SCHOOL OF ECONOMICS AND MANAGEMENT

Department of Economics
DABN01: Master Essay
June 2023

Do central bank speeches help predict monetary policy?

Evidence from the Riksbank using transfer learning technique

Author:
Quang Vo

Supervisors:
Rani Basna
Michal Kos

Contents

List of Figures

List of Tables

Introduction

For a long time, central banks used to be opaque institutions. Central bankers believed that they should say as little as possible. "Never explain, never excuse" (said Montagu Norman, the Governor of the Bank of England from 1921 to 1944) may be the best summarization of the conventional wisdom among central bankers at that time.

This view started to change in the early 1990s with the adaption of inflation targeting (initiated by New Zealand, followed by many other central banks) and a growing consensus about the importance of managing expectations in monetary policy. The motivations for higher transparency come from empirical results. Geraats (2009) finds that countries with higher transparency have experienced lower inflation. Also, Blinder et al. (2008) shows that suitable communications can move financial markets and potentially help central banks to achieve their macroeconomic objectives. The role of communication become even larger during the financial crisis of 2007-2008 as policy rates (which were already near zero) were no longer an effective monetary tool. As a result, many central banks become more transparent and pay more attention to their communications by releasing more records about internal meetings and increasing the number of speeches given by members.

Among central bank communication tools, speeches usually convey lots of useful information about the future state of the economy as they focus mainly on the overall trend rather than technical details. Given that most speeches are from economic experts, they can be used for predicting some economic indicators if they are processed by appropriate tools. The challenge lies in the unstructured nature of speeches as most of them are in the text form. Nowadays, with the development of Natural Language Processing tools as well as deep learning techniques, it is feasible to deal with the challenge.

In recent years, literature in this field blossom into two primary branches: predicting central bank policy decisions and predicting central bank communication impacts on financial markets. This paper contributes to the first branch and aims to answer the question: **How can we use central bank speeches to predict future policy deci-**

sions?.

The remaining structure of the paper is the following: In Chapter Literature Review, we summarize previous works analyzing central bank announcements. Chapter Methodology describes the data and the model employed in this paper. We do some exploratory data analysis and show our results in Chapter Result. In the last chapter, we summarize key findings and give recommendations for further research.

Literature Review

In this chapter, we review the literature analyzing central bank announcements. We find that most studies lie in two primary branches:

- Predicting future policy actions, and
- Examining the reaction of financial markets.

For the first branch, most studies focus on big central banks such as the U.S. Federal Reserve (Fed) and European Central Bank (ECB). The Natural Language Processing (NLP) techniques that these studies deploy are mainly dictionary-based.

As an illustration, Istrefi, Odendahl, and Sestieri (2023) uses topic modeling with a tone dictionary (consisting of 96 positive and 295 negative words usually used in the financial stability context) to process Fed officials' speeches. The authors aim to construct indicators to measure the intensity and tone of both Governors and Federal Reserve Board (FRB) presidents. The paper finds that a monetary policy accommodation is likely associated with a negative tone or a high number of speeches discussing financial conditions, financial stability, and regulation.

Baranowski, Bennani, and Doryń (2021) and Hubert and Labondance (2021) are also interested in the tone of central bank communications, but use policy statements instead of speeches. Baranowski, Bennani, and Doryń (2021) uses the bag-of-words approach to quantify ECB's introductory statements and shows that a tone shock can be used to predict future ECB policy decisions. On the other hand, Hubert and Labondance (2021) uses a negative-positive dictionary to quantify Fed and ECB tone and finds that the tone can be used to forecast future policy decisions and can be used to explain monetary surprises. Furthermore, the authors show that ECB tone can be used to predict its policy action three months in advance.

Some studies ignore the text part. Instead, they focus on variables that are easier to obtain and process. A typical example is Istrefi, Odendahl, and Sestieri (2022). The authors find a significant impact of speaking events (which are measured by a dummy variable equaling 1 if there is a speaking event of an ECB officer) on several dependent variables (Eonia rates, market-based inflation expectations, and sovereign bond rates). The authors conclude that communications outside of the regular meetings contain a monetary policy signal.

Apel, Grimaldi, and Hull 2019 are among very few researchers who use the deep transfer learning method. In their paper, the authors use this method to compare the usefulness of information between minutes and transcripts of the Federal Open Market Committee (FOMC). The authors claim that transcripts are more informative than minutes and a strong agreement should happen before a policy rate increase.

In this first branch, there are not many studies focusing on the Riksbank (the Swedish central bank). To the best of our knowledge, we can only find one written by Andersson, Dillen, and Sellin (2006), which uses policy signaling from speeches (a dummy variable that takes the value 0,-1, 1 depending on whether the policy rate is kept the same, decrease or increase) to predict the term structure of interest rates and find that speeches can be used to predict the longer end of the term structure.

Next, we review the parallel branch which focused mainly on the reaction of financial markets.

Similar to the first branch, most studies use dictionary-based methods to process the text part. Petropoulos and Siakoulis (2021) use a dictionary of positive and negative words and a set of machine learning algorithms (Random Forests, Extreme Gradient Boosting, Support Vector Machines, and Deep Neural Networks) to build a sentiment index in forecasting future financial market (S&P500, VIX) turmoils. Anand et al. (2021) also use a dictionary of positive and negative words to find a strong association between the movement of stock indices in six leading European countries (except for France) with the tone of either ECB or the national bank or both. In the case of France, the author points out that the stock index violation is only significantly impacted by the national bank tone. With a similar dictionary-based method, Du et al. (2023) finds that written communication of the People's Bank of China can guide the market trend in the direction that the central bank wants.

