News sentiment analysis
Project Report for NLP Course, Winter 2023

Jakub Kozieł
Warsaw University of Technology
jakub.koziel.stud@pw.edu.pl

Jakub Lis
Warsaw University of Technology
jakub.lis2.stud@pw.edu.pl

Bartosz Sawicki
Warsaw University of Technology
01151408@pw.edu.pl

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

Abstract

We perform a sentiment analysis task using news from the Slovenian Press Agency (STA) database. We distinguish between three types of sentiment: negative, neutral and positive. The task is deeply described, as well as the elements of data processing. We present the overview of State-of-the-art machine learning models used in this task and their possible extensions. Selected explainable artificial intelligence algorithms used in natural language processing are described. We share the results of the preliminary exploratory data analysis. For testing purposes, we annotated several articles and created a test dataset. We tested models for aspect-based sentiment analysis (ABSA) such as FLAN-T5 and DeBERTa and models for document-based sentiment analysis such as FinBERT, Auditor Review Sentiment Model (fine-tuned version of FinBERT), Financial-RoBERTa, and Twitter-RoBERTa. The best results were obtained for the ABSA task while using FLAN-T5 (accuracy: 0.81, f1_macro: 0.47). There was no clear winner for document-based sentiment, but for our final solution, we picked Twitter-RoBERTa (accuracy: 0.58, f1_macro: 0.46).

1 Introduction

Nowadays, the amount of data in many companies, for instance, e-commerce companies or, particularly important in this project, press agencies, is flooded with an enormous amount of data, which manual analysis performed by humans is impossible due to its time consumption. Leveraging current state-of-the-art technology in natural language processing might bring a business that cannot be overestimated. Artificial intelligence might help journalists better understand what makes an article good, bringing the agency more readers and higher income. With a tool to supervise sentiment, it is possible to compare the style in which articles of a given category are written. Insights driven by a large volume of data may serve as an input while making strategic business decisions. Such a neutral, automatic sentiment analysis tool can help write unbiased news articles free of opinion-forming traits. This can be crucial when the objectiveness of the news described is a priority.

1.1 Project goal

This project is centred on performing sentiment analysis on the text of news articles. In our case, the data comes from the Slovenian Press Agency (STA). We aimed to deliver a set of tools that analyze sentiment in various feasible scenarios. The main goal was to propose a solution that works for the whole article and its parts and can automatically assess polarity towards different issues mentioned. The project involved leveraging methods from the field of explanatory artificial intelligence to provide reasoning for model predictions. The last stage of the project encompassed the creation of reports that show news sentiment from different perspectives, which brings the business value that the solution can provide to the press agency using it.

1.2 Project overview

Sentiment analysis is a natural language processing task determining the emotional tone or sentiment a text expresses. It typically categorizes the sentiment as positive, negative, or neutral. The project aimed to assess the sentiment analysis of news articles from the STA database. We wanted to bring out the sentiment for the entire news, for parts of it, and the various issues mentioned, providing more detailed insights. We prepared an overview of state-of-the-art techniques for performing sentiment analysis tasks, mainly focusing on news sentiment analysis. We also covered techniques for explainable artificial intelligence, which we used in our project next to other visualizations of obtained predictions. We have Slovenian and English-language news, for which we performed preliminary exploratory data analysis. However, for the project solution, we focused on English news. We annotated some of the news to create a test dataset and described the process of annotating the data. We used pre-trained models and tested them on the STA news. We present the results of our testing using metrics such as accuracy, F1-score calculated for every class (negative, neutral, positive) and F1-macro, which is an average of sub-individual F1-scores. Using models selected based on the results obtained for document-based and aspect-based sentiment analysis, we created functions to generate visualizations that help supervise sentiment by category, aspect, or considering the timing of articles. Our solution also allows to examine the prediction result through the explainability techniques.

2 Related work

2.1 Sentiment analysis

In (Wankhade et al., 2022), we can find possible current approaches to the task of sentiment analysis. Sentiment analysis could be applied on several levels. Those are document level, sentence level, phrase level, and aspect level. Those approaches gradually become more and more fine-grained in the order that they are mentioned. The document-level analysis is applied to a whole document and sentence-level to each sentence. Phrase-level sentiment analysis is mining opinion within a single sentence, where one phrase could consist of single or multiple aspects. Finally, aspect-level sentiment analysis is considered, which can deal with mixed opinions about a particular thing (e.g., a service) within a single sentence and becomes crucial when one aspect is criticized, whereas another is praised in one opinion.

(Birjali et al., 2021) discusses a generic process of sentiment analysis. As described, one can distinguish three elements of data processing: Text Preprocessing, Feature Extraction, and Feature Selection. The data preprocessing step is supposed to improve the data quality by correcting spelling and grammatical errors and, in this way, reducing the noise. Secondly, as pointed out, many words do not impact text polarity and should be removed to reduce the data dimensionality. Dispensable words include articles, prepositions, punctuation, and special characters. Frequently used Python toolkits for data preprocessing are NLTK (Bird et al., 2009) and TextBlob https://textblob.readthedocs.io/en/dev/. The survey distinguishes common tasks in the preprocessing stage: tokenization, stop word removal, Part-of-Speech tagging, and lemmatization. The next discussed step is Feature extraction, with its importance in the context of sentiment analysis explicitly highlighted. This task aims to extract valuable information, such as words that express sentiment. From the sentiment analysis perspective, the following features are used: terms' presence and frequency, Parts-of-Speech (PoS) tags, opinion words and phrases, and negations. The presence and frequency of terms are general tools for information retrieval. PoS is helpful as many methods rely on adjectives in opinion mining. Opinion words and phrases are commonly used to express opinions, and lastly, the negation words (opinion shifters), e.g., not, never, and cannot. At the end of preprocessing, feature selection is used, which could be categorized into lexicon-based and statistical methods. The first one involves human work, and even though it can offer high-quality results, creating such a lexicon (or just its core to create a basis for expanding it by synonyms) is time-consuming and costly. The lexicon is supposed to be a base for the feature set of words with strong sentiment. The latter category comprises various approaches, from applying statistical measures to leveraging machine learning models.

Additionally, (Birjali et al., 2021) mentions challenges in sentiment analysis. Those include sarcasm detection (when someone is saying or writing the opposite of what they mean), negation handling (which also reverses the polarity), word sense disambiguation (word meaning depending on a context), low-resource languages (when there was poor research done so far in this language and therefore there is a lack of linguistic resources, e.g., labeled datasets).

When dealing with mixed opinions in one piece of text, one might be interested in distinguishing sentiments towards different issues mentioned. In this situation, aspect-based sentiment analysis should be used. Aspect-based sentiment analysis comprises the following tasks: identification of aspect terms, aspect categories, opinion terms, and sentiment polarities (Zhang et al., 2023). The important aspect term extraction (ATE) task aims to extract all mentioned aspect terms in the given text, which allows us to apply the subsequent task (sentiment classification) at a more fine-grained scale than the sentence level. (Liu et al., 2020) provides an overview of state-of-the-art deep learning approaches to aspect-based sentiment analysis with their evaluation on selected datasets. Mainly, the study has shown, that the task of abstract-based sentiment analysis is domain dependent. The authors provided an example, that the meaning of the word ”easy” might have an opposite meaning depending on the domain. It is used to express positive sentiment in the electronic domain while having negative meanings associated with the movie domain. Depending on the dataset's domain, a different method has achieved the best results.

2.2 State of the art

When dealing with tasks where no annotated data is possible, one often needs to rely on open-source pre-trained on different data models. Here we provide an overview of the latest accomplishments in the field of sentiment analysis.

In this work, we test two models to perform aspect-based sentiment analysis which are Flan-T5 (Chung et al., 2022) and DeBERTa (Yang and Li , 2022). Flan-T5 has been fine-tuned on more than 1000 tasks covering many languages. DeBERTa is fine-tuned with 180k examples for the ABSA dataset (including augmented data).

Document-based sentiment analysis may be performed using fine-tuned transformer models. In this work, we focus on pre-trained versions of BERT and RoBERTa. The models were fine-tuned using different datasets: FinancialBERT trained on *Financial PhraseBank* (Hazourli, 2022), auditor_sentiment - FinBERT trained on *auditor_review* (Finance Inc., 2022), twitterRoBERTa - fine-tuned on *TweetEval* (Loureiro et al., 2022), and DistilRoBERTa trained on *Financial PhraseBank* (Romero, 2023).

