, while the hypothesis is a statement that may or may not logically follow from the premise. The goal of NLI is to classify the relationship between the premise and hypothesis into one of several categories, such as **entailment** (the hypothesis logically follows from the premise), **contradiction** (the hypothesis is logically incompatible with the premise), or **neutral** (there is no logical relationship between the premise and hypothesis). As this survey will show, NLI is a fundamental task in natural language understanding and has applications in various areas even beyond natural language processing.

For instance, to leverage an NLI model to solve *text classification*, a straightforward pipeline is to generate premise-hypothesis pairs with the piece of text to classify as the premise, and different hypotheses corresponding to different classes. Finally, the class with the highest entailment score is the predicted class.

Concretely, given a piece of text t, assume the task is to classify it as either spam or non-spam. The corresponding NLI example consists of two premise-hypothesis pairs:

- Premise: $\{t\}$
- Hypothesis: $\{t\}$ *is spam.*

- Premise: $\{t\}$
- Hypothesis: $\{t\}$ *is not spam.*

Then, the class with the higher entailment score is regarded as the predicted class. Leveraging this mechanism, we can evaluate the classification performance of an *NLI* model.

3 NLI Downstream Applications

This section reviews numerous applications that benefit from a well-performing *NLI* model. We will start from the most general until the most domain-specific task, illustrating how to harness NLI within the process cleverly.

3.1 Universal Text Classifier

Text classification is arguably the most general NLP task where given a piece of text, the goal is to classify it into one of the predefined classes. The term *universal* indicates that the model can classify text from any domain and into any number of classes, whereas general text classifiers are trained specifically on a specific domain with a fixed number of classes.

The most recent work proposed in [8] builds an NLI-based universal text classifier that can perform text classification off-the-shelf without fine-tuning on the domain data. At the core of the approach, it frames text classification as a binary NLI problem (entailment, non-entailment), and generates different hypotheses through verbalization. The given text is utilized as the common premise of all premise-hypothesis pairs, while the hypotheses all follow the same textual template with only the class word differing. An illustrative example was already shown at the end of the previous section. The authors provided several variants of *DeBERTa V3* [6] models pretrained on NLI datasets and more. In particular, later in this survey, we will investigate the performance of two of these variants: one is trained on 5 NLI datasets and one is trained additionally on 28 synthetic NLI datasets derived from text classification datasets via the verbalization trick.

3.2 Hallucination Detection & Mitigation

In the context of natural language processing, hallucination refers to the phenomenon where a language model generates text or predictions that are not supported by the given context or are based on false or incorrect assumptions. This section explores how NLI can benefit the detection and mitigation of such phenomenon in summarization, knowledge-to-text generation, and large language models.

3.2.1 Hallucination Text Summarization

Text summarization is the task of distilling the key information from a longer piece of text to create a concise and coherent summary. It involves identifying the most important concepts, ideas, and arguments within the original text and presenting them in a condensed form while preserving the overall meaning and intent.

Lattimer et. al exploit NLI to detect hallucination in generated summaries [7]. Specifically, they propose *SCALE*, a training-free text chunking approach for factual inconsistency detection. Given a large source document and a summary, *SCALE* splits both into chunks consisting of one or more sentences. Chunks of the source document and of the summary will play the role of premise and hypothesis, respectively. Each hypothesis will be validated against all premises using an NLI model. The overall entailment score of a hypothesis h is the maximum entailment score of h against all source premises. As an additional feature, *SCALE* is interpretable since for each summary chunk (hypothesis) it is possible to recover the source chunk (premise) that entailed it.

3.2.2 Hallucination in Knowledge-to-Text Generation

Knowledge-to-text (K2T) generation is the task of synthesizing coherent text that accurately reflects a knowledge source, such as a database or a knowledge graph. In this context, hallucination occurs when the generated text contains inaccuracies or information not supported by the knowledge source.

