# Sentiment Analysis Towards Named Entities with Explainability Techniques
# Project Proposal for NLP Course, Winter 2024

**Patryk Rakus**
Warsaw University of Technology
patryk.rakus.stud@pw.edu.pl

**Filip Szympliński**
Warsaw University of Technology
01161601@pw.edu.pl

**Julia Kaznowska**
Warsaw University of Technology
julia.kaznowska.stud@pw.edu.pl

**Michał Tomczyk**
Warsaw University of Technology
01161608@pw.edu.pl

**supervisor: Anna Wróblewska**
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

## Abstract

This project aims to enhance sentiment analysis by focusing specifically on Named Entities (NEs), such as brands, individuals, or organizations, and examining the sentiment associated with these entities in text. Traditional sentiment analysis often lacks the precision to assess sentiment towards specific entities. Furthermore, current models are frequently "black boxes", providing little insight into how sentiment predictions are made. This research addresses these gaps by developing an NE-focused sentiment analysis system integrated with explainability techniques. The project's novelty lies in combining targeted entity sentiment analysis with interpretable machine learning methods (e.g. LIME, SHAP, and attention-based mechanisms), thus contributing to both natural language processing (NLP) and explainable AI (XAI). The expected outcome is a robust, interpretable framework that can improve accuracy and transparency in entity-specific sentiment detection, which will be beneficial in applications like brand monitoring, public relations, and risk assessment.

## 1 Introduction

### 1.1 Scientific goal of the project

The primary objective of this project is to advance sentiment analysis by focusing on the sentiments specifically directed towards NEs in textual data. Unlike traditional sentiment analysis, which evaluates the sentiment of an entire text, entity-focused sentiment analysis hones in on individual subjects (e.g., brands, individuals, organizations) within a text, distinguishing positive, negative, or neutral sentiment for each. The project aims to address the following research questions:

- Can a sentiment analysis model be designed to target Named Entities with greater precision than traditional methods?
- How can explainability techniques be incorporated to enhance transparency in sentiment prediction for NEs?

The project hypothesizes that integrating NE-focused sentiment analysis with explainable models will increase the interpretability and usability of sentiment predictions in real-world applications.

### 1.2 Justification and impact

This project is pioneering in its dual focus: entity-specific sentiment analysis and explainability. By tackling the interpretability of sentiment models at the entity level, the project not only enhances model transparency but also fills in a gap in real-world applications such as social media analysis, reputation management, and market analysis, where understanding sentiments directed at specific entities is crucial. The demand for ethical and explainable AI is growing, which makes the project timely and aligned with global trends. The project's outcomes are expected to benefit both NLP and XAI, potentially influencing future research and industry practices in both fields.

## 2 Literature review

### 2.1 State of the Art

Sentiment analysis has been a prominent research field within NLP. It focuses on classifying the emotions that were expressed in a text, starting from differentiating whether the text is positive, negative or neutral, through naming more complex emotional states like joy, sadness or anger (Pang and Lee, 2008). The techniques used for this task have changed significantly throughout the years. First attempts to sentiment analysis were based in rule-based systems, especially pattern-matching (Turney and Littman, 2003). With the introduction of more advanced machine learning techniques, classifiers and statistical models have been used more commonly (Abirami and Aa, 2016). With deep learning models and neural networks being introduced, as well as a growth in computational power, the field has changed drastically and those techniques were stared to be used (Kim, 2014).

The emergence of transformer-based architectures, has revolutionized the field. One of the most powerful one is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), which serves as the foundation for most SOTA solutions due to its bidirectional contextual embeddings enable precise sentiment classification by understanding the relationships between words and their surrounding context. Fine-tuned models like SciBERT (Beltagy et al., 2019) have shown exceptional performance in domain-specific tasks, such as analyzing sentiment toward financial organizations or scientific entities, respectively. These models enhance entity sentiment analysis by integrating named entity recognition (NER) tasks into a single transformer framework.

Generative Pre-trained Transformer (GPT) (Radford and Narasimhan, 2018) is a very impactful model that changed a lot not only in the field, but also in a day-to-day life of the public. Unlike BERT, it uses a unidirectional approach for text interpretation and generation. It has been trained on a large corpus of internet data, combining unsupervised pre-training and supervised fine-tuning. Recent models combine NER and sentiment classification in multitask learning frameworks. This approach reduces error propagation and improves sentiment attribution.

There are quite a few challenges in the field of NLP and sentiment analysis. Models have problems with the complexity of nuances and context, especially with ambiguity and linguistic context, sarcasm and irony, and multilingual or bilingual texts (Gupta et al., 2024), which diminish the accuracy of predictions. More advanced models, while improving the understanding of those challenges, are highly computationally complex, as they require substantial amount of resources and memory. Training them also involve the usage of huge datasets and powerful hardware. The challenges also remain in analyzing sentiments directed at specific entities within a text. Approaches like Aspect-Based Sentiment Analysis (ABSA) (Hua et al., 2024) have been partially effective but are limited in scope and interpretability.