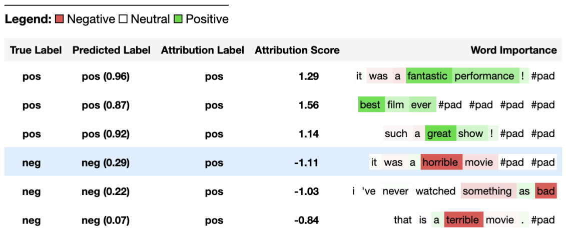
2.3 Possibilities for Slovenian language

The major part of the available pre-trained models was trained using English datasets. Multilingual models, trained on datasets of texts in different languages, are becoming more popular. Because most of our data available is in Slovenian, we tried to find a model pre-trained on documents in this language. A Slovenian NLP Benchmark is available at `https://slobench.cjvt.si/` but lacks a sentiment analysis task. We found a model pre-trained on Croatian News with metadata referring to the Slovenian language. The model is available at https://huggingface.co/FFZG-cleopatra/Croatian-Document-News-Sentiment-Classifier, but it may have low quality, as it is community-based.

Although pre-trained models for sentiment analysis in Slovenian are not widely accessible, there exist datasets for sentiment analysis in Slovenian. (Mozetič et al., 2016) prepared a dataset of tweets in this language. For our project, sentiment annotated news corpus (Bučar et al., 2018) and aspect-based sentiment news corpus (Žitnik , 2019) seem more suitable. Both datasets are publicly available. The sentiment-annotated news corpus consists of 250,000 documents with automatically detected sentiment annotation and 10,000 documents with manually detected sentiment at document, paragraph, and sentence levels. The aspect-based sentiment news corpus comprises 837 documents with 31,000 manually tagged named entities and 5-level sentiment annotation for each entity.

2.4 Explainable artificial intelligence

Explainable Artificial Intelligence (XAI) is a set of techniques enabling the interpretation of deep learning models. XAI is an emerging scientific field of research. Nevertheless, several open-source projects, like Captum (Kokhlikyan et al., 2020) or Dalex (Baniecki et al., 2021), were created to implement the most popular explanation algorithms and make them compatible with the most popular machine learning frameworks. Unfortunately, most XAI algorithms are domain-specific and work only with tabular or image data. How-



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad

Figure 1: Example of word importance obtained by using the Integrated Gradients. Source: `https://github.com/pytorch/captum/blob/master/tutorials/IMDB_TorchText_Interpret.ipynb`.

ever, few methods are model-agnostic, or their underlying assumptions are satisfied for NLP models.

In the context of Natural Language Processing and Sentiment Analysis, XAI algorithms can be used to evaluate word importance in a given model prediction. They indicate which words attribute to positive or negative prediction and to what extent. In the case of our project, the XAI diagnostics can identify words influencing predicted sentiment the most. To illustrate this capability, we included a sample output of an algorithm in Figure 1.

Most deep-learning models use gradient learning. This is exploited in the Integrated Gradients method (Sundararajan et al., 2017). The algorithm provides a way to measure feature importance by integrating the model's gradients for the input features over a path from a baseline or reference input to the actual input. For explaining NLP models, usually, the padding token acts as a baseline. After integration, it is required to sum the attribution scores across all embedding dimensions for each word/token to attain a word/token level attribution score.

Another framework suitable for explaining NLP models is Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016). The main idea behind this algorithm is to approximate the model decision boundary locally, in the neighborhood of an explained instance, by the interpretable surrogate model, for example, logistic regression. The surrogate model is then interpreted instead of the black-box deep learning model. The key challenge is to sample from the neighborhood of an instance. It is done by perturbating features of the instance. For NLP tasks, perturbation is done by removing words/tokens from the text or substituting a padding token.

3 Our dataset

3.1 Data description

In our work, we used data from the STA database, available by API access. STA database consists of articles published by the Slovenian Press Agency over the years. The API provides the list of the IDs of the news from a given day and retrieves the news text and its metadata based on the selected ID. The news is mostly in Slovenian, but English news is also available. The most important metadata that is stored are:

- authors of the news,
- headline,
- categories of the news,
- list of keywords,
- priority (1-6),
- places (including country and city),
- timestamp of the creation of the news.

Data is returned in JSON format. The authors of the news are represented only by their initials, and in the case of more than one author (which is a rather common case), they are separated by a slash. The headers are of type String and provide essential information about the content they precede. Categories are returned in a list of 2-characters string format, as one news can contain more than one category. There are 20 different categories for Slovenian news; among them, we find Kultura (Culture); Napovedi dogodkov (Schedule of Events); Mednarodna politika (International politics); Slovensko gospodarstvo (Slovenian economy); Šport (Sport), and others. For English, there are eight categories: Advisory (AD); Arts and Culture (AC); Around Slovenia (AS); Business, Finance and Economy (BE); Health, environment, science (HE); Politics (PO); Roundup (RU); Schedule of Events (SE) and Sports (ST). The list of keywords provides a more granular definition than the category. Priority means how prioritized news is, whereas four means ordinary news. Places is a list of places the news mainly concerns; it includes country, city, and codes of the country. The timestamp is in Integer format – in UNIX format and represents the specific date and time of the news creation.

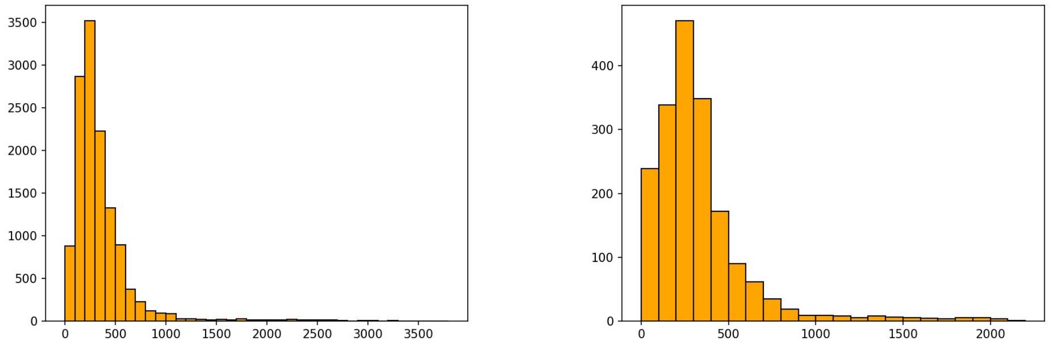
3.2 Exploratory data analysis

We downloaded data from 2 months (September and October 2023) and prepared exploratory data analysis (EDA) based on this period. It contains 12,852 Slovenian news and 1,758 English news. Figure 2 shows the number of used words in our dataset. For both Slovenian and English, more than 85% of news has fewer than 500 words, but there are some outliers in the data, and some news had more than 2,000 words or even 3,000 words for Slovenian. Figure 3 represents the number of sentences in the news. The distribution is very similar to that for the words. More than 90% of the news had less than 30 sentences. Again, we can observe outliers in the data, especially for Slovenian news, where an article had exactly 178 sentences (the article is from the Pregledi dogodkov "Event overviews" category and it summarizes events from the last week). Figure 4 shows the most common keywords of the news. We can see that the most significant number of news relate to Slovenia, regardless of the language. Other common news topics are napoved (forecast), nogomet (football), izidi (results), zda (weather), EU for Slovenian, and events, press, coverage, weather, government for English. Figure 5 shows both languages' most common news categories. We can see that for Slovenian news, the SP (Šport, "Sport"), MP (Mednarodna politika, "International politics"), and GO (Slovensko gospodarstvo, "Slovenian economy") categories prevail. For English, most news is in PO (Politics), BE (Business, Finance and Economy), and AS (Around Slovenia) categories. In data, we also have information about the places the news concerns. Figure 6 shows how much news is about a particular locality. For both languages, the most news concern is Ljubljana, the capital of Slovenia. Interestingly, the top 3 cities concerned with English news are Slovenian. However, in Slovenian news, in the top 3 there are New York and Brussels, which are not Slovenian cities.

3.3 Test data set

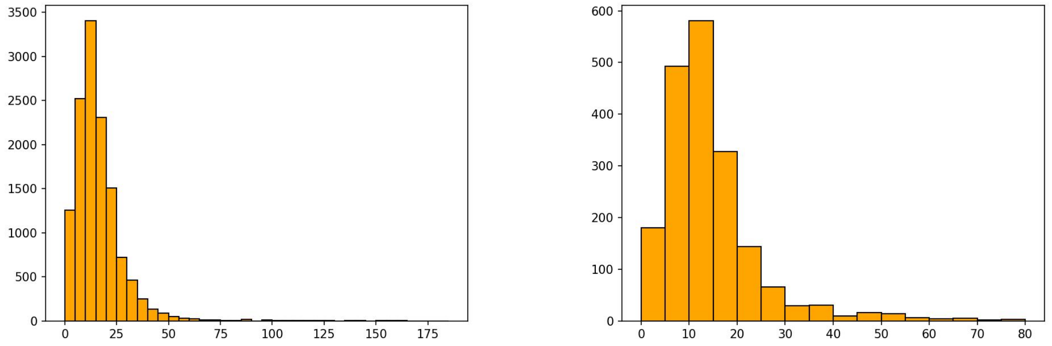
To create a test set, we decided to manually prepare labels for selected articles, which would serve as a ground truth for comparisons. We had 3 labelers taking part in assigning labels for the sentiment analysis task. For each article under review, 2 labelers have given their opinion. This ensured that at least 2 labelers were involved in assigning articles to a specific class – the possible classes: 1 - Positive, 0 - Neutral, -1 - Negative. We had 2 reviewers working on every single article to improve the certainty and quality of labels. The third labeler was involved only in case the labels of 2 previous labelers did not match. Each article has got assigned overall sentiment, sentiment towards keywords provided by journalists writing the article (available as metadata), and sentiment towards NERs extracted by us in the stage of preprocessing. This added extra work, as assigning single articles appropriate scores often meant that more than 15 labels had to be provided, which has made the process laborious. The test set was constructed of 12 articles randomly selected from the following 6 categories: Arts and Culture; Around Slovenia; Business, finance and economy; Health, environment, science; Politics; Sports. Three categories were excluded: Advisory; Roundup; Schedule of Events. Excluded categories comprise articles such as those making a function of daily digests. Assigning them a sentiment score does not make much sense, as a text in daily digests contains information from other articles summarized. This would make machine learning task difficult as sentiment from different events around the world would be blended in together. It is better to assign the sentiment to the original article directly instead of its summary that is preceded and followed by completely unrelated information concerning different topics. It was confirmed by the Slovenian Press Agency, that sentiment analysis of such articles would not bring much business value. Excluding 3 insignificant categories allowed for saving some time of the labelers (keeping in mind that assigning scores to each NER extracted could be very laborious).

This gave us a dataset of 72 articles, prepared to be used in evaluation in both tasks (document as well as aspect-based sentiment analysis). This means that for the task of document-based sentiment analysis, we had 72 labels assigned. The distribution was as follows: 36 neutral articles, 23 positive articles, and 13 negative articles. As in each article, multiple keywords and NERs can be found, 1192 labels to aspects have been assigned of which 996 were neutral, 148 were positive, and 48 were negative. Figure 7a shows the class distribution among articles' categories whereas Figure 7b shows the label distribution of classes among aspects mentioned the greatest number of times.



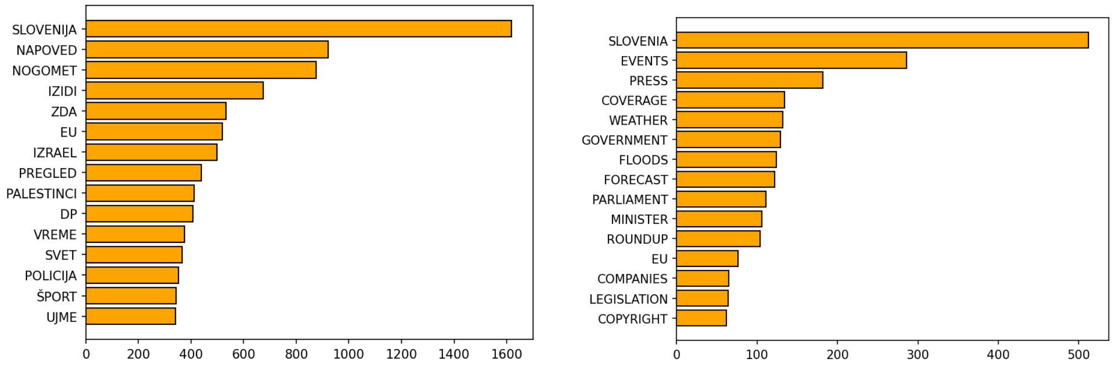
(a) **(b)**

Figure 2: Histogram of number of used words in each news for **(a)** Slovenian and **(b)** English. It was calculated based on white spaces in every news item from September to October 2023.



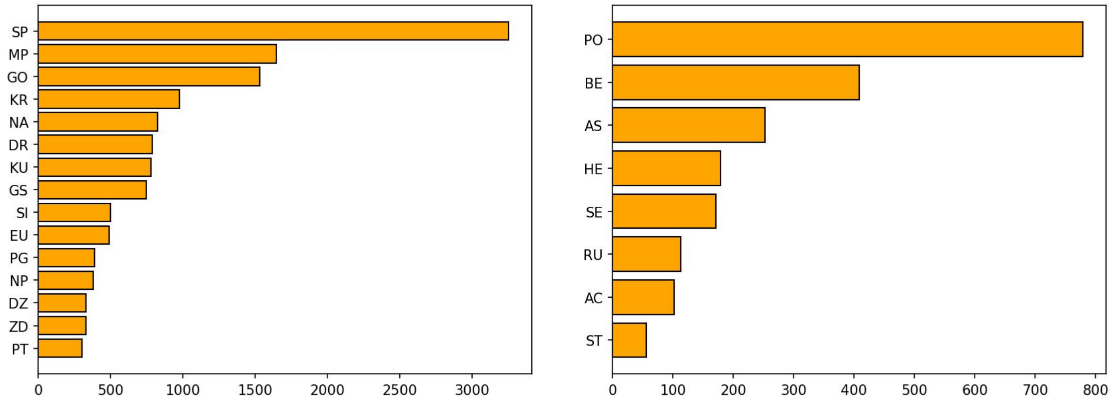
(a) **(b)**

Figure 3: Histogram of number of sentences in each news from September and October 2023 for **(a)** Slovenian and **(b)** English. It was calculated with the sent_tokenize function from the Python package nltk.tokenize.



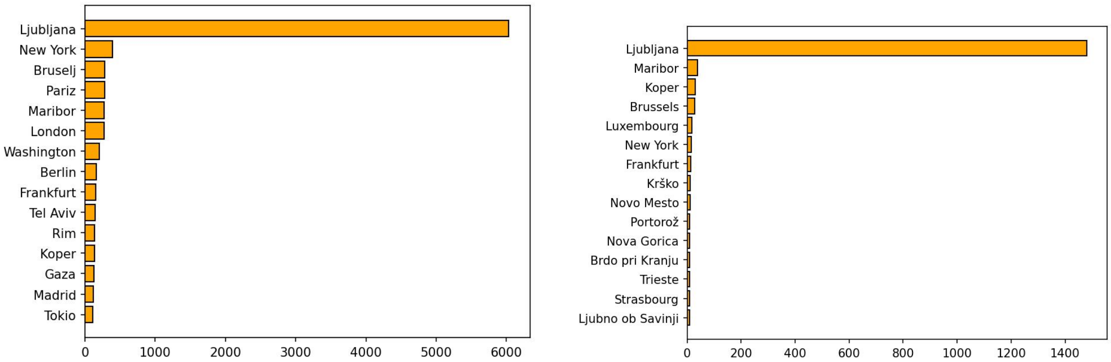
(a) **(b)**

Figure 4: Number of keyword occurrences in the news from September and October 2023 for **(a)** Slovenian and **(b)** English. Only the most common keywords are presented.



(a) (b)

Figure 5: Number of news in each category from **(a)** Slovenian, **(b)** English data from September and October 2023.



(a) (b)

Figure 6: Number of news items related to a particular city for **(a)** Slovenian, **(b)** English news items from September and October 2023. Only cities with the highest number of occurrences are presented.

We can observe that categories (AC) - Arts and Culture, as well as (AS) - Around Slovenia have no negative articles in the dataset. As the data comes from STA, it is no surprise that Slovenia is the most common keyword used, as it even has its category. Labels are mainly neutral in the test set in both cases document- and aspect-based task, with higher disparity in the case of aspects.