Qiu et al. propose *TWEAK* [10], a framework incorporating an NLI verification step within a *K2T* generator that is agnostic to the nature of the generator. They treat the knowledge source as triples $\langle$*subject, relation, object*$\rangle$, and verbalize these triplets into one piece of text, which will be the NLI premise. On the decoder side of the *K2T*

generator, at each decoding step *T*, we generate two hypotheses. First, the backward hypothesis is the text generated up to the current step *T*. Second, the forward hypothesis is the text generated thus far concatenated to the text generated afterwards. These two hypotheses are validated against the premise using a pretrained NLI model, and the overall faithfulness score is a linear combination of the two entailment scores. By executing the generation multiple times, we can pick the generated text enjoying the highest overall faithfulness. In a sense, *TWEAK* reduces the hallucination of the generated text.

3.2.3 Hallucination in LLM

Rather than focusing on a specific task, *Lei et al.* propose a novel framework [9] for refining LLM-generated text by removing hallucinations that are unsupported by the source text. The framework consists of a *detection agent* and a *mitigation agent*, responsible for detecting and removing hallucinations, respectively. Let *X* be the source text and y_{raw} be the generated text, the *detection agent* employs sentence-level and entity-level NLI and extracts all sentences deemed ungrounded and the reason why they are deemed so. Subsequently, the *mitigation agent* receives the source text *X*, the original generated text y_{raw}, the hallucination sentences with their corresponding reasons, and produces $y_{refined}$ by removing hallucinations.

3.3 Faithfulness Evaluation

NLI can also serve well for evaluation purposes. A naive way of applying NLI to faithfulness evaluation is to treat the whole source document as the premise and the summary as the hypothesis. However, existing NLI models might struggle with extremely long documents. A clever solution is proposed by *Zhang et a.* in [11] to eschew the computational workload of using the entire document as the premise. The idea is to divide the source document and the summary into sentences and conduct a more fine-grained evaluation at a sentence-to-sentence level. Starting from *M* document sentences *D* and *N* summary sentences *S*, an entailment matrix *E* of shape *MxN* is constructed via an NLI model, where $E_{i,j}$ represents the entailment score with M_i and S_j respectively being the premise and hypothesis. For each summary sentence, we rank the document sentences according to the entailment score. We then initialize the premise as an empty set and iteratively insert the most entailed document sentence to the growing premise. The goal is to obtain a shorter version of the source document and use that as the premise to validate the summary. The iterations halt when the neutral score returned by the NLI model starts to increase since it presages a shift in the entailment relation. Finally, we evaluate the faithfulness of the summary by using only the restricted document as the premise, successfully bypassing the computational infeasibility of feeding the whole document as the premise.

3.4 Hate Speech Detection

Hate speech detection is a branch of natural language processing (NLP) and computational linguistics focused on identifying and categorizing text that contains hateful or abusive language, typically targeting individuals or groups based on characteristics such as race, ethnicity, religion, gender, sexual orientation, disability, or nationality. Again, one naive way of exploiting NLI here is to feed the source text into an NLI model as the premise and verbalize "hate speech" and "non-hate speech" into hypotheses. *Goldzycher et al.* proposed in [5] a more astute framework for hypothesis engineering that enhances hate speech detection performance of pretrained NLI models. The trick is to employ the NLI model also on additional auxiliary hypotheses that complement the main hypothesis (i.e. the naive hypothesis). The authors propose 4 strategies for detecting more nuanced hate speech by checking if some groups are targeted in the text, if the text contains counter-speech (e.g. when a piece of text cites hate speech but itself is not), if the text is self-directed, and if the text contains dehumanizing comparisons. These strategies, followed by the main hypothesis, are easily extendible and are applied on a cascade as decision rules. A piece of text is classified as hate speech if any stage returns a positive outcome towards hate speech.

Goldzycher et al. also proposed enhancement tricks [4] for NLI-based speech detectors in data-scarse languages. The goal is to obtain an NLI-based hate speech detector in a language with limited NLI data on hate speech. Starting from a backbone LLM, several approaches are proposed, three of which are the most promising. First, the LLM is pretrained on English hate speech detection where data is plentiful and then is fine-tuned on target-language hate speech data. Second, the

LLM is pretrained on general NLI in the target language and then fine-tuned on target-language hate speech. Third, leverage additional auxiliary hypotheses as illustrated in the previous paragraph.