Similarly, current XAI techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) which perform perturbation of the input text sequence by hiding one word and check how the predictions differ. Also Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017), has been applied to broader classification models but those methods are rarely adapted specifically for NE-focused sentiment analysis.

Another technique used in model interpretability are conterfactual explanations (**?**). This method involve altering input text to observe changes in the output prediction. This approach identifies minimal changes required to flip sentiment predictions, offering insights into model sensitivity and robustness for input data changes.

In transformer-based models, the attention mechanisms is widely used (Voita et al., 2019). Attention weights indicate which parts of sentence the model focuses on during a prediction. While not explicitly designed for explanation, attention maps are often interpreted to infer decision-making processes. However, while attention mechanisms are effective, lack sufficient transparency in sentiment-oriented contexts.

### 2.2 Available datasets

As already mentioned, both sentiment classification and NER are important subtasks of NLP, gaining more and more popularity in recent years. Because of that, various datasets are available in the public domain, facilitating further development in these areas. It has to be noted that while sentiment analysis and NER can be combined into a single task, many available resources focus primarily on one of them. Thus, most datasets are

designed to be used in either of these tasks. Nevertheless, a few different datasets might be used for different parts of the project focused on a specific topic. Furthermore, a dataset focusing mostly on the named entities can be helpful for the task of sentiment classification and vice versa. One popular source of data, not only for the above tasks, but for the whole field of NLP, is SamEval - an annual series of workshops concentrating on sentiment analysis. Each year the creators publish a list of tasks and the datasets to be used for training and evaluating models related to them. In 2022 and 2023 two editions of the task concerning named entity recognition were published, thus providing with two good quality datasets, each consisting of 36 entity classes and 12 languages (Malmasi et al., 2022). Another task in 2022 was associated with Structured Sentiment Analysis. There, the authors provided seven datasets, with the data in five languages, all concerning sentiment classification (Barnes et al., 2022). Especially noteworthy is the OpeNER dataset (Agerri et al., 2013), focusing on both the sentiment analysis and NER. It consists of hotel reviews in six languages. Another high quality NER dataset is WNUT-2017 (Derczynski et al., 2017), collecting data from various social networks (e.g. Twitter, Reddit, YouTube and StackExchange) and focusing on different categories of named entities and identifying entities not usually considered.

## 3 Methodology and Technologies Used

### 3.1 Methodology

The methodology associated with the project consists of several key stages, each with defined steps to ensure the correctness of the results. Defining all necessary steps needed to achieve satisfactory results before the start of the implementation will help with planning the work and will allow to identify potential challenges. The overview of the methodology is as follows:

1. Data acquisition - the most important factor regarding the quality of the trained model is relevant data of high quality. The appropriate data sources need to be identified and selected. Easy access to chosen data sources must be ensured as well. Then, suitable data collection techniques must be implemented. Two possible and commonly used techniques are collecting the data from APIs and the use of web scraping. Both methods allow both real-time and batch processing, depending on the needs. Other possibility is downloading ready datasets, in the forms of files (most commonly in the .csv or .xlsx formats). On this stage, recognizing the named entities in the data should be applied (as well as filtering the data for relevance), to ensure the focus on NE-sentiment analysis. This can be achieved by using ready NLP frameworks and tools.

2. Data preprocessing - to ensure the quality of the acquired data (and thus the quality of the model training), it needs to be processed. This can consist of removing the noise (such as irrelevant characters), text tokenization, lemmatization, normalization and Named Entity tagging.

3. Model selection - choosing the right network is crucial aspect of the quality of the solution. Having the ability to use pre-trained models, it is possible to utilize state-of-the-art techniques for sentiment analysis. Identifying the most suitable model will be achieved based on various factors, including overlooking the architecture of the model, its complexity and explainability prospects, the specific dataset on which it was trained, the specific task for which it is desired and the quality of the produced results. The chosen model will additionally have to be adapted for the contextual information around named entities. This might include contextual embeddings or attention mechanisms focusing on text surrounding the entities. The primary focus will be on fine-tuning pre-trained models. However, the existing and available resources might not achieve the desired effects when focusing on the specific entities and text domain. In this case, custom models could be used to reach more satisfactory results.

4. Feature engineering and model training - In the final model classifying the sentiments, the words embeddings will be used for fine-tuning and the predictions will be made on them as well. Thus, representing the text as vector will be a crucial task during the feature extraction. For that task, another pre-trained model will be used. Acquired embeddings will be used to train the model, with the task of sentiment classification, with the number

of classes depending on the selected datasets. During and after the training, the results will be evaluated on both the test and validation datasets, using various metrics, the choice of which will depend on the specific task and the characteristics of datasets.