4 Experiments

The STA dataset (articles from the Slovenian Press Agency), which we use, has no annotated data. This means we have news available, but they do not contain information about their sentiment. We annotate the data for the test set manually to calculate the performance of the proposed final solution. Detailed information on the test set and its labeling process can be found in Section 3.3. However, regarding training, we use existing pretrained models and test how good they are without fine-tuning. For this reason, we worked only on English news, as we could annotate such data ourselves and better understand the results of predictions or XAI.

We have to face the problem of too long news. Typically, a transformer model will have a maximum input size of 512 tokens. We propose a solution to that problem. It focuses on splitting long news into parts, performing sentiment analysis on these parts, and combining them to get one output.

As per the project description, we aimed to satisfy the requirement of creating a solution capable of providing sentiment analysis of whole articles (document-based) and the different issues mentioned within them (aspect-based). We tested four different models for the document-based task: FinancialBERT, Auditor Review Sentiment Model, Financial-RoBERTa, and Twitter-RoBERTa. These models categorize into sentiments: negative, neutral, and positive. For the aspect-based task, we tested FLAN-T5 and DeBERTa. These models also return negative, neutral, or positive sentiment for each aspect indicated. Based on the tests (see Results), we selected one model for the document-based task and one for the aspect-based task, and using them, we prepared a sentiment visualization tool (using notebooks).

To perform aspect-based sentiment analysis, we first need to identify the relevant aspects of the article text. We use keywords of the article and the outputs of named entity recognition (NER) models for this task. We utilize *Babelscape/wikineural-multilingual-ner* model as it gave the most promising preliminary results, mainly as it allows for grouped entities improving NER extraction significantly. (Grouped entities are entities with names composed of multiple words). This model returns entities about which news sentiment research is most important and interesting - politicians, organizations, places, etc. Secondly, it was multilingual and trained on large corpora. Even though Slovenian was not among the languages it was trained on, having other than English languages helped. Even though articles are in English, some Slovenian entities are often mentioned. We believe that training on multiple languages still allowed for more accurate assessment when a non-English entity was mentioned. (This model was the most sensitive, returning, on average, the most entities, whereas others tended to omit some of the entities. + returned names were actually entities.)

4.1 Metrics

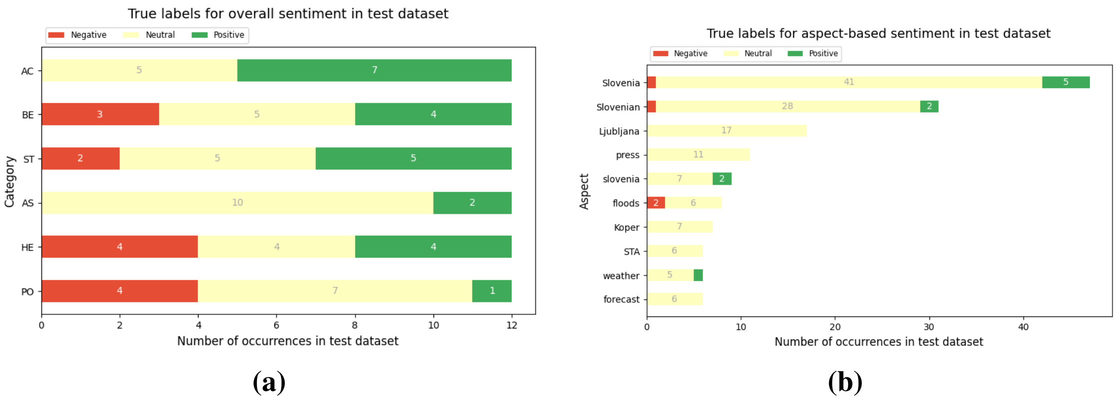
For both document-based and aspect-based analysis sentiment, we measure metrics such as accuracy and F1-score for negative, neutral, and positive classes, and we calculate F1-macro, an average of sub-individual F1-scores. A different model may be best for each metric, so we pay the most attention to accuracy when choosing the final model. Despite the prevalence of neutral sentiment and the lack of class balance - we were most concerned with the correct classification of precisely the most common class - neutral. We do not want to give an incorrect prediction suggesting that the neutral article is biased.

4.2 XAI analysis

An eXplainable Artificial Intelligence (XAI) analysis of the SiEBERT model uses the Captum library (Kokhlikyan et al., 2020). The attributions are calculated for whole articles. We prepare Integrated Gradients and LIME explanations.

In the case of NLP models, the implemented algorithms need to be customized. Integrated Gradients need to be computed in the embedding space. After generating explanations, the results must be summed to obtain token-level attributions.

We must define a perturbing and similarity function to calculate the LIME attributions. Perturbed input tokens are generated by sampling from the Bernoulli distribution with a probability



(a) **(b)**

Figure 7: Test set class distribution among articles' categories shown in figure **(a)**. Test set class distribution among most commonly used keywords and NERs shown in **(b)**.

of 0.5. We used an exponential cosine similarity function between two embeddings x_1 and x_2 defined as in Equation 1.

$$similarity(x_1, x_2) = \frac{x_1 x_2}{\max(||x_1||_2 ||x_2||_2, \epsilon)}$$
$$dist(x_1, x_2) = \exp\left(\frac{-(1 - similarity(x_1, x_2))^2}{2}\right) \tag{1}$$

A linear model with LASSO regularization is used as an interpretable model, so the amount of l_1 regularization can be controlled.

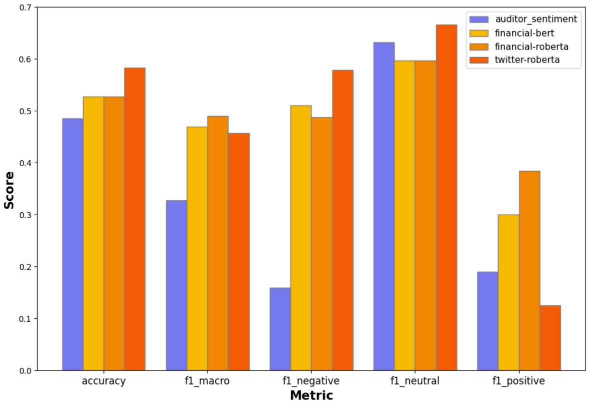


Figure 8: Document-based models' scores

5 Results and discussion

We present tests results of document-based analysis sentiment in Figure 8. Depending on the metric, the Financial-RoBERTa and Twitter-RoBERTa models turned out to be the best. However, we decided to choose the final model based on accuracy, and here, Twitter-RoBERTa was the best. Its accuracy was 0.58 and F1-macro 0.46. In Figure 10, we present more detailed results in confusion matrices. The model we chose performs the worst in detecting positive sentiment (it very often marks it as neutral). Nevertheless, since we do not want to imply bias, it is better to mistake it this way than to predict negative/positive sentiment when it is neutral.

In Figure 9, we show results of testing aspect-based models. In all selected metrics, the better model was FLAN-T5. Thus, there was no doubt to choose this model as the final solution. Its accuracy was 0.81, and F1-macro 0.47. More detailed results are presented in Figure 11 in the form of confusion matrices. We can observe one weak point of the selected model, which is that it predicted 108 positive sentiments as neutral. However, it is still better than DeBERTa.

6 Processing pipeline results

We prepare an end-to-end processing pipeline. Figure 12 presents the proposed solution architecture. This section describes the last processing layer: visualizations and XAI.

6.1 Sentiment visualizations

Our goal was to supervise sentiment in articles. To this aim, we created Python functions to visualize the statistics conveniently with the goal of ease of use. We show the operation of these functions based on English articles from 2 months - from September 1 to October 31, 2023 (a total of 1,912 articles).

Our aspect-based sentiment visualizations are based on divisions due to keywords and entities.

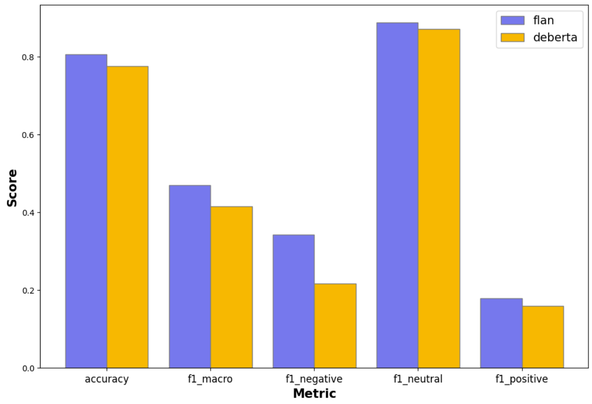


Figure 9: Aspect-based models' scores

Keywords are available in article metadata, annotated by STA, and denote what the news is about. Entities are additionally detected by the NER model so that sentiment of more aspects can be studied. Document-based sentiment visualizations use a breakdown by time of creation or categories that generally describe what the news is about (categories are less detailed than keywords; there are only 9 of them for English articles).

