YleSent – Sentiment Analysis of Yle News

Nora Heikkilä, Saija Piira, Mattis Reinola

Introduction to Data Science –course Mini Project

YleSent is a project that analyzes the sentiment of Yle news articles and shows how it changes over time. It uses a **logistic regression model** to predict whether news content is positive or negative and then visualizes average sentiment by day and month. The idea for this project came from an observation that news can often seem too negative, which may affect how people feel and think about current events. By showing clear trends, YleSent helps users see whether news coverage is mostly negative or balanced. This makes it easier for readers to view the news more critically and understand how the tone of reporting might shape public perception.

Methods

The original training data for this project came from the ScandiSent dataset, which is a multilingual sentiment analysis dataset with labeled reviews from Trustpilot. The reviews are annotated with sentiment labels (positive or negative). For the test dataset we used Yle news articles. Later we realized through model accuracy that the ScandiSent data was not compatible as training data while testing the model on Yle news articles.

We came across a model adaptation issue, since the training data differed from real Yle news in style and content. While Yle news articles are longer and written more formally, positive or negative tone is less obvious than in the short opinion-based ScandiSent reviews. To better align the model with the news domain, we first re-trained it using a dataset of 1,000 Yle news articles that were labeled via supervised annotation with ChatGPT. Since the model's accuracy was still not satisfactory, we proceeded to label 3,000 Yle news articles via supervised annotation with ChatGPT and utilized this dataset for training in place of the ScandiSent data. This improved prediction accuracy and made visualisations more meaningful. This re-training helped the model better adapt to the news domain and recognize more subtle sentiment cues efficiently.

Data Preprocessing

Before training, all Yle news articles were cleaned and lemmatized to prepare them for sentiment analysis. The raw CSV files containing article text and sentiment ratings were loaded using **pandas**. Text data was processed using the **Stanza** natural language processing library for Finnish, which provided tokenization, lemmatization, and morphological analysis.

Additionally, **NLTK** was used to remove Finnish stop words (common words such as “ja” or “on” that do not carry sentiment meaning).

Each article was converted to lowercase, stripped of punctuation and extra spaces, and lemmatized so that different word forms were reduced to a common base form (e.g., “kirjoitti” → “kirjoittaa”). The cleaned sentences were stored as lists, then saved to new CSV files for use in model training.

For the training data, after the classification, we lemmatized all the sentences of articles and also split the articles by sentences, assigning each sentence the sentiment of the whole article, in order to obtain more data to train our model on.

For the remaining data — approximately 800,000 articles — we only used the headlines, as lemmatizing the full text would have been extremely resource-intensive.

Model Training

The sentiment classification model was built using **scikit-learn**’s **Pipeline**, combining a **TF-IDF vectorizer** and a **Logistic Regression** classifier.

The final training dataset used in training consisted of 2702 Yle news articles that were preprocessed and labeled. The class distribution was relatively balanced with 57 % news labeled as positive and 43 % as negative. For each article, the text and its corresponding sentiment rating (positive or negative) were extracted. The dataset was split into training (80%) and testing (20%) subsets. The news articles used for training were selected randomly and then removed from the entire dataset.

The **TF-IDF vectorizer** transformed text into numerical features representing how important each word is across the dataset, while **Logistic Regression** was used to predict sentiment labels. After training, the final model was saved as model_logreg.joblib for later use.

Training performance was measured using **accuracy score** to ensure the model generalizes well to unseen data.

The best and latest model we ran, got the accuracy of 0.755 which was also relatively accurate according to prediction of what the method is capable of.

Sentiment Prediction

For new, unseen articles, the trained model was loaded and used to predict sentiment labels. Each article was processed almost in the same way as the training data, converted from string format into lists of sentences and the first sentence was lemmatized text. The model then predicted sentiment (positive or negative) for that first sentence for each article.

Model Evaluation

The trained model was evaluated on a separate test dataset using **accuracy** as the performance metric. This provided a basic measure of the model's performance across both positive and negative classes.

The **classification report** was generated to show per-class performance, while the **probability scores** from the logistic regression output helped assess the model's confidence in its predictions. Additionally, a **baseline accuracy** was calculated to compare the model against a simple majority-class prediction.

The evaluation results confirmed that the re-trained model improved accuracy and better captured the subtler sentiment tones typical of Finnish news articles. But using only the news data we classified with ChatGPT's help, worked a lot better than anything we tried with the tripadvisor data. While the only tripadvisor reviews created a slightly better result than random guessing and sometimes even worse, the amount of actual news data added to the model started shifting it to better and better results which is why it finally took only the classified news.

Visualization

The visualization module of **YleSent** transforms the model's predictions into a webpage with interactive and easy to understand data visualizations. It combines multiple techniques to illustrate sentiment trends and keyword importance.

First there is a **network graph** created using **PyVis**, connecting related news subjects based on their co-occurrence and sentiment. The daily and monthly average news sentiment scores are calculated from the prediction results and displayed as an interactive **time series graph** using **Plotly**. This allows users to explore how the tone of Yle news articles changes over time, identifying patterns such as consistently negative or improving sentiment periods.

In addition to the general sentiment trends, we also wanted to see if the analysis would reveal sentiment trends by different news topics. Yle's news database contained topic labels for each article, but it turned out that their use of these labels was quite inconsistent: there were many overlapping categories and the majority of the most frequent labels were related to sports. For this reason, ten common and interesting categories were picked for further investigation. The monthly average sentiment of these topics is shown in a time series graph, where the user can select what categories to view. Another graph shows the monthly number of articles in each of these categories, since the changes in volume can influence how smooth the sentiment curve appears.

Finally, two **word clouds** are generated to highlight the most influential keywords driving positive and negative classifications. These are derived from the logistic regression model's TF-IDF feature weights, showing which words most strongly contribute to each sentiment. Together, these visualizations provide a comprehensive view of sentiment distribution,

keyword influence, and topic relationships, helping users better understand how sentiment and themes evolve within Finnish news coverage.

Conclusion

The **YleSent** project successfully analyzed and visualized sentiment trends in Finnish news articles from Yle. Using a logistic regression model with TF-IDF features, the system was able to classify news sentiment and display meaningful trends over time. The visualizations—such as time series graphs, word clouds, and topic networks—helped make the results understandable and highlighted how certain topics and keywords influence overall tone.

Through model re-training and domain adaptation, prediction accuracy improved, showing that tailoring data to the specific style of news writing is essential for reliable sentiment analysis. The project demonstrates how computational methods can bring more transparency to media tone and support readers in evaluating news more critically.

Some visible patterns could also be linked to major real world events, although proving the correlation would require further analysis. The drop in sentiment at the beginning of Covid-19 pandemic around March 2020 is so clear that it is likely a reliable finding, whereas for example, Russian's annexation of Crimea in February 2014 and Donald Trump's first presidential term between January 2017 and January 2021 can be observed mainly in topic-specific trends and should be considered more speculative.

A **major limitation** was the high proportion of neutral articles in the dataset. This class imbalance caused the model to lean toward predicting neutral sentiment, reducing its ability to clearly distinguish positive and negative tones. Ideally, this issue could be mitigated through techniques such as oversampling of minority classes or filtering for articles with more emotionally charged language. Addressing this imbalance would likely improve the interpretability and reliability of future results.

Despite its challenges, the YleSent project proved to be a fascinating exploration of media tone and sentiment analysis. With more time and computational resources, it would be interesting to expand the project to include a larger labeled dataset, test more advanced language models (such as BERT or FinBERT), and explore multilingual sentiment across Nordic news sources. Additionally, examining the relationship between **sentiment and real-world news events** could provide deeper insights into how media tone reflects or shapes public perception. This could be an engaging direction for further study and analysis.