5. Model explainability - various techniques will be used to identify the features that are the most influential on the predictions. This might include various model-agnostic methods, such as SHAP or LIME. Attention analysis might be applied, as most likely the selected model there will be a variation of a transformer network. For example, visualizing the attention weights might be considered. In the recent years, various techniques focused on adding the explainability specifically to sentiment analysis models were explored. These include utilizing various data augmentation technique, such as augmenting via external knowledge or with adversarial examples (Chen and Ji, 2020). Incorporating them into the designed model should be considered as well. The usage of a different network architecture might further facilitate the explainability, like for instance Contextual Sentiment Neural Network (Ito et al., 2019). Thus, explaiability has to be taken into account on all stages of the project development, starting with model selection and preprocessing techniques and ending on the evaluation of the results.

6. Results interpretation and visualization - the achieved results will be interpreted and closely examined. Different metrics will be used for the quality of the sentiment prediction and the explainability. Evaluation of the latter aspect will focus both on the ease of interpretation by humans and quality of the representation of the model's logic. Cross-validation will be applied to ensure the correctness of the conclusions. Trend analysis will also be performed, to try to detect and understand the causes of sentiment changes, both in specific entities and overall. The trends and reached conclusions will be visualized.

### 3.2 Technologies used

The project will be made using the python programming language. Frameworks commonly chosen for data analysis, processing, deep learning and visualization will be utilized, including pandas, numpy, scikit-learn, tensorflow, keras, huggingface, matplotlib, seaborn, plotly and others. For the version control, GIT will be used. The devices used for the implementation will consist of our personal computers. To facilitate cooperation during the code writing and model training, as well as to gain access to higher computational power than private devices, Google Colab will be used.

## 4 Proof of Concept

As a first step of the practical development of our project, a Proof of Concept was prepared to ensure that the task is possible to be executed. The experiments with the available resources were also performed, which resulted in gaining initial understanding of the topic selected. The initial results will be used as a benchmark for further improvements. In line with the initial assumptions, the technologies used for this stage were python programming language with necessary packages and Google Colab platform for cooperation and accessing higher computational power.

### 4.1 Tasks performed

As an initial work with the project, two separate basic models were trained for the tasks of Sentiment Analysis and Named Entity Recognition. Pre-trained networks were selected and then fine-tuned on chosen datasets. For the model selection huggingface library was used. The networks were trained using keras library.

#### 4.1.1 Named Entity Recognition model

As an initial NER model, TFBertModel was used. It is an implementation of the BERT network from huggingface. The library contains multiple pre-trained variants of this network. In this task `bert-base-uncased` was used. The model was fine-tuned on the MultiCoNER 2022 dataset, selecting only the

data in english for faster training process. It is a common benchmark for NER models. The first step was to download and parse the data. Next, tokenization was applied (using pre-trained AutoTokenizer model from huggingface) and the correct labels were aligned. Then, batched datasets from the tensorflow library were created (with the batch size of 16), and divided into train, test and validation datasets. The pre-trained BERT model was loaded and compiled. Five epochs were trained on the train dataset, with the help of the Adam optimizer.

#### 4.1.2 Sentiment Analysis model

In this task, sentiment analysis was based on splitting the sentence. The first step was to find Named Entities using NER model. Then, the context needed to be defined. It was found by extracting neighbouring words from the sentence. This way, the input for sentiment analysis model was created. Based on that, the prediction was drawn for a given Named Entity. The model used for sentiment analysis was `DistilBERT` from huggingface.

### 4.2 Results

In order to assess the initial performance of the models, several different sentences were prepared. These sentences were created with different challenges in mind (e.g. sentence (d) tests a case where both entities are next to each other and their respective context is not as straightforward to extract).

(a) I love steve jobs, but when he created the iphone 15, it was the worst phone ever

(b) Elon Musk is lovely and I enjoy Tesla company very much

(c) I hate it when Carrefour discounts all items

(d) I absolutely loved the main character, Buzz Astral, in Toy Story, but the ending of the movie was terribly disappointing

(e) Tesla's recent quality control issues have left many customers disappointed and questioning the company's commitment to excellence

#### 4.2.1 Named Entity Recognition

The tests for Named Entity recognition yielded satisfactory results. The expected and recognised Name Entities for each sentence the presented in the table 1 below.