Firstly, we can plot aspect-based sentiment by keywords available in STA API. The examples of such plots are in Figure 13. We select the top ten keywords according to the number of appearances in all chosen articles, for instance, from a given period. With the created function, users can specify whether to use the number of articles or percentages for every keyword. We present both options – numbers for accurate figures and percentages for easy bar comparison. We can observe in Figure 13 that topics such as press, government, or minister are negative much more often than other keywords.

Keywords also can be chosen based on how often they are negative or positive. An example of such plots is in 14. Top keywords are selected based on the percentage of the desired sentiment among all articles with that keyword. Users can choose the minimum number of occurrences of a given keyword among articles to be considered. In this example, keywords that have occurred less than five times are not considered. We can see that the negative ones are usually scandals, dismissals, acts of violence, or wages. Positive are films, cycling, food, or awards.

Figure 15 shows aspect-based sentiment by entity found by the NER model. These differ from keywords which were provided in news metadata. The graphs are for the top ten entities. This way, we can see the sentiment of other aspects rather than rely on keywords alone. Slovenia, its capital, the EU, or Slovenia's prime minister, Robert Golob, appear frequently. We also see that the Prime Minister is often described negatively, but the neutral sentiment still prevails.

Another possible chart is the overall sentiment of articles by category or author. Figure 16 shows an example by category. With the presented example, one can observe which categories are more negative (Politics, Around Slovenia) and which are more positive (Arts and Culture, Sports).

Our solution also provides sentiment monitoring over time. Figure 17 shows examples of visualizations with 4-day intervals. Users can specify if they want to see the number of articles or percentages of negatives, neutrals and positives.

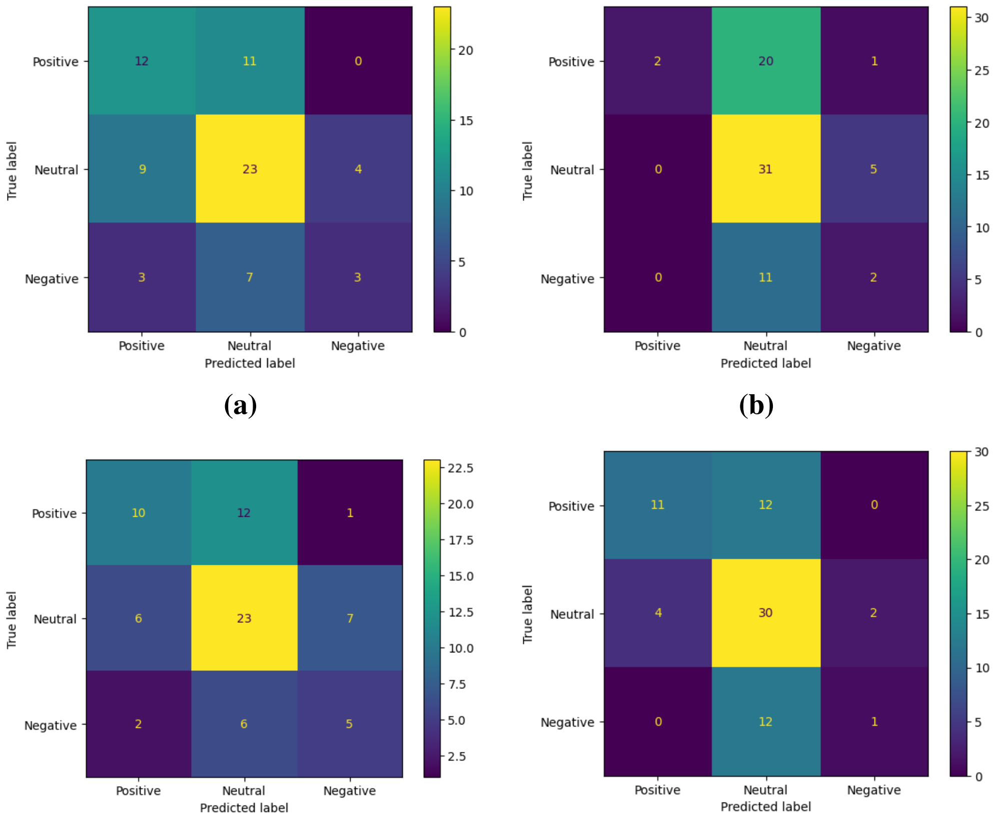
6.2 eXplainable Artificial Intelligence diagnostic

We see two possible benefits from incorporating XAI methods in our solution. Firstly, they can be used to explore the model's predictions and give insights into the importance of specific words. It will help journalists prepare unbiased texts.

The second aspect of XAI methods in the project is a better understanding of erroneous predictions. We observed that sentiment analysis of press articles is a difficult machine-learning task. The texts mostly consist of objective words with neutral sentiment; the ground-truth label strongly depends on the underlying mood of the whole article, which the models capture with difficulties.

Visualizations of Integrated Gradients attributions for a short demonstrative text and an STA article, for which the prediction was wrong, are presented in Figure 18. The true label was positive, however, the model predicted the text to be neutral. We may observe that the potential source of this error is attributing negative sentiment to words such as ”cannot” or ”translated” by the model, which in this context is unjustified.

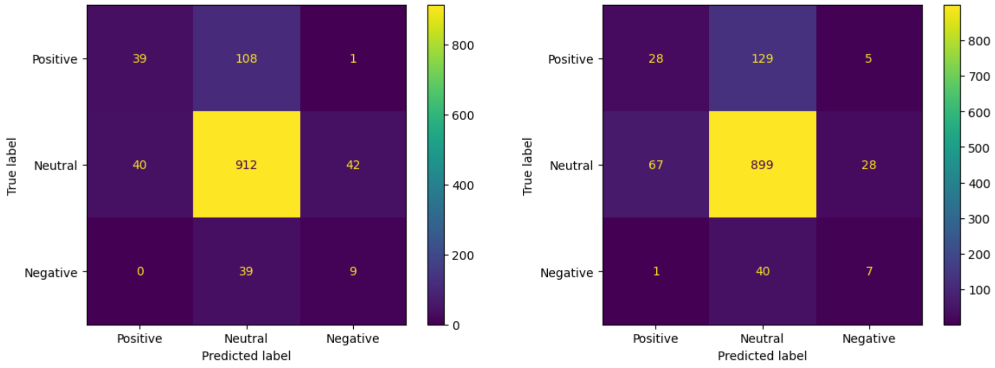
We discovered a limitation in using LIME explanations for long articles. The method works for shorter texts but does not give satisfactory results for longer texts. The problem was not solved by setting different l_1 regularization parameters; attributions of all tokens had still small magnitudes. For the visualizations, the attributions need to be multiplied. The results of the LIME explanations



(a) **(b)**

(c) **(d)**

Figure 10: Confusion matrices in document-based analysis task using **(a)** FinBERT, **(b)** Auditor Review Sentiment Model, **(c)** Financial-RoBERTa, **(d)** Twitter-RoBERTa.



(a) **(b)**

Figure 11: Confusion matrices in aspect-based analysis task using **(a)** FLAN-T5, **(b)** DeBERTa.

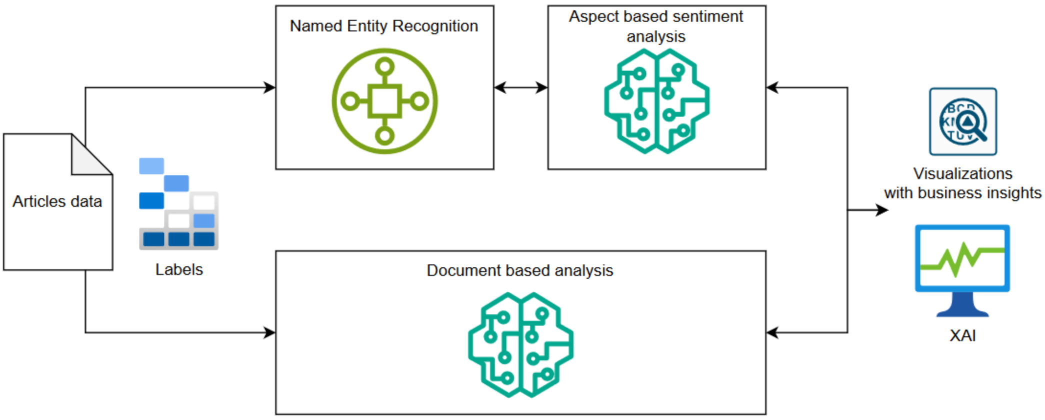
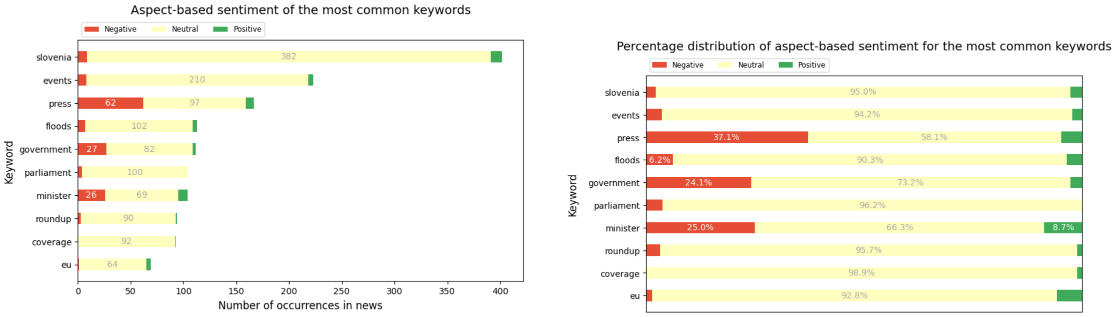


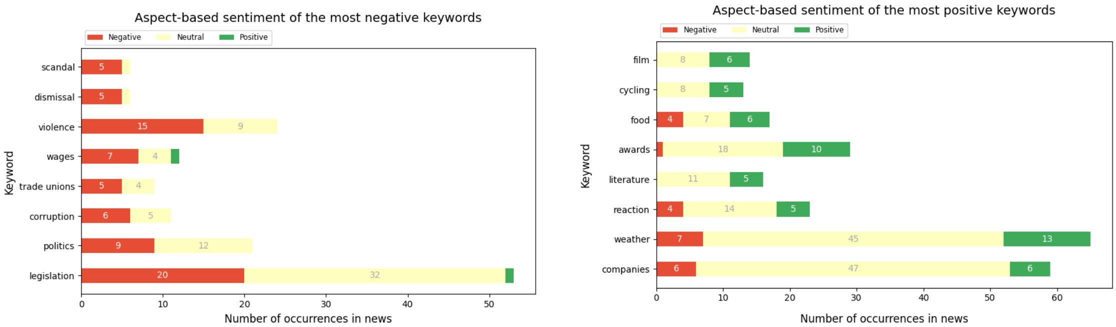
Figure 12: Architecture of implemented solution. Labels for the test dataset were manually created. Each article is processed in two ways: a) overall article sentiment is evaluated, b) named entities are recognized and sentiment toward them is calculated. Finally, analysis results are presented and XAI is performed.



(a)

(b)

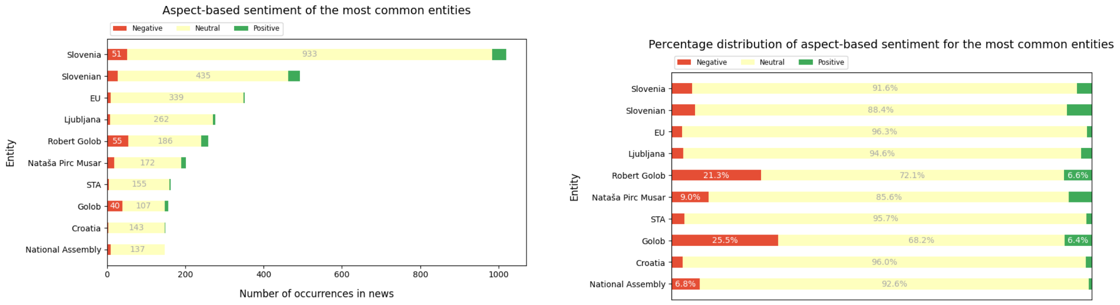
Figure 13: Aspect-based sentiment by keywords available in STA API presented with **(a)** number of articles for each sentiment and keyword, **(b)** percentage of every sentiment for a given keyword. Data is from September 1 to October 31, 2023.



(a)

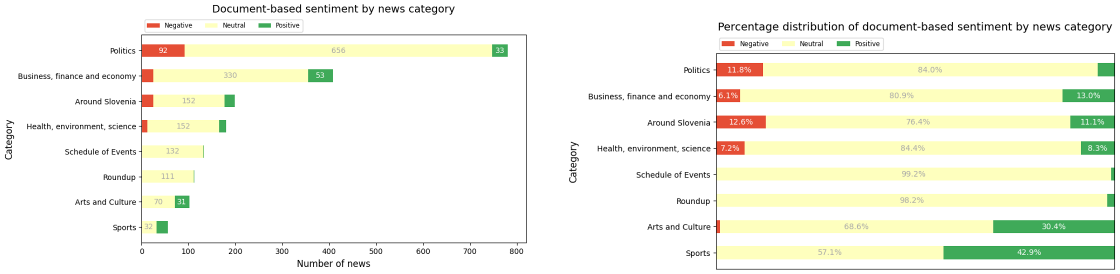
(b)

Figure 14: Most **(a)** negative and most **(b)** positive keywords. Data is from September 1 to October 31, 2023.



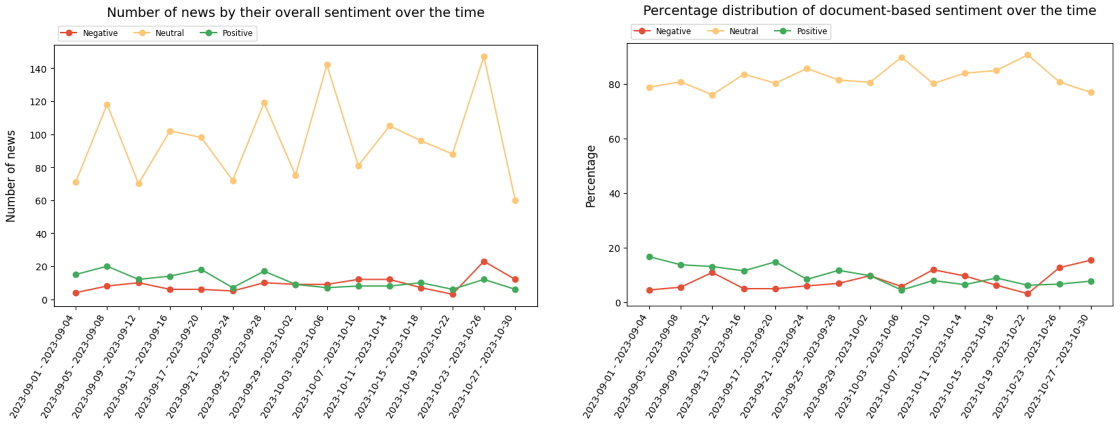
(a) **(b)**

Figure 15: Aspect-based sentiment by found entities presented with **(a)** number of articles for each sentiment and entity, **(b)** percentage of every sentiment for a given entity. Data is from September 1 to October 31, 2023.



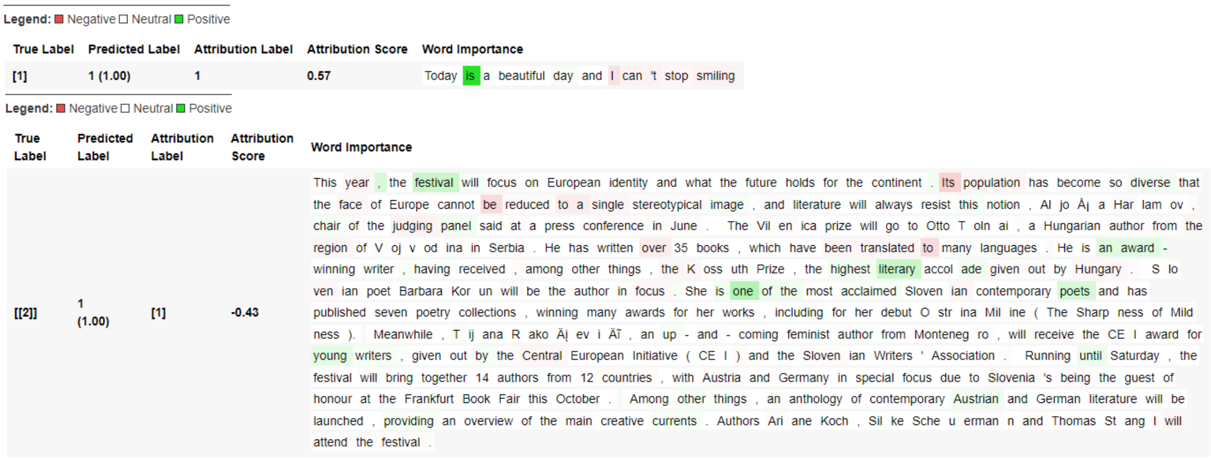
(a) **(b)**

Figure 16: Overall articles sentiment grouped by category presented with **(a)** number of articles for each sentiment and category, **(b)** percentage of every sentiment for a given category. Data is from September 1 to October 31, 2023.



(a) **(b)**

Figure 17: Overall articles sentiment according to time presented with **(a)** number of articles for each sentiment and time period, **(b)** percentage of every sentiment for a given time period. Data is from September 1 to October 31, 2023.



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
[1]	1 (1.00)	1	0.57	Today is a beautiful day and I can 't stop smiling

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
[[2]]	1 (1.00)	[1]	-0.43	This year , the festival will focus on European identity and what the future holds for the continent . Its population has become so diverse that the face of Europe cannot be reduced to a single stereotypical image , and literature will always resist this notion , Al jo Åj a Har lam ov , chair of the judging panel said at a press conference in June . The Vil en ica prize will go to Otto T oln ai , a Hungarian author from the region of V oj v od ina in Serbia . He has written over 35 books , which have been translated to many languages . He is an award - winning writer , having received , among other things , the K oss uth Prize , the highest literary accol ade given out by Hungary . S lo ven ian poet Barbara Kor un will be the author in focus . She is one of the most acclaimed Sloven ian contemporary poets and has published seven poetry collections , winning many awards for her works , including for her debut O str ina Mil ine (The Sharp ness of Mild ness). Meanwhile , T ij ana R ako Äj ev i Äĩ , an up - and - coming feminist author from Monteneg ro , will receive the CE I award for young writers , given out by the Central European Initiative (CE I) and the Sloven ian Writers ' Association . Running until Saturday , the festival will bring together 14 authors from 12 countries , with Austria and Germany in special focus due to Slovenia 's being the guest of honour at the Frankfurt Book Fair this October . Among other things , an anthology of contemporary Austrian and German literature will be launched , providing an overview of the main creative currents . Authors Ari ane Koch , Sil ke Sche u erman n and Thomas St ang l will attend the festival .

Figure 18: Integrated Gradients attributions. Short demonstrative text is presented in the upper part, and the full-length article is in the lower part. The sentiment of the article was predicted incorrectly. XAI methods may help to diagnose and understand the source of the error.

of demonstrative text and an STA article are presented in Figure 19.

7 Summary

In this work we propose an end-to-end solution for the sentiment analysis of articles. We propose to investigate the texts using both aspect- and document-based sentiment analysis as we found this approach to be the best suited for the business needs stated by the Slovenian Press Agency. To automate the aspect-based sentiment analysis for historical articles we leverage the advancements in NER for natural language processing. However, we also perform an analysis of keywords, to give journalists the possibility to state explicitly the issues towards which sentiment should be measured. They can start providing it in the metadata of articles being newly created. We have evaluated several models and chosen Twitter-RoBERTa and FLAN-T5 to be the best suited for document and abstract-based analysis respectively. We have implemented 2 XAI methods to provide explanations of the models. In this case, XAI can help firstly to understand the errors made by the model, secondly to help journalists identify and avoid emotionally charged phrases, as a step to create free of bias, reliable content. We also prepared notebooks and scripts to create visualizations that aggregate model predictions and present the results in a way that allows more convenient, business-insightful analysis of the results. These notebooks were used to analyze 1912 STA's articles from September 1 to October 31, 2023. The code we have used is made public with README.md file explaining how to leverage this solution to enable STA to reproduce our results, incorporate it, and investigate new data or develop it further. Lastly, as a result of this work, we provide a labeled test set comprising 72 articles.

8 Future work

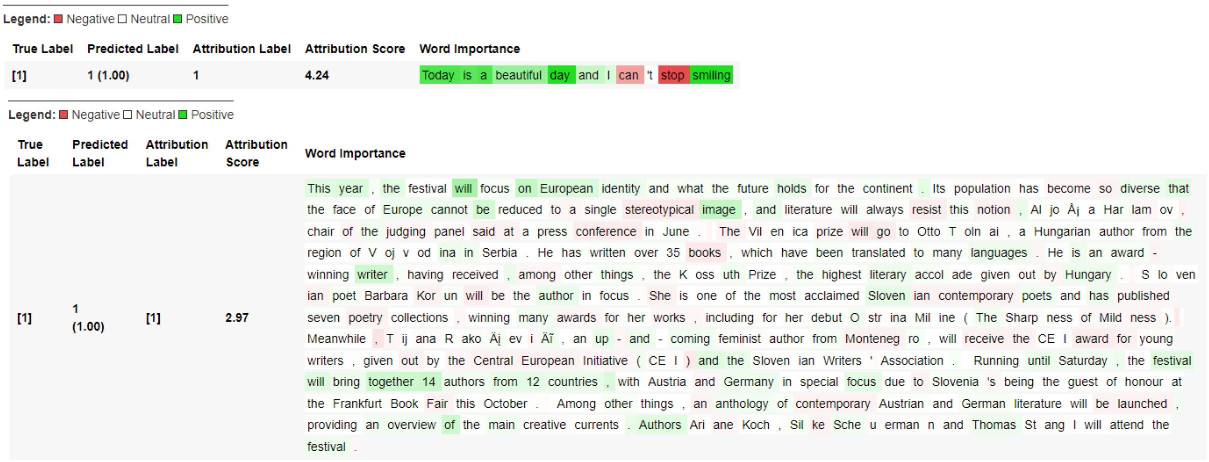
As we have described in the literature review, models trained on one domain may transfer poorly to another domain (Liu et al., 2020). As there was no labeled data available, we had no possibility of fine-tuning publicly available models to our task. This means, that once the models are fine-tuned, improvement in the results could be observed. It is worth considering having a separate model for each category of news to further boost the performance (i.e. one per Sports, one per Arts and Culture, etc.).

As the most articles published by STA, the solution could be extended to Slovenian language by testing models dedicated for this language or by experimenting with multilingual solutions.

Another direction to further develop this solution would be coupling already implemented sentiment analysis task with ML-based emotion detection which allows to analyze more complex emotions like fear, anger, sadness, love, frustration, and many more.

Contribution

Our individual contributions are presented in Table 1.



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
[1]	1 (1.00)	1	4.24	Today is a beautiful day and I can 't stop smiling

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
[1]	1 (1.00)	[1]	2.97	This year , the festival will focus on European identity and what the future holds for the continent . Its population has become so diverse that the face of Europe cannot be reduced to a single stereotypical image , and literature will always resist this notion , Al jo Åj a Har lam ov , chair of the judging panel said at a press conference in June . The Vil en ica prize will go to Otto T oln ai , a Hungarian author from the region of V oj v od ina in Serbia . He has written over 35 books , which have been translated to many languages . He is an award - winning writer , having received , among other things , the K oss uth Prize , the highest literary accol ade given out by Hungary . S lo ven ian poet Barbara Kor un will be the author in focus . She is one of the most acclaimed Sloven ian contemporary poets and has published seven poetry collections , winning many awards for her works , including for her debut O str ina Mil ine (The Sharp ness of Mild ness). Meanwhile , T ij ana R ako Äj ev i ÄÏ , an up - and - coming feminist author from Monteneg ro , will receive the CE I award for young writers , given out by the Central European Initiative (CE I) and the Sloven ian Writers ' Association . Running until Saturday , the festival will bring together 14 authors from 12 countries , with Austria and Germany in special focus due to Slovenia 's being the guest of honour at the Frankfurt Book Fair this October . Among other things , an anthology of contemporary Austrian and German literature will be launched , providing an overview of the main creative currents . Authors Ari ane Koch , Sil ke Sche u erman n and Thomas St ang l will attend the festival .

Figure 19: LIME attributions. Short demonstrative text is presented in the upper part, and the full-length article is in the lower part. The LIME attributions for the long text of an article have smaller magnitudes, so they need to be multiplied by 10 for visualizations.