3.5 Symptom Status Recognition

Deeper into the domain-specific end, NLI can even benefit the healthcare industry. Symptom status recognition is the task of recognizing the relationship between a symptom and a patient, given the medical dialog. For example, it attempts to answer questions in the form of 'Does patient *A* have symptom *S*?'. Expectedly, it is infeasible to hire human experts to recognize symptoms from a huge amount of documents. Therefore, automated techniques are highly craved, and NLI comes in handy.

Chen et al. proposed *KNSE*[1], a framework to solve symptom status recognition via NLI formulation. As a pre-step, we assume the existence of a symptom extractor that retrieves the symptom names from the medical dialog. Then, given one of the symptoms, the goal is to classify its relationship with the patient as either positive, negative, or undefined. The way to formulate this problem as an NLI task is by treating the medical dialog and symptom knowledge as the premise, whereas the hypothesis is in the form "The patient has {symptom}". More precisely, the symptom knowledge is a general description of the symptom obtained by querying *ChatGPT* or any other source. By padding the hypothesis with learnable token embeddings before and after, the authors empirically show the improved performance. The overall NLI model is trained on the *Chinese Medical Documents* Dataset and has *BERT* as the backbone and combines it with *bidirectional gated recurrent unit* and a classification head at the end.

4 Datasets and Benchmarks

Given the universality of *text classification* in natural language processing, this section reviews the datasets used for evaluating the ***universal NLI-based text classifier*** [8]. These datasets cover a broad range of domains with diverse objectives and are listed below:

Amazon polarity, IMBD, AppReviews, YelpReviews, RottenTomatoes, EmotionDAIR, EmoContext, Empathetic, Financial Phrasebank, Banking77, MASSIVE, WikiToxic, HateOffensive, HateXplain, BiasFrames Offensive, AG News, Yahoo Topics, True Teacher, Spam, Wellformed Query, Manifesto, OTU, MultiNLI, Fever NLI, ANLI, WANLI, LingNLI.

Among these datasets, some might be used for training, depending on the model configuration. ***Balanced accuracy*** is used as the evaluation metric, which is formally expressed as the mean of sensitivity and specificity:

$$\textbf{Balanced Accuracy} = \frac{Sensitivity + Specificity}{2} \tag{1}$$

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \tag{3}$$

5 Existing Code

The authors transparently provided the code for the entire pipeline which includes the following stages:

1. Data construction;
2. Data cleaning;
3. Data formatting as NLI;
4. Training and evaluation;
5. Result visualization;

Moreover, several variants of the *DeBERTaV3* [6] model are pretrained and released on *Huggingface*, along with detailed instructions for usage. These variants differ by the volume of parameters and training data. Specifically, we will evaluate

- *DeBERTa-v3 large* trained only on 5 NLI datasets
- *DeBERTa-v3 large* trained on 5 NLI datasets + 28 Synthetic NLI datasets derived from text classification data

on two additional datasets

- GoEmotions [2]
- Moral Stories [3]

6 NLI-Based Universal Text Classifier Evaluation

The choice of the two model variants is not arbitrary. The rationale is that their comparison reveals the worthiness of training the model on 28 additional synthetic NLI datasets. ***GoEmotions*** [2] is a challenging emotion classification dataset where 28 are the possible emotion categories for a piece of text. Each sample text might correctly be categorized into one or two emotions, the goal of the experimentation is to examine the accuracy and recall of the universal classifier on this task, which is a more holistic evaluation than those conducted in the original work. Secondly, the ***Moral Stories*** [3] dataset contains examples as tuples comprising:

- *norm*
- *situation*
- *intention*
- *moral action*
- *moral consequence*
- *immoral action*
- *immoral consequence*

Each example essentially describes a situation where an agent finds himself in, and two possible actions the agent can take, one of which is moral and one is immoral. To frame it as a text classification benchmark, we will create the task of classifying an action as moral or immoral given the situation, the norm, and the action taken. The reformulated dataset contains 24000 examples, each consisting of a norm, a situation, an action taken by the agent, and the gold label (moral, immoral). We concatenate all attributes except for the label as the NLI premise and verbalize the label using the template *"The action done in the text is {}"*. This task is challenging due to the necessity of moral reasoning, an abstract rather than factual concept.

7 Comparative Evaluation

For the sake of fairness of comparison, we opt to abide by the same evaluation method adopted in the original work by using ***balanced accuracy*** as the assessment metric to gauge the performance of the two pretrained models on *GoEmotions* and *Moral Stories*.

7.1 GoEmotions Results

Due to time constraints, the evaluation is not performed on the entire dataset of over 200k examples. We randomly sample 3000 examples and evaluate the balanced accuracy in the following way. First, we count the true positive, true negative, false positive, and false negative for all 28 classes. Then, for each class separately, we build its confusion matrix and the balanced accuracy. Finally, we compute overall balanced accuracy as the average balanced accuracy across classes. The above is repeated for both models and the results are depicted below.

Model	Sensitivity	Specificity	Balanced Acc.
NLI Only	0.308	0.792	**0.550**
NLI + Synthetic	0.2809	0.845	**0.563**

Table 1: Metrics Evaluation

Additionally, to gain more insights into the models' behavior, we can visualize the performances of the two models for different classes explicitly.

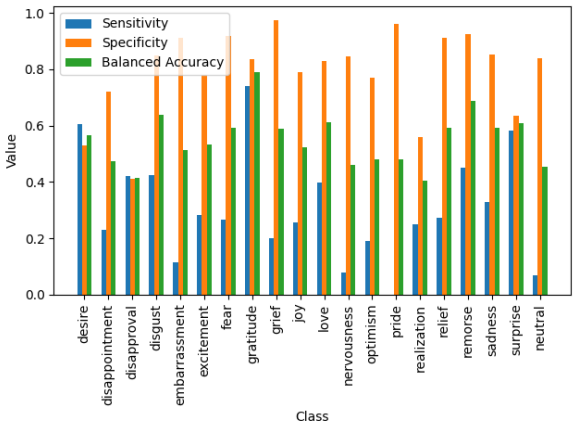

Figure 1: NLI Only Model

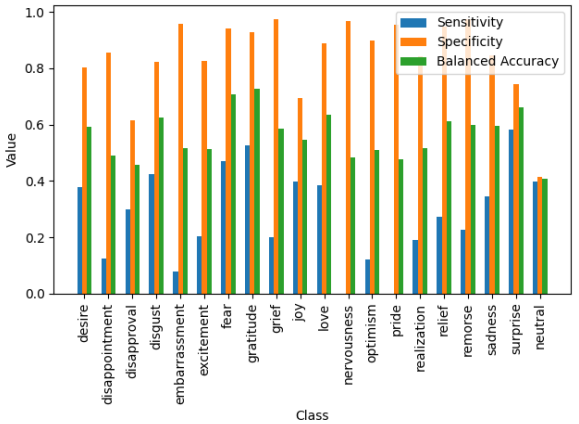

Figure 2: NLI + Synthetic Model

For a clearer comparison, we illustrate the direct comparison of the two models in terms of balanced accuracy.

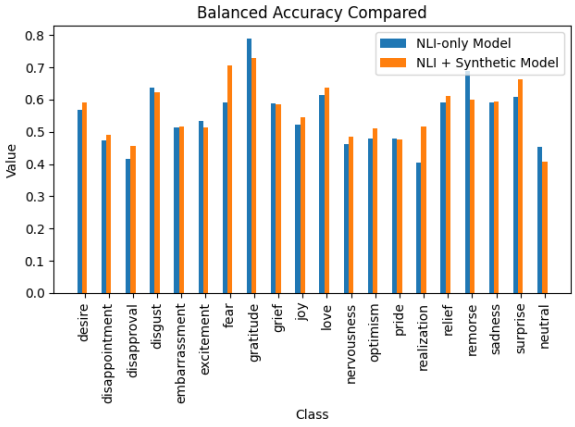

Figure 3: Balanced Accuracy Comparison

It can be observed that the model variant trained with additional synthetic NLI data tends to perform slightly better for most classes, which validates the advantage of training with additional derived data.

7.2 Moral Stories Results

Again, we first compute the confusion matrix, which this time can be easily visualized, given the binary nature of the classification task. Below are the confusion matrices of the two model variants.

	Predicted Moral	**Predicted Immoral**
Actual Moral	8939	3061
Actual Immoral	4215	7785

Table 2: Confusion Matrix of *DeBERTa-V3* trained on NLI data only

	Predicted Moral	**Predicted Immoral**
Actual Moral	10263	1737
Actual Immoral	6356	5644

Table 3: Confusion Matrix of *DeBERTa-V3* trained on NLI data + synthetic NLI data

By applying equations 2-3-1, the above tables result in the following metrics.

Model	**Sensitivity**	**Specificity**	**Balanced Acc.**
NLI Only	0.745	0.649	**0.697**
NLI + Synthetic	0.855	0.470	**0.663**

Table 4: Metrics Evaluation

7.3 Discussion

The results from the previous section, albeit seemingly unimpressive compared to task-specific state-of-the-art models, underline the versatile nature of the *NLI-based universal classifier*. Note that experiments were carried out in zero-shot settings, where the test domains were not privy to the models. Further, the two datasets [2][3] were by no means trivial in nature: *GoEmotions* contains 28 classes whereas *Moral Stories* entails abstract morality thinking. For both benchmarks, the model variant pretrained on NLI data plus additional synthetic NLI data slightly outperforms the counterpart trained only on pure NLI data. This outcome supports the intuition that pretraining NLI models with additional pseudo-NLI data derived from non-NLI datasets is a promising practice for further performance gains.

In any case, the strength of universal text classifiers lies in their versatility and ease of use, and the one examined in this survey, proposed by *Laurer et al.* [8], achieves just that.

8 Conclusion

Given the broad applicability of *natural language inference*, this survey shed light on its recent applications in an array of downstream tasks across varied domains, spanning from text classification, hallucination detection and mitigation, text faithfulness evaluation, hate speech detection, and symptom status recognition, all of which either exploit *NLI* as an intermediate step or reframe the problem into *NLI*. Finally, we assess the *NLI-based universal text classifier* proposed by *Laurer et al.* [8] on two new challenging benchmarks. The performance, although falling short compared to the state-of-the-art domain-specific classifiers, are still indicative of the versatile nature of the approach. By testing two variants of the same model, we also verified the benefit of training NLI models with synthetic NLI data extracted from non-NLI data.

References

[1] Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuanjing Huang. Knse: A knowledge-aware natural language inference framework for dialogue symptom status recognition. *arXiv preprint arXiv:2305.16833*, 2023. 4

[2] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020. 1, 4, 5, 6

[3] Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*, 2020. 1, 4, 5, 6

[4] Janis Goldzycher, Moritz Preisig, Chantal Amrhein, and Gerold Schneider. Evaluating the effectiveness of natural language inference for hate speech detection in languages with limited labeled data. *arXiv preprint arXiv:2306.03722*, 2023. 3

[5] Janis Goldzycher and Gerold Schneider. Hypothesis engineering for zero-shot hate speech detection. *arXiv preprint arXiv:2210.00910*, 2022. 3

[6] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. 2, 4

[7] Barrett Martin Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. Fast and accurate factual inconsistency detection over long documents. *arXiv preprint arXiv:2310.13189*, 2023. 2

[8] Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*, 2023. 1, 2, 4, 6

[9] Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023. 3

[10] Yifu Qiu, Varun Embar, Shay B Cohen, and Benjamin Han. Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. *arXiv preprint arXiv:2311.09467*, 2023. 2

[11] Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. *arXiv preprint arXiv:2402.17630*, 2024. 3