Table 1: NER Test Results: Abbreviations used: **S** = Sentence, **NE** = Named Entity, **Recog. NE** = Recognised Named Entity

| **S** | **NE** | **Recog. NE** |
|---|---|---|
| a | steve jobs, iphone 15 | steve jobs, iphone 15 |
| b | Elon Musk, Tesla company | Elon Musk, Tesla company |
| c | Carrefour | Carrefour |
| d | Buzz Astral, Toy Story | Buzz Astral, Toy Story |
| e | Tesla | - |

The model correctly recognised Named Entities in most of the occurances. The only exception was "Tesla" in the last sentence. This case will be invetigated in the later stages of the project.

#### 4.2.2 Sentiment Analysis

The sentences from **a** to **d** were subjected to sentiment analysis. The results are presented in the table 2 below.

Table 2: Sentiment Test Results: Abbreviations used: **S** = Sentence, **NE** = Named Entity, **Es** = Expected sentiment, **Ps** = Predicted sentiment, **Pos** = Positive, **Neg** = Negative

| **S** | **NE** | **Es** | **Ps** | **Score** |
|---|---|---|---|---|
| a | steve jobs | Pos | Pos | 99.65% |
| a | iphone 15 | Neg | Neg | 96.95% |
| b | Elon Musk | Pos | Pos | 99.98% |
| b | Tesla company | Pos | Pos | 99.98% |
| c | Carrefour | Neg | Neg | 99.84% |
| d | Buzz Astral | Pos | Pos | 99.97% |
| d | Toy Story | Neg | Neg | 94.21% |

The sentiment predictions were correct in every test. The confidence levels were high, which makes the results very promising.

## 5 Work Plan

To ensure timely and successful completion of the project, the following detailed

work plan is being proposed. The timeline is divided into clear milestones to meet the requirements of the Proof of Concept (PoC), preliminary report deadline (11.12.2024), and the final submission deadline (22.01.2025).

**Data Acquisition and Preprocessing (04.12.2024 - 10.12.2024)**

- Identify and acquire datasets related to Named Entity Recognition (NER) and sentiment analysis (e.g., OpeNER, SemEval datasets).
- Perform initial data preprocessing steps:
  - Noise removal (e.g., irrelevant characters, URLs, stop words).
  - Tokenization, lemmatization, and normalization.
  - Tagging Named Entities in the datasets.

**Proof of Concept (PoC) Development (11.12.2024)**

- Develop a preliminary version of the model:
  - Fine-tune a basic pre-trained model (e.g., BERT, RoBERTa) on a subset of the data.
  - Implement basic sentiment classification for Named Entities.
- Conduct initial evaluations to validate feasibility.
- Draft the preliminary report based on findings.

**Model Selection and Enhancement (12.12.2024 - 20.12.2024)**

- Experiment with various pre-trained models
- Integrate Named Entity Recognition (NER) and sentiment analysis tasks.
- Fine-tune selected models on the full dataset.

**Explainability Integration (21.12.2024 - 03.01.2025)**

- Incorporate explainability techniques, such as:
  - LIME and SHAP for feature importance visualization.
  - Attention weight visualizations in transformer-based models.
  - Experiment with counterfactual explanations for improved interpretability.

**Evaluation and Refinement (04.01.2025 - 15.01.2025)**

- Evaluate the model's performance using metrics like:
  - Sentiment classification accuracy, precision, recall, and F1-score.
  - Explainability evaluation based on ease of interpretation and feature importance relevance.
- Fine-tune model parameters for better performance.

**Report Writing and Presentation Preparation (16.01.2025 - 21.01.2025)**

- Prepare the final project report.
- Create a presentation summarizing the project for the submission deadline.

**Submission and Final Presentation (22.01.2025)**

- Submit the final report, code, and presentation.
- Deliver the presentation to demonstrate the results and conclusions.

## References


[Abirami and Aa2016] A.M. Abirami and Askarunisa Aa. 2016. Feature based sentiment analysis for service reviews. 22:650–670, 01.

[Agerri et al.2013] Rodrigo Agerri, Montse Cuadros, Seán Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento de Lenguaje Natural*, 51:215–218, 09.

[Barnes et al.2022] Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International*

*Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States, July. Association for Computational Linguistics.

[Beltagy et al.2019] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

[Chen and Ji2020] Hanjie Chen and Yangfeng Ji. 2020. Improving the explainability of neural sentiment classifiers via data augmentation.

[Derczynski et al.2017] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Gupta et al.2024] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. Comprehensive study on sentiment analysis: From rule-based to modern llm based system.

[Hua et al.2024] Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11), September.

[Ito et al.2019] Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Kiyoshi Izumi, and Tatsuo Yamashita. 2019. Csnn: Contextual sentiment neural network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1126–1131.

[Kim2014] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

[Lundberg and Lee2017] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

[Malmasi et al.2022] Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition.

[Pang and Lee2008] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 01.

[Radford and Narasimhan2018] Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

[Ribeiro et al.2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

[Turney and Littman2003] Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *CoRR*, cs.CL/0309034.

[Voita et al.2019] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned.