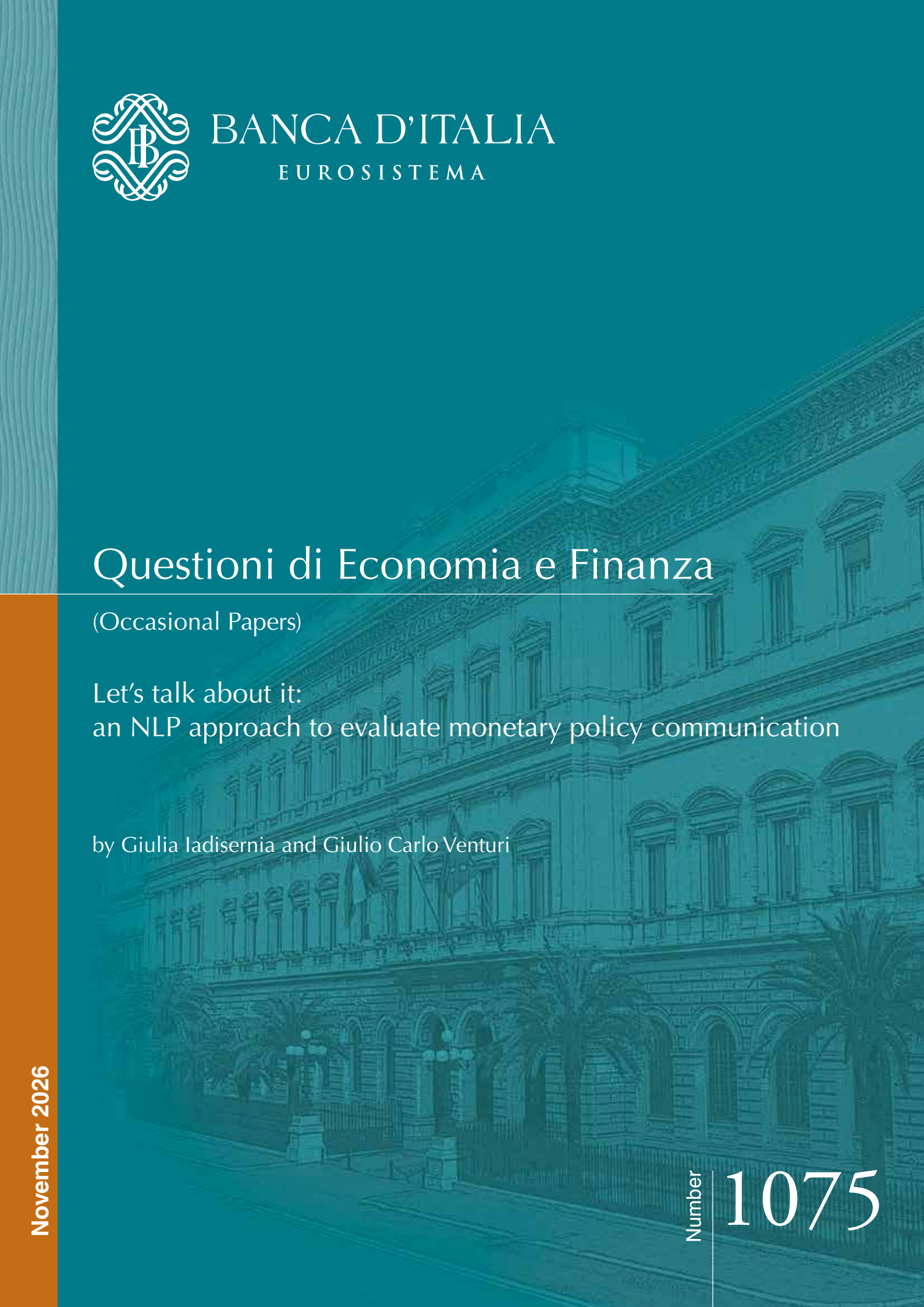
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

Let's talk about it: an NLP approach to evaluate monetary policy communication

by Giulia Iadisernia and Giulio Carlo Venturi

November 2026

Number 1075

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Let's talk about it:
an NLP approach to evaluate monetary policy communication

by Giulia Iadisernia and Giulio Carlo Venturi

Number 1075 – November 2026

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of Banca d'Italia and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within Banca d'Italia, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of Banca d'Italia

LET'S TALK ABOUT IT:
AN NLP APPROACH TO EVALUATE MONETARY POLICY COMMUNICATION

by Giulia Iadisernia* and Giulio Carlo Venturi**

Abstract

We develop an NLP pipeline to compute the textual tone score index of the ECB's monetary policy communication, quantifying the degree of hawkishness or dovishness. The pipeline leverages RoBERTa Transformer models, and is trained and evaluated on a sample of transcripts of ECB press conferences following policy rate announcements from 2000 to 2026. The index displays a coherent evolution of the ECB's monetary policy stance, as the series co-moves with ECB inflation projections, ECB policy rates and the ECB shadow rate. Moreover, the index allows us to retrieve an economically acceptable effect of monetary policy on inflation, thus proving that the index meaningfully summarizes the ECB's evaluation of the underlying economic environment. The pipeline is used to empirically estimate the ECB's loss function regarding inflation, validating the assumption of symmetry around the inflation target.

JEL Classification: C45, E52, E58.

Keywords: monetary policy communication, natural language processing, central bank loss function.

DOI: 10.32057/0.QEF.2026.1075

* Department of Computer, Control and Management Engineering, Sapienza University of Rome; iadisernia@diag.uniroma1.it

** Economic Outlook and Monetary Policy Directorate, Banca d'Italia; giuliocarlo.venturi@bancaditalia.it

1 Introduction[1]

"It is widely accepted that the ability of a central bank to affect the economy depends critically on its ability to influence market expectations about the future path of overnight interest rates" (Blinder et al., 2008, p. 913). Therefore, central bank communication regarding monetary policy is a tool of utmost importance to achieve the desired policy goals. However, empirical evaluation of the impact of communication is complex, because the comprehension of human language is inherently nuanced and subjective.

We propose a methodology to translate transcripts of monetary policy communication into a quantitative measure suitable for econometric analysis. Specifically, we develop a Natural Language Processing (NLP) pipeline that takes as input the raw transcript of a central bank official statement and produces as output a *textual tone score*, that is a numerical quantification summarizing the degree of hawkishness or dovishness conveyed in the communication.

Our pipeline builds on the RoBERTa model (Liu et al., 2019), a state-of-the-art Transformer architecture pre-trained on large text corpora to generate meaningful numerical representations of text. Intuitively, the model maps "human language" (i.e. words) into "machine language" (i.e. numbers) that can be used for downstream analytical tasks. Specifically, the pipeline consists of four steps. (I) It pre-processes the raw document and retains only economics-related sentences. (II) It identifies which of these sentences are specifically relevant to monetary policy. (III) It classifies the tone of each relevant sentence into five ordered categories: strongly dovish, dovish, neutral, hawkish, strongly hawkish. (IV) It aggregates sentence-level classifications into a single textual tone score for the document. Steps (II) and (III) rely on supervised learning. The authors independently annotated a corpus of sentences for monetary policy relevance and textual tone, and the Transformer models are trained on this labeled data to learn how to classify new sentences accordingly.

We apply this pipeline to a corpus of 270 ECB press conferences held after policy rate announcements between 2000 and 2026. Using these data, we construct a textual tone score index, i.e. a time series that tracks the evolution of communication tone over time. The index recovers a coherent evolution of the ECB policy stance, as the series exhibits a close correlation with ECB projections of the euro area Harmonized Index of Consumer Prices (HICP) inflation rate, the ECB policy rates and the ECB shadow rate computed by Krippner (2013).

We show that our index meaningfully summarizes the ECB's evaluation of the underlying economic environment, because the index allows us to recover an economically acceptable effect of monetary policy on inflation. Indeed, the index mitigates the endogeneity between policy actions and targets, as pointed out by the seminal work by Romer and Romer (2004), summarizing the central bank's evaluation of the economic environment.

Exploiting our pipeline, we recover a non-parametric empirical estimate of the ECB loss function regarding inflation, validating the theoretical assumption that the function is symmetric around the inflation target. This finding validates the canonical choice in monetary policy theory of a quadratic loss function, which implies that both upward and downward deviations from the inflation target are symmetrically costly for the central bank. Our result contests the main claim of Haavio et al. (2026), who find that the ECB loss function regarding inflation is asymmetric with greater aversion to inflation above the inflation target.

The remainder of the paper is organized as follows. Section 2 reviews previous attempts to

[1]The authors thank Alex Tagliabracci for providing ECB forecasts time series and colleagues at Banca d'Italia ART Unit for useful discussion.

evaluate monetary policy communication. Section 3 describes the data employed to implement and exploit our NLP pipeline. Section 4 explains in detail how the pipeline is assembled, trained, and executed. Section 5 exploits the textual tone score index to estimate the impact of monetary policy on inflation. Section 6 exploits the pipeline to recover a non-parametric empirical estimate of the ECB loss function regarding inflation.

2 Related Work

The study of monetary policy communication has several dimensions both in terms of goals and tools. Some studies evaluate the tone of communication, others analyze authorship style, still others assess the similarity of statements. Techniques employed range from traditional text analysis methods to more advanced natural language processing approaches. In this section, we focus on studies that are most closely related to our application, namely those aimed at quantifying the hawkish–dovish tone of central bank communication.

Among the simplest and most widely used approaches are *Bag-of-Words* (BoW) methods, which represent a document as a collection of word counts or frequencies and compute the tone based on the relative frequency of terms associated with tightening or easing monetary policy (Picault et al., 2022; Bennani and Neuenkirch, 2017; Baranowski et al., 2021). Despite their simplicity and interpretability, these approaches cannot properly capture context or semantic nuance, because they treat text as unordered collections of words or phrases. For example, a dictionary-based method that flags the bigram *"raise rates"* as hawkish would assign the same classification to the sentences *"we will raise rates"* and *"we will not raise rates"*, not taking into account the negation (Pfeifer and Marohl, 2023).

A major breakthrough to overcome these limitations came with the introduction of the Transformer architecture by Vaswani et al. (2017), and the subsequent development of BERT (Bidirectional Encoder Representations from Transformers) by Google AI (Devlin et al., 2018). Transformer-based models generate bidirectional contextual embeddings, capturing word order, syntax, and long-range semantic relationships, so that the representation of a word varies depending on the sentence in which it appears. Transformer-based models have rapidly become the state-of-the-art in natural language processing and they have proven particularly well suited for capturing key aspects of policy communication.

Pfeifer and Marohl (2023) and Gambacorta et al. (2024) employ RoBERTa-based architectures to classify the monetary policy stance of central bank communication at the sentence level. The former trained a model on communication from several central banks for a binary classification task (positive vs. negative). The latter focused on Federal Open Market Committee statements and formulated a three-class classification problem (dovish, hawkish, neutral).

Other works go a step further and construct aggregate indices that measure the overall hawkishness–dovishness of an entire communication. Shah et al. (2023) develop a monetary policy stance index from Fed documents based on a three-class sentiment RoBERTa-large model. Lupton et al. (2023) and Lupton and Weitzenfeld (2024) employ five-class sentiment BERT-based models to assess the tone of communications by the ECB, the Federal Reserve, and the Bank of England. In a similar vein, Hallmark et al. (2024) compute the tone of ECB communications based on a five-class RoBERTa model trained on Bloomberg news headlines.

The rapid rise of generative AI models was reflected in the study of monetary policy communication as well. Geiger et al. (2025) assess the sentiment of ECB communications, crafting prompts

for the Llama 3.1 model. However, Gambacorta et al. (2024); Shah et al. (2023); Lupton and Weitzenfeld (2024) show that BERT-based models can still provide superior performance.

Our pipeline takes these contributions as a starting point for further improvement. We expand the dictionary of Shah et al. (2023) to identify economically relevant sentences in step (I). We assess the relevance of each sentence for monetary policy in step (II) following Lupton et al. (2023); Lupton and Weitzenfeld (2024). We prefer RoBERTa models over gen-AI options, because they ensure greater accountability, provide cost-effectiveness, and perform as well or better. We adopt a fine five-class sentiment classification in step (III) to capture the full gradation of policy stance, as in Lupton et al. (2023); Lupton and Weitzenfeld (2024); Hallmark et al. (2024). We develop a metric to address the ordinal and imbalanced nature of the data, inspired by Hallmark et al. (2024). We augment each sentence with information about the inflation environment to reduce ambiguity, as in Lupton et al. (2023); Lupton and Weitzenfeld (2024); Geiger et al. (2025).

Using our tone index to estimate the effect of monetary policy on inflation is inspired by Romer and Romer (2004) and Aruoba and Drechsel (2026). The idea of purging monetary policy from the systematic component endogenous to inflation can be traced back to the seminal contribution by Romer and Romer (2004). Their method relies on central bank forecasts to elicit the economic environment to which monetary policy is responding. Aruoba and Drechsel (2026) show that central bank communication contains valuable information regarding the central bank's assessment of the economy on top of central bank's forecasts.