Team member	Tasks
Jakub Koziel	SOTA in sentiment analysis task (5h), ABSA implementation (8h), NER implementation (6h), data preprocessing and test set preparation (4h), evaluation of different models (2h), metric calculation (3h), data labeling (3h), solution concept (3h), reviewing other teams (2h), incorporating reviewers remarks from rebuttals (3h), writing the report (8h)
Jakub Lis	Data acquisition (4h), Data description (4h), exploratory data analysis (6h), data labeling (6h), visualizations with model predictions (6h), solution concept (3h), reviewing other teams (2h), incorporating reviewers remarks from rebuttals (4h), editing milestone presentations (4h), writing the report (8h)
Bartosz Sawicki	Explainable artificial intelligence research (4h), XAI implementation (6h), implementation of document based sentiment analysis(8h), evaluation of different models (2h), data labeling (6h), solution concept (2h), reproducibility check list (3h), reviewing other teams (2h), incorporating reviewers remarks from rebuttals (3h), code refactoring (2h), writing the report (8h)

Table 1: Contributions of team members.

Acknowledgments

We thank the Slovenian Press Agency (STA) for sharing access to its API.

References

[Birjali et al.2021] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane 2021. *A comprehensive survey on sentiment analysis: Approaches, challenges and trends*, Knowledge-Based Systems.

[Wankhade et al.2022] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, Chaitanya Kulkarni 2022. *A survey on sentiment analysis methods, applications, and challenges*, Artificial Intelligence Review.

[Zhang et al.2023] Wenxuan Zhang and Xin Li and Yang Deng and Lidong Bing and Wai Lam 2023. *A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges*, IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 11, pp. 11019-11038.

[Liu et al.2020] Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah 2020. *A survey on sentiment analysis methods, applications, and challenges*, Aspect-Based Sentiment Analysis: A Survey of Deep Learning Methods," in IEEE Transactions on Computational Social Systems, vol. 7, no. 6, pp. 1358-1375.

[Bird et al.2009] Steven Bird, Edward Loper and Ewan Klein 2009. *A survey on sentiment analysis methods, applications, and challenges*, Natural Language Processing with Python. O'Reilly Media Inc.

[Demszky et al.2020] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, Sujith Ravi 2020. *GoEmotions: A Dataset of Fine-Grained Emotions*. ArXive preprint.

[Žitnik 2019] Žitnik, Slavko 2019. *Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0*. Slovenian language resource repository CLARIN.SI.

[Bučar et al.2018] Bučar, J., Žnidaršič, M., Povh, J. 2018. *Annotated news corpora and a lexicon for sentiment analysis in Slovene*. Lang Resources & Evaluation, 52, 895–919 (2018).

[Kim et al.2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres 2018. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. Proceedings of the 35th International Conference on Machine Learning.

[Ribeiro et al.2016] Ribeiro Marco Tulio, Sameer Singh, Carlos Guestrin 2016. *" Why should i trust you?" Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

[Sundararajan et al.2017] Sundararajan, Mukund, Ankur Taly, Qiqi Yan. 2017. *Axiomatic attribution for deep networks*. International conference on machine learning.

[Baniecki et al.2021] Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, Przemyslaw Biecek. 2021. *dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python*. Journal of Machine Learning Research.

[Kokhlikyan et al.2020] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, Orion Reblitz-Richardson. 2020. *Captum: A unified and generic model interpretability library for PyTorch*. ArXive preprint.

[Hartman et al.2023] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. *More than a Feeling: Accuracy and Application of Sentiment Analysis*. International Journal of Research in Marketing.

[Yang and Li 2022] Heng Yang and Ke Li. 2022. *Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning*. International Journal of Research in Marketing.

[He et al.2021] Pengcheng He and Jianfeng Gao and Weizhu Chen. 2021. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*.

[Sanh et al.2019] Victor Sanh and Lysandre Debut and Julien Chaumond and Thomas Wolf. 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. International Journal of Research in Marketing.

[Augustyniak et al.2023] Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, Tomasz Kajdanowicz. 2023. *Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark*. arXiv.

[Mozetič et al.2016] Mozetič, Igor and Grčar, Miha and Smailović, Jasmina. 2016. *Multilingual Twitter Sentiment Classification: The Role of Human Annotators*. PLOS ONE 11. 5(2016): 1-26.

[Hazourli 2022] Ahmed Rachid Hazourli. 2022. *FinancialBERT for Sentiment Analysis*. `https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis`.

[Finance Inc.2022] Rajiv Shah and Derek Thomas. 2022. *Auditor Review Sentiment Model*. `https://huggingface.co/FinanceInc/auditor_sentiment_finetuned`.

[Loureiro et al.2022] Loureiro, Daniel and Barbieri, Francesco and Neves, Leonardo and Espinosa Anke, Luis and Camacho-collados, Jose. 2022. *Twitter-roBERTa-base for Sentiment Analysis*. `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest`.

[Romero 2023] Manuel Romero 2023. *DistilRoberta-financial-sentiment*. `https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis`.

[Chung et al.2022] Chung, Hyung Won and Hou, Le and Longpre, Shayne and Zoph, Barret and Tay, Yi and Fedus, William and Li, Eric and Wang, Xuezhi and Dehghani, Mostafa and Brahma, Siddhartha and Webson, Albert and Gu, Shixiang Shane and Dai, Zhuyun and Suzgun, Mirac and Chen, Xinyun and Chowdhery, Aakanksha and Narang, Sharan and Mishra, Gaurav and Yu, Adams and Zhao, Vincent and Huang, Yanping and Dai, Andrew and Yu, Hongkun and Petrov, Slav and Chi, Ed H. and Dean, Jeff and Devlin, Jacob and Roberts, Adam and Zhou, Denny and Le, Quoc V. and Wei, Jason 2022. *Scaling Instruction-Finetuned Language Models*. Creative Commons Attribution 4.0 International. `https://huggingface.co/google/flan-t5-base`.

Rebuttal

Many thanks to our reviewers for providing us with valuable feedback and helping us, to achieve in this project something remarkable! The reviews were very insightful and we are also glad, that our efforts did not go unnoticed. We were happy to read all the strong points of our work highlighted. Below we refer only to the weak points mentioned or suggestions, providing our response which sometimes is also the outcome of a discussion with the team that was reviewing us.

Review rebuttal - Team 1

Weak point/Suggestion	Our response
Bigger test set could be built, to achive more reliable results (but you mentioned that it is not easy due to the time schedule)	We have extended the test set from 24 to 72 articles since the first version of the final report.
It would be interesting to see how the SiEBERT outperforms the other models for sentiment analysis task (e.g. DistilBERT). For example a comparative visualization/table could be presented proving that your model choice was actually very good.	Since the first version of the report we have extended our experiments and tested more different models for each task. We have eliminated SiEBERT, as this model was not designed to provide netural sentiment (only positive and negative). Instead, we have tested FinBERT, Auditor Review Sentiment Model (a fine-tuned version of FinBERT), Financial-RoBERTa, and Twitter-RoBERTa. For abstract-based sentiment analysis we have tested FLAN-T5 to compare it with previously tested DeBERTa. So since this valuable remark from reviewers, 5 more models were tested during the last weeks of the project.
In Table 1 you show the confusion matrix for ABSA. It would be interesting to see the data instances which were mismatched completely (e.g. instead of positive label, negative one was assigned). Maybe there could be discovered limitation of the model where it could be possibly improved.	We provide such an example in Figure 18.
The work is limited to only one solution, without comparing the practical results to other solutions from literature.	As we have explained in the Related works, the results of sentiment analysis are hardly comparable between datasets. Instead of comparing our results with the work of others, we compare different models on our test set. In terms of comparing to a work that would also try to tackle the problem of article sentiment analysis, we have not found anything recent to compare to.
The test set is limited to english language only. Therefore its hard to tell how good would be the performance for the whole STA (Slovenian Press Agency) dataset.	This limitation of the work is true. In future works, we propose extending it to the Slovenian language as we did not manage to do that given the time constraints.

Review rebuttal - Team 6

Weak point/Suggestion	Our response
there is no clear hypothesis stated nor research questions (it should be emphasized) hypothesis is not clearly stated	We have extended the introduction Section 1 with 2 additional subsections: Project goal and project overview.
The dataset could be described in more detail, what the STA database is etc.	We have extended the description of the database in Section 3.
Plots are unreadable, axes not signed.	We increased the font size of the labels and legends to make them more visible. We added titles and axis labels to the solution visualizations, and in the case of EDA, we enhanced the figures' captions to accurately describe what we are showing.