3 Dataset

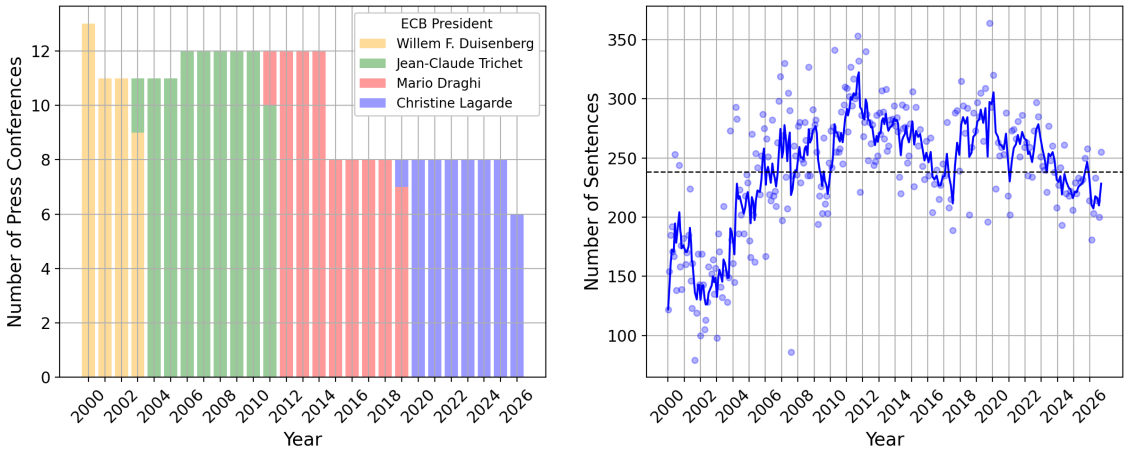
The ECB Governing Council gathers at regular intervals to assess economic and monetary developments and take monetary policy decisions. After each decision, the ECB president, together with the vice-president, holds a press conference, which lasts approximately 60 minutes and consists of two parts: an introductory statement to motivate the decisions and a Q&A session to interact with journalists. As shown in Figure 1(a), the number of press conferences varied over time, because the Council gathered roughly on a monthly basis before 2015, while meetings have been taking place approximately every six weeks since then. Transcripts were scraped directly from the ECB website.[2] As a result, we obtained a dataset containing 270 documents from 2000 to 2026; each document contains on average about 240 sentences, as indicated by the dashed black line in Figure 1(b). We retrieve from the ECB Statistical Data Warehouse the ECB policy rates and the quarterly euro area HICP inflation projections[3]; we retrieve from Bloomberg the monthly euro area HICP index.

4 Textual Tone Score Index

In this section, we describe the procedure to construct the *textual tone score index*. The main steps are briefly summarized below and illustrated in Figure 2.

- *Step 1.* An initial *relevance filtering step*, in which we retain only economics-related sentences in ECB communications.
- *Step 2.* A *relevance classification model* that classifies each economics-related sentence as either relevant to monetary policy or not.

[2] `https://www.ecb.europa.eu/press/press_conference/monetary-policy-statement/html/index.en.html`.
[3] ECB Macroeconomic Projection Database series MPD.Q.U2.HIC.A



(a) ECB Press Conferences per Year

(b) Total Number of Sentences per Communication

Figure 1: Overview of the ECB press conference dataset. Panel 1(a) shows the number of ECB press conferences held each year from 2000 to 2026, bars are colored according to the ECB president in office. Panel 1(b) shows the number of sentences per press conference (scatter points) together with an exponentially weighted moving average. The dashed black line denotes the overall average length of ECB press conferences, equal to roughly 240 sentences.

- *Step 3.* A *sentiment classification model* that assigns a sentiment (textual tone) label to each sentence.
- *Step 4.* The computation of the *textual tone score* for each ECB communication, defined as a weighted average of sentence-level textual tone scores using the relevance weights.

4.1 Step 1: First-Pass Relevance Filtering

The purpose of this initial filtering step is to identify and remove economically irrelevant sentences, that is, sentences that are not useful for computing the textual tone score, such as greetings or questions from journalists. This step serves two main purposes. First, it improves computational efficiency by preventing these sentences from being passed to the more expensive classification models. Second, it reduces noise by removing completely irrelevant sentences, such as "*We are now ready to take your questions*", before the relevance classification step, ensuring that the *relevance model* only distinguishes between sentences that are relevant to monetary policy and those that are not. On average, 45% of the sentences in each document are removed (approximately 110 out of 240). The filtering procedure follows five steps described below and details are provided in Appendix B:

1. *Sentence splitting.* We split the full document text (monetary policy statement and Q&A transcript) into sentences using "." and "?", which mark the end of a statement sentence or a question.

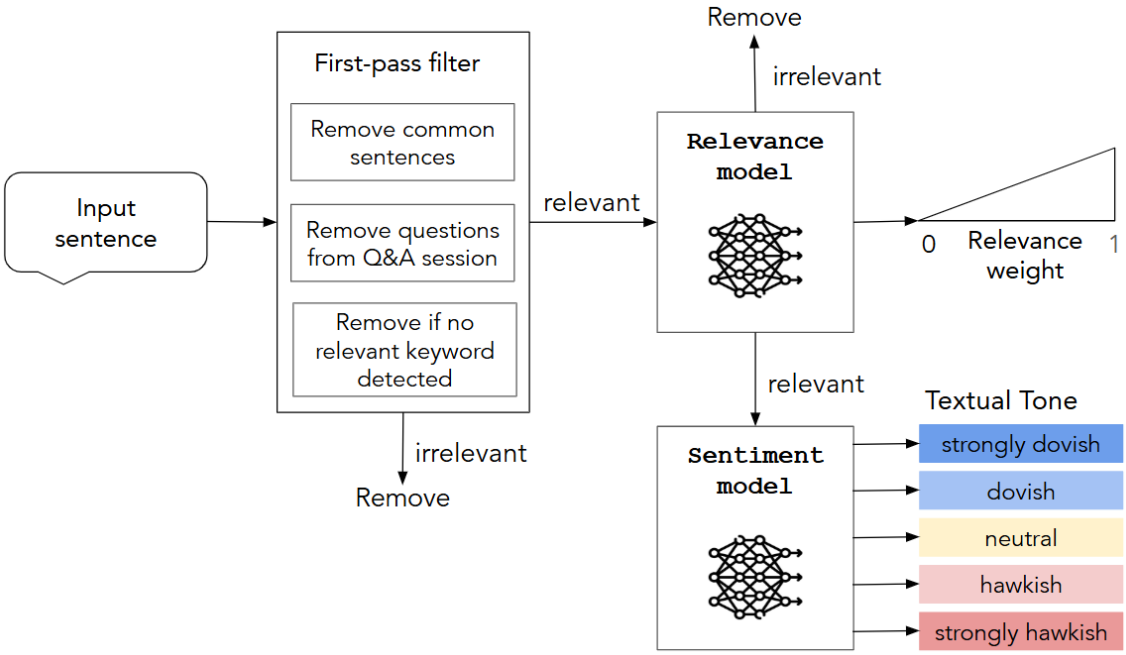


Figure 2: Overview of the NLP pipeline for sentiment classification.

2. *Sentence merging using connector keywords.* We merge adjacent sentences when the second sentence begins with a *connector keyword*. We do this because the meaning of a passage may be distributed across multiple sentences. If treated separately, individual sentences may lose their meaning, because, for example, the second sentence might omit the subject and refer implicitly to the previous one.

3. *Removal of common sentences.* We remove overly common sentences, such as introductory greetings.

4. *Removal of journalists' questions.* We remove all questions asked by journalists during the Q&A session.

5. *Economics-related keyword filtering.* We apply a last filtering step by retaining only sentences that contain at least one *economics-related keyword* from a predefined dictionary, adapted from Shah et al. (2023) to better reflect ECB-specific terminology.

The effectiveness of these steps was validated through a series of manual checks on randomly selected samples of both removed and retained sentences, ensuring that the filtering procedure behaved as intended.

4.2 Preparatory Steps for Classification (Steps 2 and 3)

Before proceeding with the relevance and sentiment classification tasks, we first prepare the data for training and introduce the performance evaluation metrics used to assess model performance.

4.2.1 Data Preparation

For a classification task, we require a labeled dataset of the form

$$\mathcal{D} = \{(x_i, y_i),\ 1 \leq i \leq \mathcal{N}\}, \tag{1}$$

where x_i is the i-th sentence, y_i is the corresponding classification label used to train the models, and $\mathcal{N}$ is the total number of labeled elements in the dataset. Given the full set of document texts $\mathcal{X} = \{X_1, X_2, \ldots\}$, we then proceed with the *manual labeling* of sentences from a selected subset of documents in $\mathcal{X}$.

Two annotators assigned a value $y_i \in \{-2, -1, 0, 1, 2, 3\}$ to each sentence. The labels are defined as follows, and further details are provided in Appendix C:

- *Label -2 (Strongly Dovish)*: Sentences that explicitly indicate a monetary policy action aimed at stimulating the economy and/or raising inflation, such as lowering interest rates or expanding the central bank's balance sheet. These statements reflect a strongly accommodative policy stance with a clear intention to act. For example,
  - "*Looking ahead, our monetary policy stance will remain accommodative for as long as necessary, in line with the forward guidance provided in July.*" (October 2, 2013)
  - "*In any case, we will conduct net asset purchases under the PEPP until the Governing Council judges that the coronavirus crisis phase is over.*" (June 4, 2020)
- *Label -1 (Dovish)*: Sentences that describe weak economic conditions, such as slow growth, high unemployment, or inflation below target. These reflect negative economic considerations, but do not indicate a concrete policy action. For example,
  - "*The risks surrounding the economic outlook for the euro area remain on the downside.*" (October 2, 2014)
  - "*Underlying price pressures in the euro area are expected to remain subdued over the medium term.*" (August 1, 2013)
- *Label 0 (Neutral)*: Sentences that are neither dovish nor hawkish, providing general information without suggesting a direction for monetary policy. For example,
  - "*Medium to long-term inflation expectations continue to be firmly anchored in line with price stability.*" (October 2, 2013)
  - "*The new Eurosystem staff projections show headline inflation averaging 2.1 per cent in 2025, 1.9 per cent in 2026, 1.8 per cent in 2027 and 2.0 per cent in 2028.*" (December 18, 2025)
- *Label 1 (Hawkish)*: Sentences that highlight strong economic conditions, rising inflation, or low unemployment. These reflect positive economic considerations, but do not explicitly signal a tightening action. For example,
  - "*The risks to the inflation outlook continue to be on the upside and have intensified, particularly in the short term.*" (July 21, 2022)
  - "*To sum up, annual inflation rates are projected to remain elevated in 2006 and 2007, with risks to this outlook remaining clearly on the upside.*" (November 2, 2006)

- *Label 2 (Strongly Hawkish)*: Sentences that explicitly indicate a tightening of monetary policy, such as raising interest rates, in response to high inflation or an economy growing faster than desired. These statements reflect a strongly restrictive policy stance with a clear intention to act. For example,
  - "*Our future decisions will ensure that the key ECB interest rates will be set at sufficiently restrictive levels for as long as necessary to achieve a timely return of inflation to our two per cent medium-term target.*" (July 27, 2023)
  - "*We took today's decision, and expect to raise interest rates further, to ensure the timely return of inflation to our two per cent medium-term inflation target.*" (October 27, 2022)
- *Label 3 (Irrelevant to Monetary Policy)*: Sentences that do not pertain to monetary policy and are included for training the relevance classification model. The initial relevance filter (Step 1) is not sufficient to exclude all sentences that, although related to the broader economy, are not specifically relevant to monetary policy. Therefore, an additional classification step is required to further distinguish between sentences that are useful for computing the tone (i.e., those directly related to monetary policy) and those that are not. For example,
  - "*To maintain the credibility of the fiscal policy framework it is essential to fully abide by its rules and to implement them in every respect.*" (October 2, 2003)
  - "*National fiscal and structural policies should aim at making the economy more productive and competitive, which would help to raise potential growth and reduce price pressures in the medium term.*" (July 18, 2024)

From the corpus of 270 press conference transcripts, we randomly selected approximately four documents per year, for a total of 104 documents (about 39% of the transcript corpus), for manual labeling as described above.

Before annotation, we applied the filtering procedure reported in Section 4.1 in order to exclude sentences that are completely irrelevant and will not be used in either the relevance or sentiment classification tasks. After filtering, the labeled dataset $\mathcal{D}$ consists of approximately 13,500 sentences. As illustrated in Figure 3, the class distribution appears roughly normal, with the extreme classes accounting for about 5% of the labeled dataset $\mathcal{D}$.

Another issue concerns the *ambiguity* of certain sentences, whose interpretation depends on the prevailing economic context. This aspect was taken into account during manual labeling, as the annotators had access to the relevant macroeconomic information at the time of each press conference. However, a classification model does not have access to this contextual knowledge unless it is explicitly provided with additional information. For example, the sentence "*The Governing Council is determined to ensure that inflation returns to its 2% medium-term target in a timely manner*" can be interpreted as strongly hawkish when inflation is well above the target, but it may convey a dovish tone when inflation is below target.

To address this issue, we concatenate[4] to each sentence an inflation context label, which specifies the inflation environment with respect to the ECB inflation target. Initially, the ECB price stability mandate was articulated as a "medium-term inflation below 2%"; the target was first modified in May 2003 as "medium-term inflation below, but close to, 2%"; it was finally reviewed in July 2021 as "symmetric 2% medium-term inflation". Since the "medium term" evaluation horizon is inherently

[4]The context label c_i is then concatenated to the sentence text x_i to form an augmented input $x_i^c = x_i \ \langle /s \rangle \ c_i$. This results in an enriched labeled dataset $\mathcal{D}^c = \{(x_i^c, y_i),\ 1 \leq i \leq \mathcal{N}\}$.

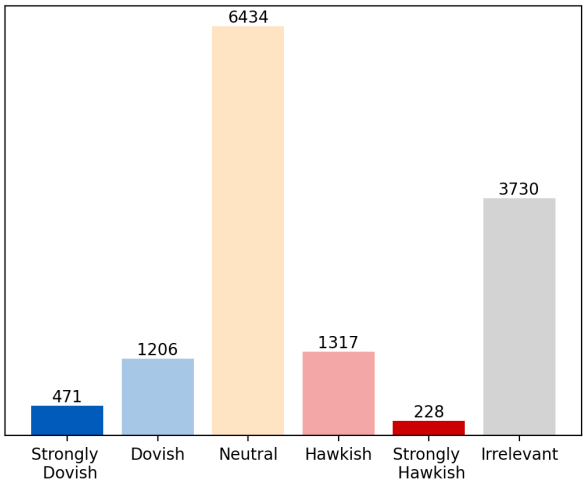


Figure 3: Distribution of sentence-level labels in the manually annotated dataset $\mathcal{D}$. The figure highlights the strong class imbalance, with the extreme labels (-2 and 2) accounting for only a small fraction of the observations.

vague, we assign to each press conference the highest reading of the closest quarterly ECB HICP inflation forecasts for the next four quarters, $\pi_{Q1:Q4}^{forecast}$. The inflation context label states whether this inflation reading is below, at or above the prevailing target, which is specified as follows and results in the inflation target time line displayed in Figure 4.

$$\text{Inflation Target} = \begin{cases} \pi_{Q1:Q4}^{forecast} < 2.00\% & \text{Before May 2003,} \\ 1.80\% \leq \pi_{Q1:Q4}^{forecast} < 2.00\% & \text{Between May 2003 and July 2021,} \\ \pi_{Q1:Q4}^{forecast} = 2.00\% & \text{After July 2021} \end{cases}$$

The final step in preparing the data for training is randomly splitting the manually labeled dataset into training, validation, and test sets.[5] Given the strong class imbalance, we apply a stratified split to ensure that each subset preserves the same class proportions as the full dataset $\mathcal{D}^c$. This prevents the validation and test sets from containing too few observations from minority classes, such as -2 and 2. We denote the training, validation, and test sets as $\mathcal{D}_{\text{Tr}}^c$, $\mathcal{D}_{\text{V}}^c$, and $\mathcal{D}_{\text{Te}}^c$, respectively.

4.2.2 Performance Evaluation Metrics

We require an evaluation metric to assess the performance of a model, that is, its ability to correctly classify observations and to generalize to new and unseen data drawn from the same distribution.

[5]The training set (in-sample) is used to train the model and learn patterns from the data. The validation set (out-of-sample) is used for model selection and hyperparameter tuning, helping to prevent overfitting to the training data. The test set (out-of-sample) is reserved for an unbiased evaluation of the final model's performance.

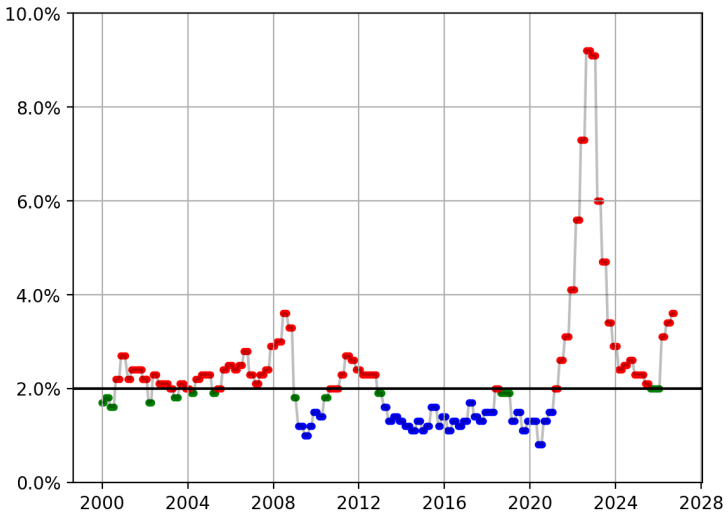


Figure 4: Inflation timeline used to construct the context labels, which help resolve sentiment ambiguity in sentence classification. Red dots represent an inflation reading above the target, green at the target, blue below the target.

Number of press conference texts	104
Share of press conference texts	39%
Number of labeled sentences	≈13,500
Total number of labels	6
Number of sentiment labels	5

Table 1: Overview of the dataset used for training and evaluation.

For classification tasks, a standard performance metric is *accuracy*, which is defined as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{2}$$

where y_i and $\hat{y}_i$ denote the true and predicted labels for observation x_i, respectively, and $\mathbb{I}(\cdot)$ is the indicator function. However, as discussed in Section 4.2.1, the class distribution in our dataset is highly imbalanced: among sentences carrying a sentiment label (-2 to 2), the neutral class represents approximately two thirds of the observations, while the extreme sentiment classes account for only about 7% (about 48% and 5%, respectively, of the full labeled dataset $\mathcal{D}$, which also includes sentences irrelevant to monetary policy). Standard accuracy assigns equal weight to each observation rather than to each class, implying that errors in minority classes contribute less to the overall score. As a result, accuracy tends to favor models that perform well on majority classes at the expense of rarer but important categories (such as -2 and 2). To address this issue, we can adopt *balanced accuracy*, which assigns equal importance to each class by averaging class-wise

recall.[6] In the multiclass case, balanced accuracy is defined as

$$\text{Balanced Accuracy} = \frac{1}{\mathcal{C}} \sum_{k=1}^{\mathcal{C}} \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{3}$$

where $\mathcal{C}$ is the number of classes, and TP_k and FN_k denote the number of true positives and false negatives for class k, respectively. This metric ensures that performance on minority classes contributes equally to the final evaluation, making it more appropriate for our imbalanced classification setting. To interpret the accuracy of a model we can compare it to the baseline accuracy. For binary classification, the baseline accuracy is 50%, while for a 5-class classification problem, the baseline is 20%. Any accuracy above the baseline shows an improvement over random guessing.

Another important consideration is that sentiment classes are *ordered* rather than nominal:

"strongly dovish" < "dovish" < "neutral" < "hawkish" < "strongly hawkish".

Consequently, not all misclassifications are equally severe. For example, classifying a *strongly dovish* (-2) sentence as *dovish* (-1) is substantially less problematic than classifying it as *strongly hawkish* (2). This introduces an *ordinality* issue in addition to class imbalance, implying that an appropriate evaluation metric should penalize classification errors more heavily the larger the distance between y_i and $\hat{y}_i$ in the ordered scale. A commonly used metric for ordinal classification problems is the *Quadratic Weighted Kappa* (QWK). It measures the level of *agreement* between two raters—here, the human annotators and the classification model—who classify items into an ordered set of categories. This measure ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement no better than random. In our setting, higher values of QWK correspond to better classification performance, with 1 representing perfect classification. The QWK is defined as:

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{4}$$

where

- O is the *observed confusion matrix* comparing predicted and true labels. It is a $\mathcal{C} \times \mathcal{C}$ matrix, where diagonal elements correspond to normalized counts of correctly classified observations and off-diagonal elements to misclassifications.
- W is the *weight matrix*, whose entries quantify the penalty associated with each type of disagreement. Each element is defined as

$$W_{i,j} = \frac{(i-j)^2}{(\mathcal{C}-1)^2}. \tag{5}$$

  The diagonal entries are equal to zero (when $i = j$), reflecting the absence of penalty for correct classifications. The weights lie in the interval $[0, 1]$ and increase with the severity of the error, measured by the distance between the predicted class i and the true class j on the ordinal scale. When $\mathcal{C} = 5$, the possible penalty values are $\{0, 0.0625, 0.25, 0.5625, 1\}$, with the best-case scenario corresponding to $i = j$ and the worst-case scenario occurring when $|i - j| = \mathcal{C} - 1$.[7]

[6] Also known as True Positive Rate (TPR).

[7] Labels are mapped to integer indices as $-2 \mapsto 0,\ -1 \mapsto 1,\ 0 \mapsto 2,\ 1 \mapsto 3,\ 2 \mapsto 4$.

- E denotes the *expected confusion matrix* under chance agreement, computed by preserving the marginal class distributions.

While QWK is more robust than standard accuracy in the presence of class imbalance, it is not explicitly designed to address it. To account for both ordinality and class imbalance, we introduce a balanced variant of this metric, which we refer to as *Balanced Quadratic Weighted Kappa* (BQWK). The BQWK is computed by evaluating the QWK separately for each class and then averaging the resulting values across the $\mathcal{C}$ classes. This ensures that each class contributes equally to the final score, independently of its frequency, while preserving the ordinal structure of the sentiment labels.

4.3 Step 2: Relevance Model

After the first-pass filtering and data preparation steps, we proceed to the next stage of the NLP pipeline, namely *relevance classification*, where sentences are classified as either relevant to monetary policy or not.

We begin by considering the context-augmented labeled dataset $\mathcal{D}^c$ and removing duplicate sentences, as several sentences recur across different monetary policy statements. After deduplication, the dataset contains 12,899 observations. Next, we remap the original labels y_i to construct a *binary relevance target*. Specifically, sentences originally labeled as 3 are mapped to 0, while the remaining labels associated with sentiment values from -2 to 2 are mapped to 1, indicating relevance to monetary policy. Finally, we split the dataset into training, validation, and test sets using an 80%–10%–10% split, obtaining the datasets $\mathcal{D}_{\text{Tr}}$, $\mathcal{D}_{\text{V}}$, and $\mathcal{D}_{\text{Te}}$, with sizes 10,319, 1,290, and 1,290, respectively. The training set contains 2,911 not-relevant and 7,408 relevant sentences. The split is performed in a stratified manner to preserve the same class proportions across all subsets (Sadaiyandi et al., 2023).

With the dataset prepared, we proceed to the classification stage. We use the Transformer-based model *RoBERTa*[8] (see Appendix A.1 for an overview of Transformers and the RoBERTa architecture), implemented via the `transformers` library from Hugging Face. Specifically, we adopt the base version of the model, which consists of 12 layers and 125 million parameters.[9] We load both the pretrained RoBERTa model and its corresponding tokenizer.[10] The tokenizer is applied to the sentences in the training, validation, and test sets, with a maximum sequence length of 256 tokens. Sentences exceeding this length (which are rare) are truncated, while shorter sentences are padded to ensure a uniform input length across all inputs.[11]

The RoBERTa model is pretrained on large-scale, general-purpose text corpora, but it must be adapted to our specific dataset and to the binary relevance classification task. This adaptation is performed through *fine-tuning* (see Appendix A.5 for details), which consists of retraining a subset of the model parameters using a small learning rate. Training is carried out using *mini-batch gradient descent*. At each iteration, a mini-batch of 32 sentences is randomly sampled from the training set $\mathcal{D}_{\text{Tr}}$ and used to compute the gradient of the loss function and update the model parameters. We optimize a *weighted cross-entropy loss* function to account for class imbalance. For

[8] `https://huggingface.co/FacebookAI/roberta-base`. Two versions are available, base and large, differing in the number of parameters. We found that the base version performs slightly better for the relevance classification task.

[9] We also experimented with the large version of RoBERTa, but it did not yield measurable performance improvements, so we retained the smaller model.

[10] The tokenizer converts the sentence text into tokens, that is, basic textual units such as words or subwords.

[11] Padding consists of adding a special padding token with value 1 to the end of shorter sequences so that all inputs have the same length, which is required for efficient batch processing in Transformer models.

a binary classification problem with classes $\mathcal{C} = \{0, 1\}$ and a mini-batch $\mathcal{B}$, the loss is defined as

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[w_1 \cdot y_i \log(\hat{p}_i) + w_0 \cdot (1 - y_i) \log(1 - \hat{p}_i) \right], \tag{6}$$

where $y_i \in \{0, 1\}$ is the true label, $\hat{p}_i$ is the predicted probability that sentence x_i belongs to class 1, and w_1 and w_0 are the class weights for the positive (relevant) and negative (not-relevant) classes. The weights are computed as[12]

$$w_c = \frac{|\mathcal{D}_{Tr}|}{|\mathcal{C}| \cdot |\{i \in \mathcal{D}_{Tr} : i = c\}|}, \quad c \in \mathcal{C}. \tag{7}$$

In our setting, the relevant class ($c = 1$) is the majority class, with weight

$$w_1 = \frac{10{,}319}{2 \cdot 7{,}408} \approx 0.70,$$

while the not-relevant class ($c = 0$) is the minority class, with weight

$$w_0 = \frac{10{,}319}{2 \cdot 2{,}911} \approx 1.77.$$

Assigning a higher weight to the minority class increases the penalty associated with misclassifications of not-relevant sentences, thus mitigating the effects of class imbalance during training.

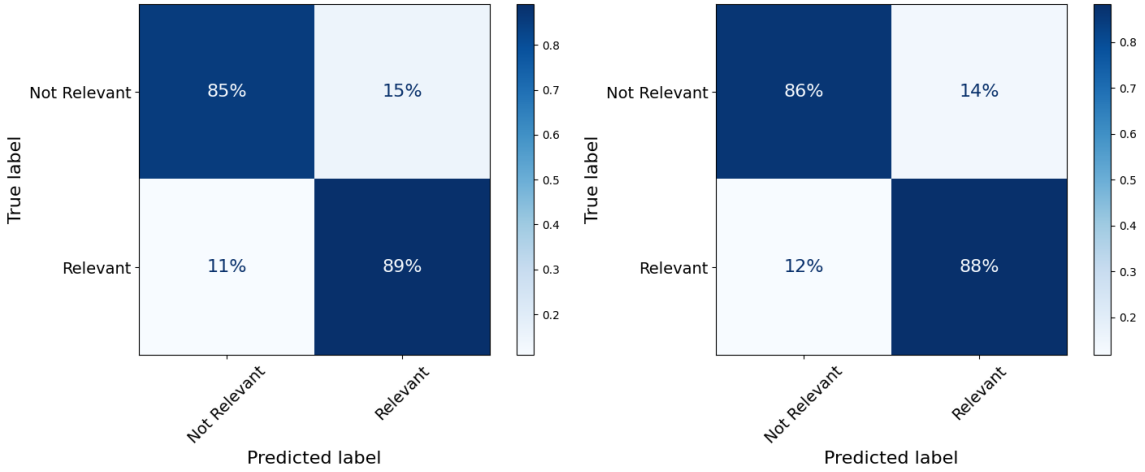
The training procedure also requires the selection of several *hyperparameters*, which can heavily influence model performance. The hyperparameters we consider are:

- Number of epochs, set to 10.
- Learning rate, for which we tested three values: $1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}$.
- Number of frozen layers l, i.e., layers whose weights remain fixed while the last $12 - l$ layers are fine-tuned. We tested $l \in \{0, 3, 6, 9\}$.

We train the model on all combinations of learning rate and frozen layers to identify the best configuration. The *best model* is selected as the one achieving the highest balanced accuracy on the validation set $\mathcal{D}_V$, which contains data unseen during training. To prevent overfitting, we implement *early stopping* with a *patience* of 3 epochs. This means that if the balanced accuracy on the validation set does not improve for 3 consecutive epochs, training is stopped and the model weights corresponding to the highest validation balanced accuracy are restored.

The best relevance model is obtained with 0 frozen layers and a learning rate of 1×10^{-5}, achieving a balanced accuracy of 0.8713 on the validation set. The model is then evaluated on the test set, containing entirely unseen data, yielding a balanced accuracy of 0.8725, which demonstrates strong generalization. Figure 5 shows the confusion matrices for the validation and test sets.

[12] The class weights are computed as in `sklearn.utils.class_weight.compute_class_weight`.



(a) Validation Set Confusion Matrix

(b) Test Set Confusion Matrix

Figure 5: Classification performance of the RoBERTa-base relevance model. The left panel shows the confusion matrix on the validation set, while the right panel reports the confusion matrix on the test set.

4.4 Step 3: Sentiment Model

Following the relevance classification model, the next step in the NLP pipeline is *sentiment classification*, where sentences that are relevant to monetary policy are assigned to one of five sentiment classes, as described in Section 4.2.1. We begin by considering the context-augmented dataset $\mathcal{D}^c$ and remove duplicate sentences. We then retain only sentences with sentiment labels ranging from -2 to 2, discarding those labeled as 3. To facilitate model training, we remap the original sentiment labels according to the transformation

$$-2 \mapsto 0,\ -1 \mapsto 1,\ 0 \mapsto 2,\ 1 \mapsto 3,\ 2 \mapsto 4.$$

As discussed previously, the sentiment classes are strongly imbalanced, with approximately 67% of sentences belonging to the neutral class and only about 7% belonging to the extreme classes (*strongly dovish* and *strongly hawkish*). The dataset is then split into training, validation, and test sets using a 70%–15%–15% split, performed in a stratified manner to preserve class proportions across all subsets. This results in the training, validation, and test datasets $\mathcal{D}_{Tr}$, $\mathcal{D}_V$, and $\mathcal{D}_{Te}$, with sizes 6,482, 1,389, and 1,389, respectively.

For this task we experimented with different variants of the Transformer-based RoBERTa model,[13] and, unlike the relevance classification task, we employ the *large* version of RoBERTa, which consists of 24 layers and approximately 355 million parameters. In contrast to the relevance task, the larger model yields better performance for sentiment classification, likely due to the greater complexity of the task, which benefits from increased model capacity. After loading the model and

[13]We also experimented with the central-bank language models (CB-LMs) proposed by Gambacorta et al. (2024), which are RoBERTa models further pre-trained on a large corpus of central bank speeches, policy documents, and research papers. However, after fine-tuning on our labeled dataset, these models performed worse on the sentiment classification task than the standard RoBERTa-large model.

its tokenizer, we tokenize all sentences in the training, validation, and test sets, using a maximum sequence length of 256 tokens, as in the relevance classification step.

The model is fine-tuned on the sentiment dataset and adapted to the new task using mini-batch gradient descent. At each training iteration, a mini-batch of 32 sentences is randomly sampled from the training set and used to update the model parameters. We optimize a *weighted cross-entropy loss* function to account for class imbalance.[14] For a 5-class classification problem with classes $\mathcal{C} = \{0, 1, 2, 3, 4\}$ and a mini-batch $\mathcal{B}$, the weighted cross-entropy loss is defined as

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{c \in \mathcal{C}} w_c \cdot \mathbb{I}\{y_i = c\} \cdot \log\left(\hat{p}_{i,c}\right), \tag{8}$$

where $\hat{p}_{i,c}$ is the predicted probability that sentence x_i belongs to class c, y_i is the ground truth label of sentence x_i, w_c is the class-specific weight, and $\mathbb{I}\{y_i = c\}$ is an indicator function equal to 1 if $y_i = c$ and 0 otherwise. This loss function is a direct generalization of the one used for the relevance model to the multiclass setting. Class weights are computed analogously in order to account for class imbalance. Specifically,

$$w_0 = \frac{6{,}482}{5 \cdot 305} = 4.25, w_1 = \frac{6{,}482}{5 \cdot 812} = 1.60, w_2 = \frac{6{,}482}{5 \cdot 4{,}340} = 0.30, w_3 = \frac{6{,}482}{5 \cdot 897} = 1.45, w_4 = \frac{6{,}482}{5 \cdot 128} = 10.13.$$

As expected, minority classes receive substantially larger weights, ensuring that misclassifications in these classes contribute more heavily to the loss and mitigating the strong class imbalance present in the data.

The hyperparameters that we consider are:

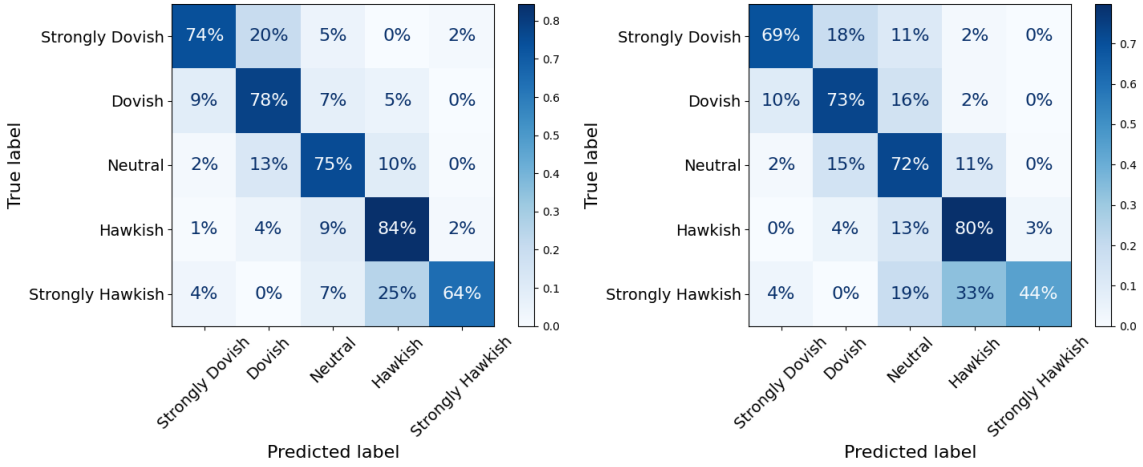
- Number of epochs, set to 30.
- Learning rate, for which we tested three values: $1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}$.
- Number of frozen layers l, i.e., layers whose weights remain fixed while the remaining $24 - l$ layers are fine-tuned. We tested $l \in \{3, 6, 9, 12, 15, 18\}$.

We train the model on all combinations of learning rate and frozen layers to identify the best configuration. The *best model* is selected as the one achieving the highest BQWK on the validation set $\mathcal{D}_V$. To prevent overfitting, we implement *early stopping* with a *patience* of 3 epochs,[15] and restore the model weights corresponding to the highest validation BQWK.

The best-performing sentiment classification model is obtained with 15 frozen layers and a learning rate of 3×10^{-5}, achieving a BQWK score of 0.7560 on the validation set and 0.7234 on the test set. Figure 6 reports the corresponding confusion matrices for the validation and test datasets. Importantly, the upper-right and lower-left regions of the matrices are largely empty, suggesting that severe misclassifications, such as labeling a *strongly dovish* sentence as *strongly hawkish*, are rare. This pattern indicates that the inclusion of context labels is effective in mitigating extreme ordinal errors.

[14] We also tried the CORN loss `https://raschka-research-group.github.io/coral-pytorch/api_subpackages/coral_pytorch.losses/`, which explicitly accounts for class ordinality as discussed in Section 4.2.2, but we did not observe performance improvements.

[15] We set the patience to three epochs, one more than in the relevance classification task, to account for the greater complexity of the sentiment classification problem. Given the larger number of classes and the more nuanced distinctions between them, the model may require additional training time to converge to a stable optimum.



(a) Validation Set Confusion Matrix

(b) Test Set Confusion Matrix

Figure 6: Classification performance of the RoBERTa-large sentiment model. The left panel shows the confusion matrix on the validation set, while the right panel reports the confusion matrix on the test set.

4.5 Step 4: Textual Tone Score Index Computation

Now that the main components of the NLP pipeline have been introduced, the final step is to combine them into a single end-to-end procedure. For each press conference, we feed the full text of the introductory statement together with the transcript of the Q&A session into the pipeline. As a result, for each press conference we obtain a label for every sentence in the text. Sentences that are relevant to monetary policy are assigned a sentiment score ranging from -2 to 2, while all remaining sentences are assigned the label 3 indicating irrelevance for monetary policy.

Figure 7 reports the raw counts of sentence-level sentiment labels over time. Sustained numbers of *hawkish* and *strongly hawkish* sentences before the great financial crisis and the sovereign debt crisis reflect overheating inflation fueled primarily by increasing oil prices. The most prominent increase in hawkish language occurs in 2022–23, coinciding with the sharp rise in inflation. Conversely, *dovish* and *strongly dovish* sentences are observed during periods of economic downturn, most notably during the great financial crisis, the sovereign debt crisis and the Covid-19 pandemic. Moreover, the use of accommodative language between 2014 and 2021 reflects the adoption of forward guidance and asset purchases as unconventional policies to stimulate the economy.

Based on these sentence-level classifications, we compute a single *textual tone score* for each press conference, which summarizes the overall dovish–hawkish tone of the communication. The tone score for the i-th press conference is defined as

$$t_i = \frac{1}{\mathcal{N}} \sum_{j=1}^{\mathcal{N}} w_j^{rel} \cdot \hat{y}_j, \tag{9}$$

where $\mathcal{N}$ is the total number of relevant sentences in the press conference, $\hat{y}_j$ is the sentiment label predicted for sentence j, and w_j^{rel} is a *relevance weight* derived from the relevance classification

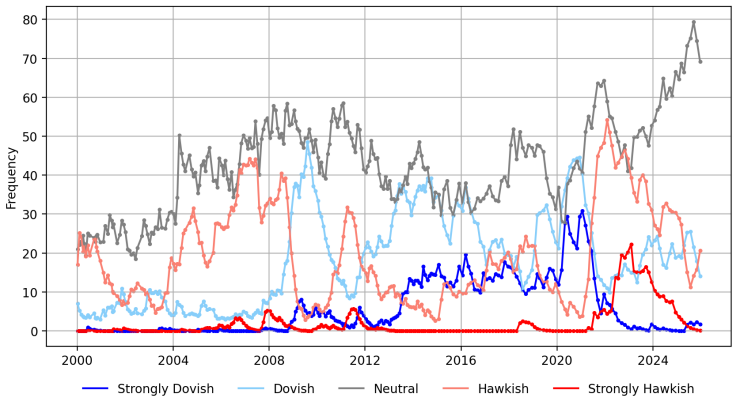


Figure 7: Raw counts of sentence-level sentiment labels, ranging from *strongly dovish* to *strongly hawkish*, for each press conference.

model's *confidence score*.[16]

We compute the textual tone score for each of the 270 press conferences held between January 2000 and September 2026. In theory, the score ranges from -2 to 2, corresponding to the extreme case in which all monetary policy-relevant sentences receive a relevance weight of 1 and are classified as either *strongly dovish* or *strongly hawkish*. In practice, the observed range is considerably narrower, spanning from -1.05 (April 30, 2020), the most dovish communication in the sample, to 0.67 (December 15, 2022), the most hawkish one.

The resulting textual tone index is plotted against context HICP inflation (cf. Figure 4) in Figure 8, and against the ECB policy rate and the shadow rate computed by Krippner (2013) in Figure 9. Overall, the index co-moves with both inflation and the policy rate, suggesting that the constructed textual tone measure captures meaningful variations in the monetary policy stance over time. Focusing on the policy rate, a preliminary visual inspection shows that changes in the communication tone anticipate variations of the policy rate, implying that communication prepares the ground for monetary policy action. Moreover, the shadow rate estimates capture the accommodation performed by the ECB through unconventional policy measures after the policy rate had hit the effective zero lower bound.

We next examine the textual tone score at a topic-specific level, focusing on four main dimensions: *monetary policy*, *inflation*, *economic activity*, and *financial and monetary conditions*. This analysis allows us to compare how the tone associated with different topics evolves over time and to point out potential discrepancies across thematic dimensions. For each topic, we define a set of keywords that identify the corresponding thematic content. Each sentence in the press conferences is then associated with one or more topics based on the presence of these keywords. The keyword dictionaries contain 170 entries for monetary policy, 108 for inflation, 409 for economic activity, and 224 for financial and monetary conditions.[17] We compute the topic-specific textual tone scores in the same manner as the overall tone score, but restricting the calculation to sentences that contain

[16] The confidence score reflects the model's estimated probability that a sentence is relevant for monetary policy.

[17] Examples of monetary policy keywords include "QE", "pre-commit", and "interest rate". Inflation-related keywords include "HICP", "price pressure", and "deflation". Growth-related keywords include "GDP", "dampen growth", and "labour market", while finance-related keywords include "financial markets", "bond", and "stock price".

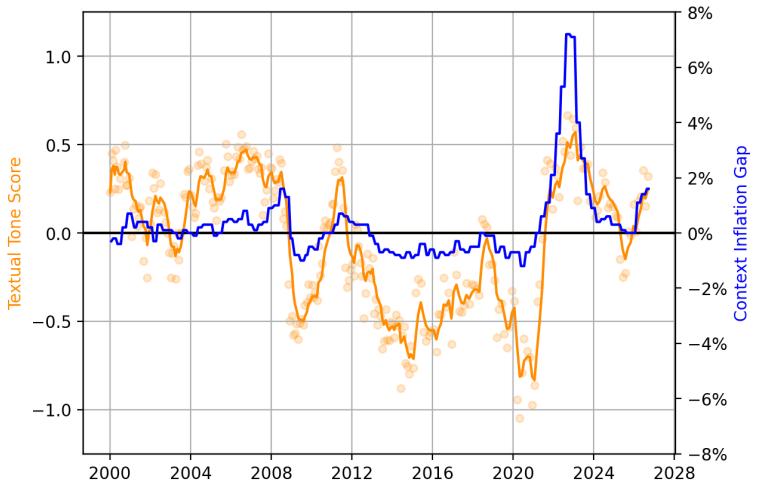


Figure 8: Textual tone score (orange) and context inflation gap (blue). The orange line represents the exponentially weighted moving average of the tone score with a half-life of 60 days, while the orange dots denote the raw tone scores computed for individual press conferences. The context inflation gap is computed with respect to the 2% target.

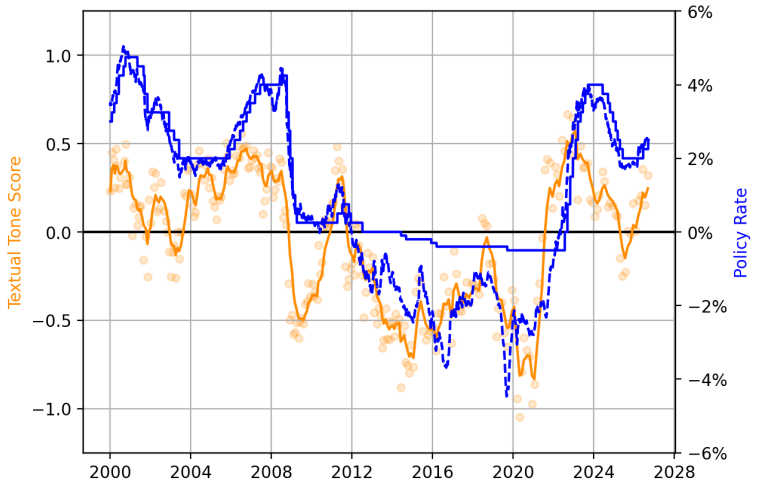


Figure 9: Textual tone score (orange), ECB policy rate (blue) and ECB shadow rate computed by Krippner (2013) (dashed blue). The ECB policy rate is the ECB marginal refinancing operation rate until 08-Oct-2008, when full allotment refinancing operations were announced, and the ECB deposit rate thereafter. The orange line represents the exponentially weighted moving average of the tone score with a half-life of 60 days, while the orange dots denote the raw tone scores computed for individual press conferences.

at least one keyword associated with the given topic.

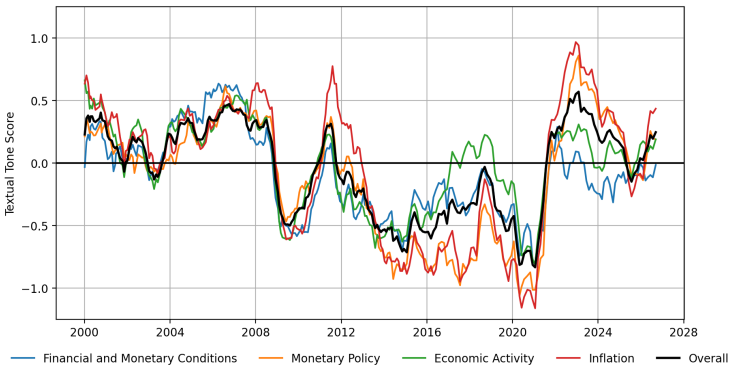


Figure 10: Comparison of textual tone scores associated with different topic categories. The black line represents the overall textual tone index, while the blue, orange, green, and red lines report the tone scores associated with financial and monetary conditions, monetary policy, economic activity, and inflation, respectively.

Figure 10 displays the four topic-specific tone indices together with the overall tone index. It can be noted that after 2022 communication turned hawkish with regard to inflation and monetary policy, while growth received a more moderate assessment. This is consistent with the utmost urgency of the fight against extreme inflationary pressures in the context of the fragile post-Covid recovery. Symmetrically, between 2016 and 2020 communication turned extremely dovish with regard to inflation and monetary policy, while growth assessment improved. This is consistent with the systematic deployment of unconventional monetary policies to stabilize inflation at the target amid improvements in the growth outlook. Overall, it can be noted that monetary policy and inflation broadly moved together, reflecting the price stability mandate of the ECB.

5 A Macroeconomic Validation

Figures 8 and 9 show a meaningful correlation between the tone score index and the context inflation rate, the ECB policy rate and the shadow rate, thus providing a first visual validation that the tone score index successfully captures the ECB monetary policy stance. In this section, we provide a further "macroeconomic" validation inspired by Romer and Romer (2004), exploiting the fact that, since ECB press conferences are primarily intended to explain and clarify the economic motivations behind each monetary policy decision, the tone score index should summarize the central bank's assessment of the economic outlook.

The stream of literature stemming from the seminal contribution of Romer and Romer (2004) acknowledges that the policy action and the policy objective are endogenous, because changes in the economic outlook drive both the evolution of inflation and the central bank reaction. Theoretically, the central bank conducts policy according to an implicit rule

$$\Delta i_t = f(\Omega_t) + \varepsilon_t^{MP}, \tag{10}$$

where changes in the policy instrument i_t are determined by (I) a systematic component $f(\Omega_t)$

depending on the central bank's assessment of the economic outlook Ω_t and (II) a non-systematic monetary policy component ε_t^{MP}.

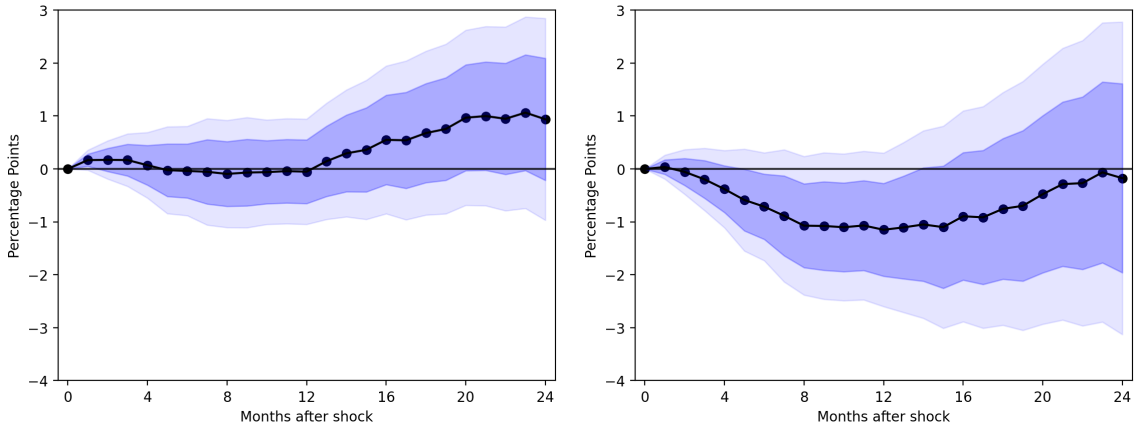
According to this framework, a naive regression of inflation on changes in the policy rate is affected by an omitted variable bias. To see this, we first estimate the following local projection

$$\Delta^m p_{t+m} = \alpha^m + \beta^m \Delta i_t + \sum_{j=1}^{24} \gamma_j^m \Delta^j p_{t-j} + \varepsilon_t^m, \tag{11}$$

where $\Delta^m p_{t+m}$ is the m-month ahead change in the log-HICP price index, Δi_t is the change in the ECB deposit rate decided at the time t meeting by the Governing Council, and $\sum_{j=1}^{24} \Delta^j p_{t-j}$ is a series of past log-HICP price index changes to control for the dynamics of inflation. Since this specification does not control for the economic outlook Ω_t driving both the price level and the policy action, it suffers from omitted variable bias. Moreover, since economic fundamentals are positively correlated with changes in the policy rate through the systematic component, β^m coefficients will be biased upward. Indeed, while economic theory predicts a negative relation between the policy rate and inflation, biased point estimates of β^m are often positive and never statistically different from zero (see Figure 11(a)).

To recover a valid estimate of β^m, one should replace Δi_t in equation (11) with the non-systematic component ε_t^{MP}, which is uncorrelated with economic fundamentals by definition. Romer and Romer (2004) recover the non-systematic component assuming that the systematic component $f(\Omega_t)$ is represented by (I) the previous policy rate to account for mean-reverting behavior i_{t-1} and (II) central bank's forecasts of key macroeconomic indicators $\mathbf{X_t}$, so that they estimate

$$\Delta i_t = \alpha + \beta i_{t-1} + \gamma \mathbf{X_t} + \varepsilon_t^{MP} \tag{12}$$



(a) Naive regression

(b) Purged regression

Figure 11: Estimation of equation (11) (panel a) and (14) (panel b). Dots represent point estimates of coefficient β^m, dark (light) areas depict 16-84% (5-95%) confidence intervals

Since ECB press conferences are primarily devoted to the explanation of the central bank's assessment of the economic outlook underlying the monetary decision, we assume that the tone

score index provides a meaningful proxy for that multi-faceted assessment based on macroeconomic forecasts and economic indicators. Therefore, we replace central bank's forecasts in the original Romer and Romer (2004) specification (12) with the tone score index. Specifically, we regress changes in the ECB deposit rate on the previous policy rate and our textual tone score index

$$\Delta i_t = \alpha + \beta i_{t-1} + \gamma Score_t + \varepsilon_t^{MP} \tag{13}$$

Our estimates confirm the mean-reverting behavior of policy and, most importantly, we recover the expected positive relation between the policy instrument and the tone index, suggesting that the central bank raises the deposit rate as economic fundamentals get stronger.

Variable	Coefficient	Estimates	Statistics	
Constant	α	0.06***	R^2	0.21
i_{t-1}	β	−0.04***	N. Obs.	270
$Score_t$	γ	0.24***		

Table 2: Estimation results of equation (13). * denotes significance at 10%, ** at 5%, *** at 1% based on HAC standard errors.

Exploiting the residuals of specification (13), $\hat{\varepsilon}_t^{MP}$, we are finally able to estimate the local projection

$$\Delta^m p_{t+m} = \alpha^m + \beta^m \hat{\varepsilon}_t^{MP} + \sum_{j=1}^{24} \gamma_j^m \Delta^j p_{t-j} + \varepsilon_t^m, \tag{14}$$

which recovers an economically acceptable negative relation between inflation and the policy instrument (Figure 11(b)). According to our estimates, a 25 basis point shock to the policy rate reduces HICP inflation by 0.27 percentage points over the next twelve months[18] ($\beta^{12} \times \Delta i_t = -1.07 \times 0.25 \approx -0.27$). Overall, these results corroborate the validity of the tone score index as a credible summary of the ECB Governing Council assessment of the euro area economic outlook.

A concern regarding this validation methodology could arise from the fact that the tone score index might provide a biased proxy of the central bank's assessment of the economic outlook, because the President not only discusses the economic environment in press conferences, but also provides forward guidance regarding monetary policy. A first reassurance comes from the fact that forward guidance sentences represent a tiny fraction of relevant sentences in a press conference, thus marginally contributing to the overall tone score of a transcript. To dispel doubts, we evaluate again the purging equation (13) replacing the overall tone score ($Score_t$) with (I) the financial-economic tone score ($FinEcoScore_t$, computed excluding sentences regarding monetary policy) and (II) the economic tone score ($EcoScore_t$, computed excluding sentences regarding monetary policy and finance). Results are virtually unchanged with $\beta^{12} = -0.99$ for the financial-economic tone score and $\beta^{12} = -1.02$ for the economic tone score, compared to $\beta^{12} = -1.07$ for the overall tone score.

[18] A 25 basis point hike is found in the literature to reduce inflation after 12 months by roughly 0.10-0.30% (see Havránek and Rusnak, 2013; Bobasu et al., 2025).

6 The ECB Loss Function

It is a standard assumption in monetary policy theory that the central bank adjusts its policy tools to minimize a quadratic loss function depending on the output gap and the inflation rate (Walsh, 1995). Since the ECB's primary mandate is price stability as stated in Article 127 of the Treaty on the Functioning of the European Union, the ECB loss function can be simplified to a quadratic function depending on the deviation of the inflation rate from the inflation target. Such a functional-form assumption grants mathematical tractability and implies that both upward and downward deviations from the inflation target are symmetrically costly for the central bank.

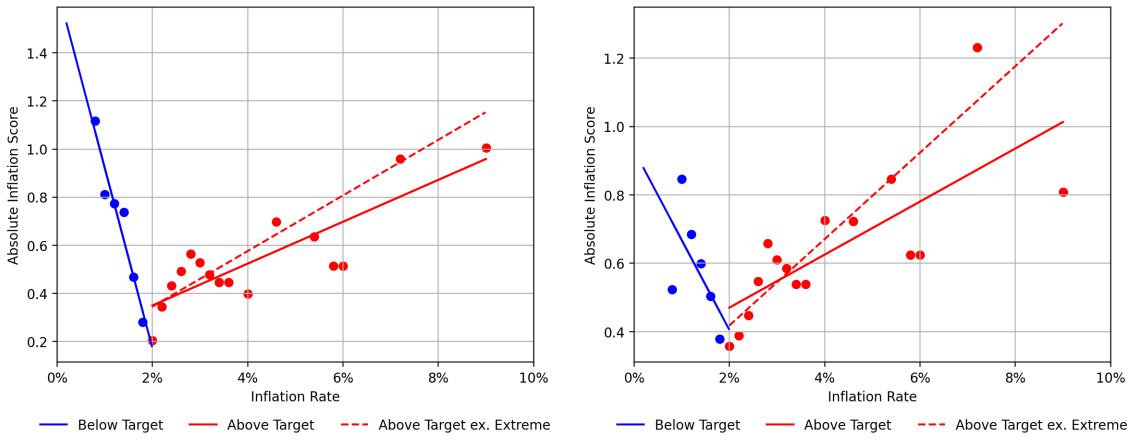
Haavio et al. (2026) provide a non-parametric empirical estimate of the ECB's loss function, computing the textual tone of ECB introductory statements using both lexicon-based and transformer-based techniques. Their methodology relies on the compelling assumption that, as inflation gets further away from the target and the loss function grows bigger, the central bank inflation-related textual tone gets more "concerned". In practice, this implies that the inflation-related textual tone rises in absolute terms, because the more inflation drops below (rises above) the target the more dovish (hawkish) the tone should become. Their main finding is that the ECB loss function is asymmetric with greater aversion to inflation above the inflation target. However, this result looks puzzling, given both the ECB mandate and the policy effort unfolded as policy rates hit the effective zero lower bound.

We exploit our textual analysis methodology to replicate the analysis in Haavio et al. (2026). We start by filtering press conference transcripts to identify inflation-related sentences according to two separate and incrementally narrower selection criteria. A first loose filter selects sentences containing inflation-related keywords (this is effectively the sample used to compute the inflation textual tone score in Figure 10); a second narrow filter selects sentences containing only inflation-related keywords (i.e. sentences containing other keywords for other topics are excluded). Each press conference is associated with the prevailing inflation context reading (see Figure 4) and we group press conferences in 0.2 percentage point wide buckets (e.g. $2.2 \leq \pi < 2.4$). For each bucket we compute the average inflation textual tone score in absolute terms.

Figure 12(a) displays the resulting inflation-tone data points based on the loose sentence filter. Fitting a linear regression to low-inflation (below 2%, in blue) and to high-inflation (above 2%, in red) data points, we note some degree of asymmetry with greater aversion to inflation below the inflation target. Lower aversion above the target might be explained by high-inflation data points after July 2021, when the ECB did not immediately respond to mounting inflation pressures to foster the post-pandemic recovery. Excluding these points above 4% context inflation, asymmetry decreases slightly (dashed red line).

Since sentences selected by the loose filter might contain other topics as well, we focus in Figure 12(b) on the subsample of sentences selected by the narrow filter containing inflation-related keywords only. Indeed, the central bank might acknowledge in the same sentence that, for example, due to a stagflationary energy shock prices are rising and economic activity is contracting. As a result, the textual tone score of the selected sentences might be biased. On this narrowed sample, asymmetry is highly reduced and the loss function is virtually symmetric after excluding high-inflation data points.

These results contrast with the findings of Haavio et al. (2026) and support the canonical theoretical assumption that the central bank loss function is symmetrical. We attribute differences to potentially three sources. (I) We believe that our identification of inflation-related sentences is narrower and more reliable: Haavio et al. (2026) exploit Latent Dirichlet Allocation to detect



(a) Loose filter

(b) Narrow filter

Figure 12: Non-parametric estimation of ECB's loss function. Dots represent the average absolute inflation textual tone score for each inflation bucket; lines represent a linear regression fit. The dashed line represents the fit on the high-inflation sample excluding high-inflation observations after July 2021 with context inflation above 4%. Panel (a) is based on the loose inflation-related sentence filter, panel (b) is based on the narrow inflation-related sentence filter.

topics, while we rely on a carefully handcrafted dictionary of keywords. (II) We evaluate the entire press conference transcript, instead of just the introductory statement. (III) We believe our textual tone computation methodology is more accurate thanks to more refined training and design.

7 Conclusion

In this paper we developed a pipeline based on state-of-the-art NLP techniques to assess the sentiment of monetary policy communication, focusing on ECB press conferences following rate decisions. We show that our methodology provides a meaningful summary of the central bank's assessment of the economic outlook, because our textual tone score index allows us to recover a theoretically valid impact of monetary policy on inflation. We exploit our pipeline to recover a non-parametric empirical estimate of the ECB loss function, validating the assumption of symmetry around the inflation target. Future research should exploit this technology to evaluate the impact of communication on financial markets.

References

Aruoba, S. B. and Drechsel, T. (2026). Identifying monetary policy shocks: A natural language approach. *American Economic Journal: Macroeconomics*.

Baranowski, P., Bennani, H., and Doryń, W. (2021). Do the ECB's introductory statements help predict monetary policy? Evidence from a tone analysis. *European Journal of Political Economy*, 66:101964.

Bennani, H. and Neuenkirch, M. (2017). The (home) bias of European central bankers: new evidence based on speeches. *Applied Economics*, 49(11):1114–1131.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–945.

Bobasu, A., Ciccarelli, M., and Notarpietro, A. (2025). Monetary policy transmission: A reference guide through ESCB models and empirical benchmarks. *ECB Occasional Paper Series*, (377).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gambacorta, L., Kwon, B., Park, T., Patelli, P., and Zhu, S. (2024). *CB-LMs: Language models for central banking*. Bank for International Settlements.

Geiger, F., Kanelis, D., Lieberknecht, P., and Sola, D. (2025). Monetary-Intelligent Language Agent (MILA). *Bundesbank Technical Paper*.

Haavio, M., Heikkinen, J., Jalasjoki, P., Kilponen, J., Paloviita, M., and Vänni, I. (2026). Reading between the lines—uncovering asymmetry in the central bank loss function. *Journal of Money, Credit and Banking*.

Hallmark, N., Galvao, A., and Powell, D. (2024). Bloomberg Economics' ECBspeak index. Technical report, Bloomberg.

Havránek, T. and Rusnak, M. (2013). Transmission lags of monetary policy: A meta-analysis. *International Journal of Central Banking*, 9(4):39–75.

Krippner, L. (2013). Measuring the stance of monetary policy in zero lower bound environments. *Economics Letters*, 118(1):135–138.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lupton, J., Atlas, D. M., and Weitzenfeld, D. (2023). Welcome to the machine: NLP of central bank speak. Technical report, J.P.Morgan.

Lupton, J. and Weitzenfeld, D. (2024). Listen up: Upgrading J.P. Morgan's central bank NLP machine. Technical report, J.P.Morgan.

Pfeifer, M. and Marohl, V. P. (2023). CentralBankRoBERTa: A fine-tuned large language model for central bank communications. *The Journal of Finance and Data Science*, 9:100114.

Picault, M., Pinter, J., and Renault, T. (2022). Media sentiment on monetary policy: Determinants and relevance for inflation expectations. *Journal of International Money and Finance*, 124:102626.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Sadaiyandi, J., Arumugam, P., Sangaiah, A. K., and Zhang, C. (2023). Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset. *Electronics*, 12(21):4423.

Shah, A., Paturi, S., and Chava, S. (2023). Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Walsh, C. E. (1995). Optimal contracts for central bankers. *The American Economic Review*, pages 150–167.

A Appendix

A.1 Transformer-based Models for Classification

Our NLP pipeline consists of two classification steps: relevance classification and sentiment classification. In the first step, sentences from ECB press conferences are classified according to whether they are relevant to monetary policy. In the second step, the relevant sentences are assigned a sentiment label reflecting their perceived dovish–hawkish tone.

Working with text is challenging, as machines cannot directly understand natural language. Therefore, text must be converted into numerical values that preserve its meaning and structure. These numerical representations are known as *embeddings*, which map words or sentences into vectors of numbers in a continuous space.

Transformer-based models such as BERT, RoBERTa, and GPT are designed to produce high-quality embeddings. During training, these models learn how to encode semantic meaning, capture word order, and model contextual dependencies between tokens. The embeddings produced by Transformer models are *contextual* rather than static, meaning that the representation of a word depends on its surrounding context. For example, the word *"bank"* conveys different meanings in the sentences *"He sat by the river bank"* and *"She deposited money in the bank"*. Transformer models can capture this difference by assigning different embeddings to the word in each sentence.

Given their strong performance in a wide range of downstream NLP tasks, we adopt RoBERTa models in this study. In this Appendix, we briefly describe the main components underlying our approach. Appendix A.2 introduces the Transformer architecture, Appendix A.3 discusses BERT, Appendix A.4 describes RoBERTa, and Appendix A.5 explains how Transformer models are fine-tuned for specific tasks.

A.2 The Transformer

We first briefly describe the Transformer architecture, as introduced by Vaswani et al. (2017). A Transformer-based model can have an *encoder-decoder*, *encoder-only* or *decoder-only* structure. The model proposed in the aforementioned work explains the canonical encoder-decoder model architecture, displayed in Figure 13. The block on the left is the *encoder*, whose outputs are fed to the *decoder*, the block on the right. Since RoBERTa is based on an *encoder-only* Transformer architecture, we focus our discussion on the encoder component.

An encoder block is composed of a stack of L identical layers, 12 for the `BERT-base` model and 24 for the `BERT-large` model. "Each layer has two sub-layers. The first is a *Multi-Head self-Attention mechanism* (MHA), and the second is a simple, position-wise *fully connected feed-forward network* $g(\cdot)$. We employ a *residual connection*[19] around each of the two sub-layers, followed by

[19] A *residual or skip connection* connects the output of a previous layer with the output of a future layer, skipping any transformation in between. In the context of a Transformer encoder layer, with the first residual connection, the input to the first sub-layer X connects to the output of the first sub-layer MHA(X), and with the second residual connection, the output of the first sub-layer, after applying layer normalization, connects to the output of the second sub-layer g(H).

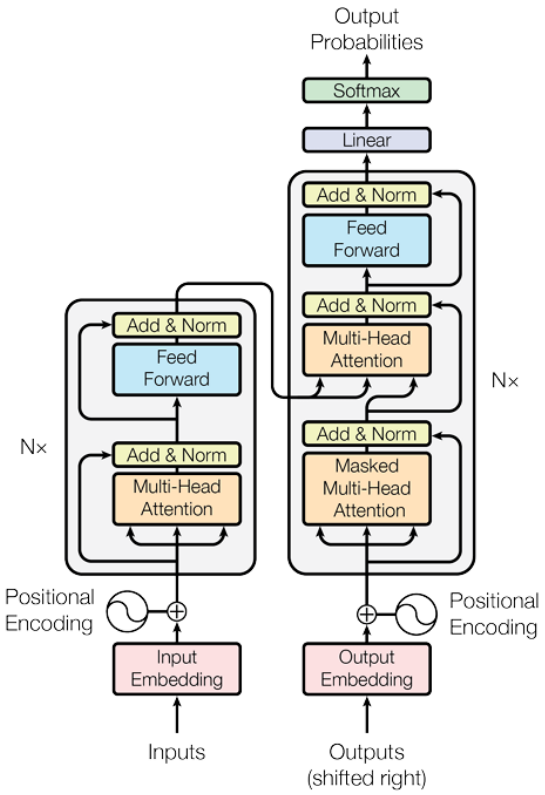


Figure 13: The Transformer - model architecture. Source: Vaswani et al. (2017).

layer normalization[20]" Vaswani et al. (2017), as shown in the equations below

$$\begin{aligned} \mathbf{H} &= \text{MHA}(\mathbf{X}) \\ \mathbf{H} &= \text{LayerNorm}(\mathbf{H} + \mathbf{X}) \\ \mathbf{F} &= g(\mathbf{H}) \\ \mathbf{H} &= \text{LayerNorm}(\mathbf{F} + \mathbf{H}). \end{aligned} \tag{15}$$

The core of the Transformer model is the *self-attention mechanism*, which allows the model to

[20] *Layer normalization* is a variant of batch normalization that normalizes the inputs across the features for each data sample rather than across the batch. First, we normalize the input features

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

where μ and σ^2 are respectively the mean and variance across the feature dimension, and ϵ is a small constant for numerical stability. Secondly, we apply the learned scale α and shift β parameters to $\hat{x}_i$ and obtain

$$\text{output}_i = \alpha \cdot \hat{x}_i + \beta$$

focus on different parts of the input sequence, mimicking the way human attention works. This mechanism works by computing *attention scores* between pairs of tokens $(\mathbf{x_i}, \mathbf{x_j})$ in the sequence. Specifically, the *scaled dot-product attention* function $\alpha(\cdot, \cdot)$ is used

$$\alpha(\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{\sqrt{d}} \mathbf{x_i}^\intercal \mathbf{x_j}, \tag{16}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are two d-dimensional vectors. Now, we can define the *self-attention layer* (for a single token i) as

$$\mathbf{h_i} = \sum_{j=1}^{n} \text{Softmax}_j \left(\frac{1}{\sqrt{d}} \mathbf{x_i}^\intercal \mathbf{x_j} \right) \mathbf{x_j}, \tag{17}$$

where the Softmax function ensures that the attention scores are normalized, converting them into probabilities, also called *attention weights*. This formulation of the self-attention layer lacks trainable parameters. Hence, we must introduce 3 *trainable matrices*, namely $\mathbf{W}_q \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_k \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$, such that we can define the following vectors

$$\mathbf{q_i} = \mathbf{W}_q^\intercal \cdot \mathbf{x_i}, \ \mathbf{k_j} = \mathbf{W}_k^\intercal \cdot \mathbf{x_j}, \ \mathbf{v_j} = \mathbf{W}_v^\intercal \cdot \mathbf{x_j}, \tag{18}$$

which we call *query*, *key* and *value*. We substitute vectors $\mathbf{q_i}, \mathbf{k_j}$ and $\mathbf{v_j}$ in equation (17) and obtain

$$\mathbf{h_i} = \sum_{j=1}^{n} \text{Softmax}_j \left(\frac{1}{\sqrt{d_k}} \mathbf{q_i}^\intercal \mathbf{k_j} \right) \mathbf{v_j}, \tag{19}$$

where the attention scores are calculated based on the dot product of query and key vectors, scaled by the square root of their dimension. Finally, we can rewrite the self-attention layer in vectorized form by stacking the n input vectors $\mathbf{x_i}$ into a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$.

$$\text{Attention Scores} = \frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}} \in \mathbb{R}^{n \times n} \tag{20}$$

$$\text{Attention Weights} = \text{Softmax}\left(\text{Attention Scores}\right) \in \mathbb{R}^{n \times n} \tag{21}$$

$$\mathbf{H} = \text{Attention Weights} \cdot \mathbf{V} \in \mathbb{R}^{n \times d_v}, \tag{22}$$

where $\mathbf{H}$ is the output of the self-attention layer, $\mathbf{Q} = \mathbf{X}\mathbf{W}_q \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} = \mathbf{X}\mathbf{W}_k \in \mathbb{R}^{n \times d_k}$ and $\mathbf{V} = \mathbf{X}\mathbf{W}_v \in \mathbb{R}^{n \times d_v}$.

To enhance the model's capability, Transformers employ *Multi-Head Attention* (MHA), which allows the model to capture various aspects of the input sequence simultaneously, e.g. different meanings of a word in different contexts. The MHA layer works by computing $t = 1, \ldots, h$ separate sets of keys, queries and values, with h being the number of *attention heads*

$$\mathbf{Q}_t = \mathbf{X}\mathbf{W}_{q,t}, \ \mathbf{K}_t = \mathbf{X}\mathbf{W}_{k,t}, \ \mathbf{V}_t = \mathbf{X}\mathbf{W}_{v,t} \tag{23}$$

$$\mathbf{H}_t = \text{Softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\intercal}{\sqrt{d_k}} \right) \mathbf{V}_t. \tag{24}$$

The outputs from all attention heads $\mathbf{H}_t$ are concatenated over the embedding dimension and projected again using the *trainable output matrix* $\mathbf{W}_o \in \mathbb{R}^{hd_v \times d_o}$, where d_o is the output size

$$\text{MHA}(\mathbf{X}) = [\mathbf{H}_1, \ldots, \mathbf{H}_h] \cdot \mathbf{W}_o \in \mathbb{R}^{n \times d_o}. \tag{25}$$

Figure 13 highlights another key component of Transformer models known as *positional encodings or embeddings*. These positional encodings, represented by the matrix $\mathbf{E}$, are either concatenated or summed with the token embeddings $\mathbf{Y}$ before being fed into the first Multi-Head Attention (MHA) layer. The purpose of positional encodings is to provide information about the position of each token within the input sequence, which allows the model to capture the order of tokens. The Transformer model uses sinusoidal functions to generate positional embeddings. The idea is to alternate sine and cosine functions of the same frequency w_j

$$w_j = \frac{1}{10000^{\frac{2j}{d}}}, \tag{26}$$

where d is the dimension of both the positional encoding and token embedding vectors. For position i and dimension j we get

$$[\mathbf{E}]_{i,2j} = \sin\left(i \cdot w_j\right), \; [\mathbf{E}]_{i,2j+1} = \cos\left(i \cdot w_j\right). \tag{27}$$

The positional embeddings are then added to the token embeddings to obtain the input matrix $\mathbf{X}$

$$[\mathbf{X}]_{i,:} = [\mathbf{Y}]_{i,:} + [\mathbf{E}]_{i,:}, \tag{28}$$

where each row $[\mathbf{E}]_{i,:}$ uniquely encodes the position of every element in the sequence. This combined representation incorporates the content of the token and its position in the sequence.

A.3 BERT

BERT (*Bidirectional Encoder Representation from Transformers*) is a multi-layer bidirectional Transformer encoder "designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" Devlin et al. (2018). The usage of BERT involves two main steps: *pre-training* and *fine-tuning*. As described in Section 4, our relevance and sentiment text classification require pre-trained Transformer-based models with BERT architecture. We adapted these models to our specific task through fine-tuning. In this Section, however, we will explain the pre-training procedure for BERT, which enables the model to learn general language representations from extensive text corpora. BERT is pre-trained via two unsupervised tasks: *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP).

MLM enables BERT to transition from unidirectional to bidirectional representations. This method involves *masking* certain tokens in the input text and then training the model to predict these masked tokens based on the context provided by the remaining words in the sentence. Unlike traditional language models that predict the next word in a sequence (which is unidirectional), Masked LM allows BERT to utilize context from both directions (left and right) around a masked token. In the BERT paper, approximately 15% of the tokens in each input sequence are masked. "If the i-th token is chosen, we replace the i-th token with (1) the `[MASK]` token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time[21]. Then, T_i[22] will

[21] To bias the representation towards the actual observed word.

[22] The final hidden representation for the i-th input token.

be used to predict the original token with cross-entropy loss" Devlin et al. (2018). Suppose that the unlabeled sentence is "*i like this book*", and during the random masking procedure, we choose the 4-th token. The masking procedure can be illustrated with the following example:

- 80% of the time we replace token "*book*" with `[MASK]`.

$$\textit{"i like this book"} \longrightarrow \textit{"i like this [MASK]"}$$

- 10% of the time we replace token "*book*" with a random word, e.g. "*apple*".

$$\textit{"i like this book"} \longrightarrow \textit{"i like this apple"}$$

- 10% of the time we keep the 4-th token unchanged.

$$\textit{"i like this book"} \longrightarrow \textit{"i like this book"}$$

On the other hand, NSP helps BERT learn the relationships between sentences, a capability not captured by language modeling alone but essential for tasks such as Question Answering (QA) and Natural Language Inference (NLI). In NSP, the two sentences are separated by a special token `[SEP]` and preceded by the `[CLS]` token used for classification tasks

Input: `[CLS]` Sentence A `[SEP]` Sentence B `[SEP]`.

For each sentence pair (Sentence A, Sentence B) (1) 50% of the time *Sentence B* is the actual sentence that follows *Sentence A* in the text (labeled as *"IsNext"*), (2) the remaining 50% of the time *Sentence B* is chosen at random from the corpus (labeled as *"NotNext"*). For instance,

(1) Sentence B: *"he bought a gallon `[MASK]` milk"* is the true sentence that follows Sentence A: *"the man went to `[MASK]` store"*. The label is *"IsNext"*.

(2) Sentence B: *"my dog `[MASK]` barking"* is a random sentence from the corpus. The label is *"NotNext"*.

A.4 RoBERTa

RoBERTa (*Robustly optimized BERT approach*) is an improved version of BERT, sharing the same architecture but trained with several improvements:

(1) *Extended training.* RoBERTa was trained with much larger mini-batches (8,000 sequences vs. BERT's 256), for up to 500K steps, and over more data (160GB, about 10 times more than for BERT).

(2) *Removal of the Next Sentence Prediction (NSP) objective.* RoBERTa was trained exclusively on the Masked Language Modeling (MLM) objective, as its authors found that eliminating NSP either matched or slightly improved performance on downstream tasks.

(3) *Training on longer sequences.* RoBERTa was trained with sequences of up to 512 tokens, while BERT was trained mostly on shorter sequences of 128 tokens.

(4) *Dynamic masking.* BERT uses static masking: masks are generated once during data preprocessing (in the replication of Liu et al. (2019), the training data are duplicated 10 times so that each sequence is masked in 10 different ways over 40 epochs). RoBERTa instead uses dynamic masking, generating a new masking pattern every time a sequence is fed to the model.

We use both the base and large variants of RoBERTa, whose architectural characteristics are reported in Table 3. The base model has 12 layers and fewer parameters, while the large model has higher capacity and is therefore better suited for more complex tasks such as sentiment classification compared to relevance classification.

Model	Layers	Attention heads	Hidden size	Number of Parameters	Vocabulary size
RoBERTa-base	12	12	768	∼125M	50,265
RoBERTa-large	24	16	1024	∼355M	50,265

Table 3: Architecture characteristics of the RoBERTa models. Hidden size is the dimension of the token representations; both models accept input sequences of up to 512 tokens. Number of parameters refers to the embedding and Transformer model parameters.

A.5 Fine-Tuning Pre-Trained Transformer-based Models

Training a deep neural network typically requires a large amount of data. When such extensive datasets are unavailable, we can leverage pre-trained models (e.g. Transformers) and adapt them for our specific task. This approach is known as *Transfer Learning*. As the name suggests, rather than starting from scratch, Transfer Learning involves transferring the knowledge acquired by a large, pre-trained model and applying it to a different but related problem. As shown in Figure 14, the process begins with a pre-trained model $\mathrm{M}_{\mathcal{D}}$ (e.g. RoBERTa), which has already been trained on a large general-purpose dataset $\mathcal{D}$ (e.g. as described in Section A.3). We can then adapt this model to a new task t using a task-specific dataset $\mathcal{D}_t$, creating a new model $\mathrm{M}_{\mathcal{D}_t}$. This process can be repeated multiple times across different tasks and datasets.

The simplest and most common way to perform Transfer Learning is through *Fine-Tuning*. Fine-tuning adapts a pre-trained model to a new task by *re-training some of its weights* using a task-specific dataset—in our case, the corpus of ECB press conference texts—to improve its performance on that specific task. Typically, the final layer of the model is replaced with a new layer that suits the new task. For example, we might add a *classification head* with randomly initialized weights that generate probabilities for the target classes.

A Transformer model is composed of multiple layers, each capturing different levels of information. The earlier layers focus on extracting general, low-level features, while the later layers extract more task-specific, high-level features. Thus, during fine-tuning, it is common to "freeze" the earlier layers, leaving only the later layers trainable (as shown in Figure 16). This allows the model to adapt to the new task without losing the general knowledge it gained during pre-training. Fine-tuning is usually done with a low learning rate to avoid wiping out all the pre-trained knowledge[23].

[23]This phenomenon is known as *catastrophic forgetting*.

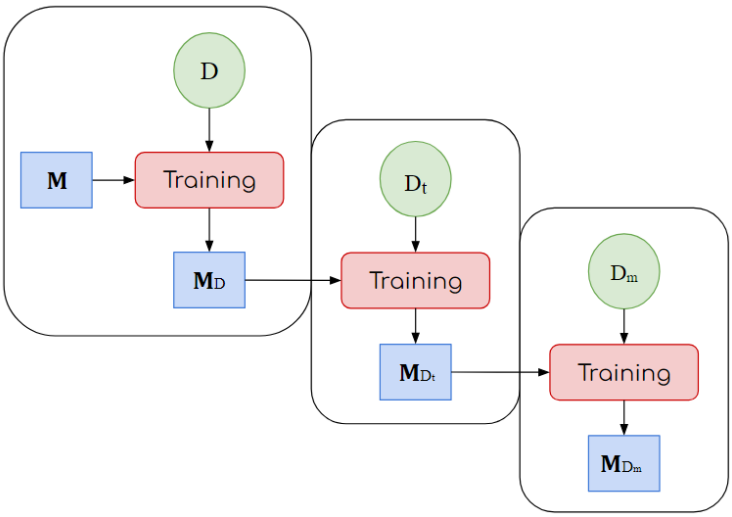


Figure 14: Transfer Learning.

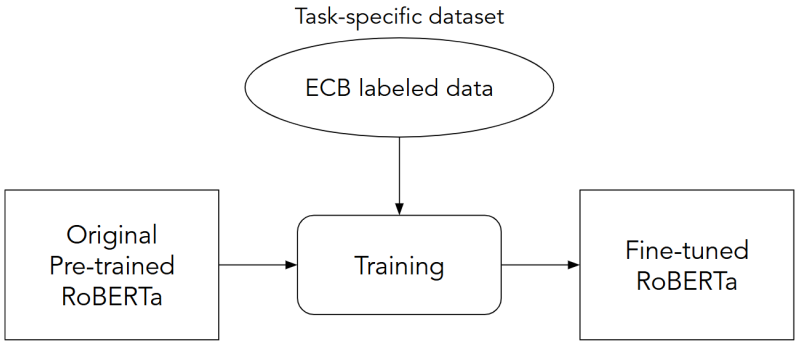


Figure 15: Adapting a Transformer-based model to the ECB dataset through fine-tuning.

Building on this framework, Figure 17 illustrates how RoBERTa is adapted for the sentence-level sentiment classification task in our setting. Each context-augmented sentence is fed into the pre-trained RoBERTa model, which is subsequently fine-tuned to specialize the general-purpose language representations for monetary policy sentiment classification. During fine-tuning, only a subset of the model parameters is updated, typically corresponding to the last $L - l$ Transformer layers, while the first l layers are kept frozen to preserve the general linguistic knowledge learned during pre-training. The model processes the input sentence by generating contextualized embeddings for each token, where the representation of each token depends on the full sentence context. The hidden representation of the special classification token $\langle s \rangle$[24] from the final Transformer layer is used as a fixed-length sentence representation. To perform classification, a task-specific five-class *classification head* is added on top of the model. This head consists of a dense layer with a non-

[24]This token signals the start of the sentence.

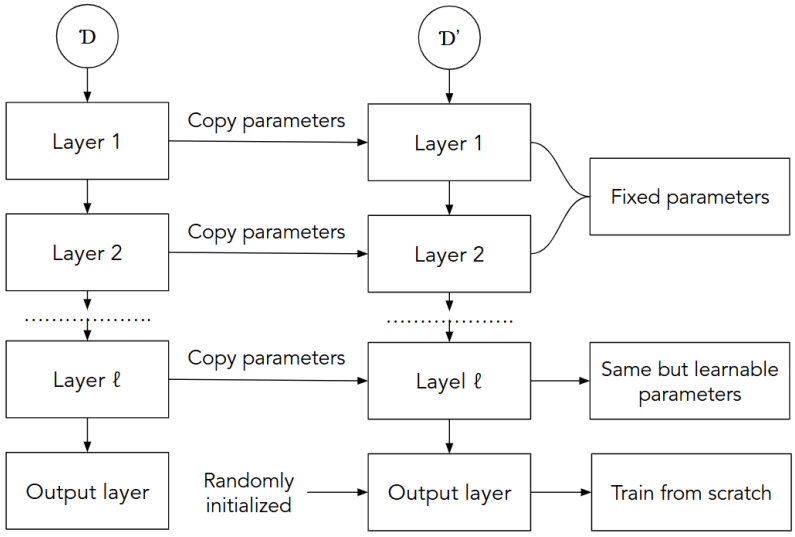


Figure 16: Fine-Tuning.

linear activation, followed by a dropout layer and an output projection layer that maps the sentence representation to class logits[25]. The parameters of the classification head are randomly initialized and trained from scratch during fine-tuning. The head outputs a probability distribution over the five sentiment classes, and the class with the highest predicted probability is selected as the final sentiment label for the sentence.

[25]Raw, un-normalized scores output by the model before being transformed into a probability distribution via the softmax function.

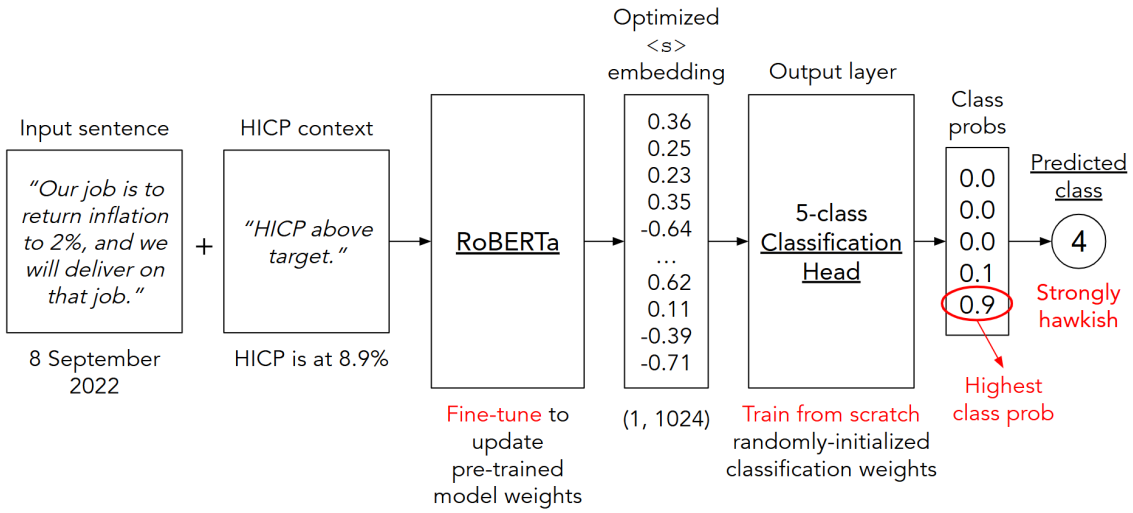


Figure 17: Fine-tuning RoBERTa for classification: an example.

B First-pass Relevance Filter

This Appendix lists the keywords used to perform the first-pass relevance filter of Section 4.1.

Category	Terms
Connectors	accordingly, although, and, at the same time, but, consequently, furthermore, hence, however, in addition, indeed, instead, moreover, nevertheless, nonetheless, on the other hand, rather, subsequently, therefore, though, yet
Common	celebrate, further information, good afternoon, happy new year, i will now outline, i will now report, ladies and gentlemen, let me now, merry christmas, please, press release, publish, scheduled, the governing council today decided, to be held, we decided to, website, will be released
Questions	a question on, add-on question, ask you a question, could you, do you have, do you think, does that, first question is on, follow-up question, i am wondering, i have a question, i have a quick question, I have a two-pronged question, i have got a quick question, i have one question, i have three questions, i have two questions, i just wanted to ask, i want to ask, i wanted to ask, i wonder whether, i would like to ask, i've got a question, i've got two questions, if i may, if you please, in your opinion, my first question, my question, my second question, my third question, one last question, one question is, question:, second question is on, that is the first question, the first one is on, the first question is, the first question was, the next question is, the next question regards, the other question is, the other question regards, the question is, the second one is on, the second question is, the second question was, the third question is, the third question was, what do you, what is your, what would, would you, you have mentioned, you mentioned, you said, you were saying, your last press conference, your press conference
Economic	accommodation, accommodative, adverse factor, anchor, APP, balance sheet, bank, bank rate, challenge, consumption, credit, crisis, CSPP, cut, cycle, cyclical, data, decelerate, decline, decrease, deficit, deflation, demand, deposit, disinflation, disrupt, dove, downward, drop, ease, easing, economic, economic activity, economy, employment, exchange rate, expand, export, fall, fell, financial, financial sector, financing, fiscal, forecast, forward guidance, fund rate, GDP, good news, good place, grow, growth, hawk, headline, HICP, high, hike, hold, improve, increase, inflation, inflation expectation, interest rate, inventor, investment, job market, job retention, labor, labour, lending, liquid, liquidity, loan, low, market, monetary, monetary policies, monetary policy, monetary policy transmission, negative, non-accelerating, outlook, pause, pausing, PEPP, pessimistic, PMI, pmi indicator, portfolio, positive, pressure, price, productivity, projection, QE, QT, quantitative easing, quantitative tightening, raise, rapid, rate, rate path, real sector, rebound, recovery, reduce, refinancing, reinvest, repay, required, reserve, resilient, restrictive, rise, rising, risk, savings, services, slack, slow, sluggish, stable, stagflation, stimulus, strength, strong, subdue, supply chain, tariff, tighten, trade, uncertainty, unemployment, upward, urgent, wage, weak, well positioned, well-positioned, zero lower bound

Table 4: Economics-related keywords used for relevance filtering in Step 1.

C Data Labeling Examples

All sentences in the dataset were independently annotated by two annotators. Each annotator assigned one of six labels to every sentence, reflecting the monetary policy stance indicated. In cases of disagreement, a final label was jointly agreed upon. The labels are defined as follows:

- *Label -2 (Strongly Dovish):* Sentences that explicitly indicate a monetary policy action aimed at stimulating the economy and/or raising inflation, such as lowering interest rates or expanding the central bank's balance sheet. These statements reflect a strongly accommodative policy stance with a clear intention to act.
- *Label -1 (Dovish):* Sentences that describe weak economic conditions, such as slow growth, high unemployment, or inflation below target. These reflect negative economic considerations but do not indicate a concrete policy action.
- *Label 0 (Neutral):* Sentences that are neither dovish nor hawkish, providing general information without suggesting a direction for monetary policy.
- *Label 1 (Hawkish):* Sentences that highlight strong economic conditions, rising inflation, or low unemployment. These reflect positive economic considerations but do not explicitly signal a tightening action.
- *Label 2 (Strongly Hawkish):* Sentences that explicitly indicate a tightening of monetary policy, such as raising interest rates, in response to high inflation or an economy growing faster than desired. These statements reflect a strongly restrictive policy stance with a clear intention to act.
- *Label 3 (Irrelevant):* Sentences that do not pertain to monetary policy and are included for training the relevance classification model.

The following tables contain 10 sentence labeling examples per category. Each sentence is accompanied by a *context* label indicating whether inflation was *below target*, *at target*, or *above target* at the time of the communication. This contextual information is particularly useful when annotating ambiguous sentences, as the interpretation of the tone (dovish, hawkish, neutral) can depend on the prevailing inflationary environment.

Table 5: Strongly dovish examples.

Sentence	**Date**	**Context**
Overall, this will allow our monetary policy stance to remain accommodative.	07-Mar-2013	HICP below target
So, I will repeat what we stated, and this was unanimous: the Governing Council confirms that the key ECB interest rates, including the rate on the deposit facility, will remain at present or lower levels for an extended period of time.	01-Aug-2013	HICP below target

Continued on next page

Sentence	Date	Context
So the final decision is to stand ready with a programme that would help to strengthen our accommodative monetary policy stance, injecting money into the real economy; And in so doing, it would lead to a reconstruction of a market that has disappeared with the crisis, also for good reasons.	07-Aug-2014	HICP below target
We just want to be compliant with our mandate, which is to drive inflation back to below, but close to, 2% in the medium term.	22-Oct-2015	HICP below target
Regarding non-standard monetary policy measures, we confirm that our net asset purchases, at the current monthly pace of €30 billion, are intended to run until the end of September 2018, or beyond, if necessary, and in any case until the Governing Council sees a sustained adjustment in the path of inflation consistent with its inflation aim.	26-Apr-2018	HICP below target
We continue to expect them to remain at their present levels at least through the summer of 2019, and in any case for as long as necessary to ensure the continued sustained convergence of inflation to levels that are below, but close to, 2% over the medium term.	13-Sep-2018	HICP at target
The Governing Council also underlined the need for a highly accommodative stance of monetary policy for a prolonged period of time, as inflation rates, both realised and projected, have been persistently below levels that are in line with its aim; Accordingly, if the medium-term inflation outlook continues to fall short of our aim, the Governing Council is determined to act, in line with its commitment to symmetry in the inflation aim.	25-Jul-2019	HICP below target
It is clearly with the purpose of getting back to the pre-COVID inflation path that we use this tool as well as all the other tools of the package.	04-Jun-2020	HICP below target
We will continue our purchases under the pandemic emergency purchase programme (PEPP) with a total envelope of €1,350 billion.	29-Oct-2020	HICP below target
In this environment ample monetary stimulus remains essential to preserve favourable financing conditions over the pandemic period for all sectors of the economy.	21-Jan-2021	HICP below target

Table 6: Dovish examples.

Sentence	Date	Context
It is expected, therefore, that economic growth will also remain subdued in the coming months.	05-Dec-2002	HICP above target

Continued on next page

Sentence	Date	Context
In the Governing Council's view, the risks surrounding this broadly favourable outlook for economic growth over the coming years lie mainly on the downside.	11-Jan-2007	HICP above target
We observed a worsening, both in hard data and survey data – worsening Purchasing Managers' Indices (PMIs), especially in manufacturing, in new orders growth, etc.	03-Nov-2011	HICP above target
To a large extent, subdued loan growth reflects the current cyclical situation, heightened risk aversion and the ongoing adjustment in the balance sheets of households and enterprises, all of which weigh on credit demand.	02-Aug-2012	HICP above target
In relation to demand, the economy is indeed weak, and it is weak in its domestic components: consumption and, especially, fixed investment.	02-May-2013	HICP below target
Compared with the March 2020 ECB staff macroeconomic projections, the outlook for HICP inflation has been revised downwards by 0.8 percentage points in 2020, 0.6 percentage points in 2021 and 0.3 percentage points in 2022.	04-Jun-2020	HICP below target
Underlying inflation has also eased, reinforcing the signs that price pressures have weakened, and inflation expectations have declined at all horizons.	06-Jun-2024	HICP above target
Overall, credit dynamics remain weak.	18-Jul-2024	HICP above target
Borrowing costs have gone up in response to the fiscal news.	06-Mar-2025	HICP above target
Forward-looking indicators, such as the ECB's wage tracker and surveys on wage expectations, point to slower wage growth over the remainder of the year and the first half of 2026.	30-Oct-2025	HICP at target

Table 7: Neutral examples.

Sentence	Date	Context
While recent oil price developments imply some short-term upward pressure on inflation rates, the broad picture of annual HICP inflation hovering around 2% during the remainder of this year remains unchanged.	04-Sep-2003	HICP above target
We will continue to carefully monitor all developments which might affect our assessment.	02-Oct-2003	HICP above target
We are looking very carefully at all survey data and all the other information we extract from the financial markets.	08-Nov-2007	HICP above target
We believe that the current monetary policy stance will contribute to achieving this objective.	08-May-2008	HICP above target
Risks to the outlook for inflation are broadly balanced.	02-Jul-2009	HICP below target
Inflation expectations remain firmly anchored in line with our aim of keeping inflation rates below, but close to, 2% over the medium term.	13-Jan-2011	HICP above target

Continued on next page

Sentence	Date	Context
Incoming information and analysis have further underpinned our previous assessment.	05-Sep-2013	HICP below target
The Governing Council sees both upside and downside risks to the outlook for price developments as limited and broadly balanced over the medium term.	03-Apr-2014	HICP below target
We say that the exchange rate is not a policy target, and it's not a policy target; However, we say the exchange rate is also important for growth and price stability.	02-Jun-2016	HICP below target
We will follow a data-dependent and meeting-by-meeting approach to determining the appropriate monetary policy stance.	12-Dec-2024	HICP above target

Table 8: Hawkish examples.

Sentence	Date	Context
As I mentioned earlier, inflation rates may remain above 2% for longer than was expected just a few months ago, owing to unexpected developments in energy prices and the past decline in the euro.	02-Nov-2000	HICP above target
As a result of the persistently strong growth in M3 over the past few years, there remains substantially more liquidity in the euro area than is needed to finance non-inflationary economic growth.	03-Feb-2005	HICP above target
Global growth remains solid, providing a favourable environment for euro area exports.	03-Feb-2005	HICP above target
The monetary analysis confirms the prevailing upside risks to price stability at medium to longer horizons.	12-Apr-2007	HICP above target
Loans to non-financial corporations remain very dynamic and we are at the level of 12.6% of growth over the last twelve months.	05-Jul-2007	HICP above target
The latest information confirms that economic activity in the euro area continued to expand around the turn of the year.	04-Feb-2010	HICP below target
We continue to see upward pressure on overall inflation, mainly owing to energy and commodity prices.	05-May-2011	HICP above target
A gradual recovery should start later in 2013 as our accommodative monetary policy stance and significant improvement in financial market confidence work their way through to private domestic expenditure, and a strengthening of foreign demand should support export growth.	06-Dec-2012	HICP at target
Wages are still rising at an elevated pace, making up for the past inflation surge.	06-Jun-2024	HICP above target
The robust labour market, rising real incomes and solid private sector balance sheets continue to support consumption.	24-Jul-2025	HICP below target

Table 9: Strongly hawkish examples.

Sentence	**Date**	**Context**
On the basis of our assessment, and by acting in a firm and timely manner, we will ensure that risks to price stability over the medium term do not materialise and that medium-term inflation expectations remain firmly anchored in line with price stability.	06-Sep-2007	HICP above target
Monetary policy will do all that is needed to maintain price stability in the euro area over the medium term.	02-Sep-2010	HICP above target
We remain strongly determined to secure a firm anchoring of inflation expectations in the euro area in line with our aim of maintaining inflation rates below, but close to, 2% over the medium term.	09-Jun-2011	HICP above target
We anticipate that, after September 2018, subject to incoming data confirming our medium-term inflation outlook, we will reduce the monthly pace of the net asset purchases to €15 billion until the end of December 2018 and then end net purchases.	26-Jul-2018	HICP above target
We are on our path to exit negative interest rates soon.	09-Jun-2022	HICP above target
As a result of this assessment, the Governing Council concluded that those conditions have been satisfied; Accordingly, and in line with our policy sequencing, we intend to raise the key ECB interest rates by 25 basis points at our July monetary policy meeting.	09-Jun-2022	HICP above target
It takes frontloading, it will take further hikes in the next several meetings of a magnitude and at a pace that will be determined meeting by meeting and on the basis of the data that we receive; But we are certainly heading there because that is where we believe that we will actually deliver on our 2% medium-term target as indicated in our strategy.	08-Sep-2022	HICP above target
We also decided that this was not an isolated decision, but that we would raise interest rates further.	08-Sep-2022	HICP above target
We stand ready to adjust all of our instruments within our mandate to ensure that inflation returns to our medium-term inflation target.	27-Oct-2022	HICP above target
Unless there was a material change to our baseline, we will continue to hike at our next meeting.	15-Jun-2023	HICP above target