There are also some studies ignoring the text part when analyzing central bank communication. Brubakk, Ellen, and Xu (2021) use published interest rates forecasted by the Norway and Sweden central banks to measure the impact of communications on the market yield curve on the announced date. The authors find that the key driver that moves the market rate is the forward guidance of these banks. Using a similar event-study approach, J. Liu et al. (2022) shows that communications can effectively influence the bond market.

This paper contributes mainly to the first branch. Specifically, our research question is "Can the Riksbank speeches be used to predict the bank's future policy decisions?". The contributions of this paper are: First, this study focuses solely on the Riksbank; and second, this paper uses deep transfer learning techniques instead of dictionary-based methods.

Methodology

In this chapter, we first describe the variables. Then, we discuss key steps in our model pipeline.

3.1 Variable measurements

There are two variables in our model:

- Central bank speeches, and
- Policy decision

3.1.1 Central bank speeches

There are some reasons explaining the importance of central bank speeches compared to other communication types. First, speeches are released more frequently than reports or meeting minutes. Second, due to the flexibility in format, information extracted from speeches can have a greater variety. Finally, speeches are usually about overall trends rather than technical details, which should give clues for future predictions.

In this paper, we ignore images and tables in the speeches as a typical speech usually doesn't contain such information. The text part in speeches is the only part we use.

Besides data from the Riksbank, we also use speeches from ECB for the training step. The reason is that there are not many speeches from the Riksbank (with a yearly average of around 20). ECB, on the other hand, releases around 80 speeches per year. Moreover, the high correlation in monetary decisions between these two institutions also motivates us to train our model with ECB data. Such a close relationship is understandable as the Swedish business cycle has been closely correlated with the EU economies (Söderström 2008) and there is a spillover effect from ECB monetary policy to the Riksbank (Ellen, Jansen, and Midthjell 2020).

Next, we describe the way we collect data. For the Riksbank speeches dataset, we scrape data from the Riksbank website with the help of the Beautiful Soup package (a Python package specializing in parsing HTML and XML documents). Due to the availability, we can only get data from January 2002 to April 2023. As for the ECB speeches dataset, we download it from the ECB website, filtering data in the same period as the Riksbank dataset.

3.1.2 Policy decision

One of the most important decisions of central banks is adjusting policy rates. A higher (or lower) policy rate leads to commercial banks in turn increasing (or decreasing) their borrowing rates, which can limit (or stimulate) economic activities. Given the importance of this decision, we use policy rate change as a proxy for the policy decision variable. In this paper, the dependent variable $\tilde{y}_t$ is the discrete transformation of the policy rate change y_t with:

$$y_t = r_{t+3} - r_t$$

where r_{t+3} is the policy rate at time t + 3 (three months after time t), and r_t is the policy rate at time t.

Our dependent variable $\tilde{y}_t$ is then defined as:

$$\tilde{y}_t = \begin{cases} 0, & \text{if } y_t < 0 \text{ (decrease)} \\ 1, & \text{if } y_t = 0 \text{ (no change)} \\ 2, & \text{if } y_t > 0 \text{ (increase)} \end{cases}$$

There are several reasons for the choice of the three-month time frame. First, the Executive Board of the Riksbank holds five meetings per year (before 2020, it held six meetings), which mean a policy rate change (if any) is likely to happen in two to three months. Second, in the study of Hubert and Labondance (2021), the authors find that ECB policy rate decisions can be predicted three months in advance. Given the assumption about the similarity of the ECB and the Riksbank policy decisions (as mentioned before), we believe a three-month period is a suitable choice.

Next, we describe how we collect policy rate datasets. As mentioned earlier, we need more data beyond what we can get from the Riksbank to train our model. Therefore, we also need two policy rate datasets: the Riksbank policy rate and the ECB policy rate.

For the Riksbank policy rate dataset, we download it from the Riksbank website. Due to the availability of speech data (which has the oldest month of January 2002), our policy rate dataset is also collected from 2002.

As for the ECB, there are more different types of policy rates that we can download:

– The deposit facility rate: the rate that banks may make an overnight deposit in the Europe banking system.

– Main refinancing operation rate: the rate for operations providing the bulk of liquidity to the Europe banking system.

– Marginal lending facility rate: the rate for banks to get overnight credit.

In this paper, we only use the deposit facility rate as the proxy for ECB's policy rate for simplicity reasons. Furthermore, we believe among the three interest rate types above, the deposit facility rate is more compatible with the Riksbank policy rate definition (the Riksbank policy rate is the rate that other commercial banks can borrow from and deposit in the Riksbank).

3.2 Model

In this section, we summarize the key steps that we implement. Our machine-learning workflow starts with the training phase. Then, the learned models are sent to the testing phase to evaluate the model performance on unseen data. The overall workflow is in Figure 3.1.

3.2.1 Input

There are two types of features in the model: text feature and past policy rate feature.

For the text feature, we include ECB speeches in the training phase, while we only test on the Riksbank speeches. This can help our model become more robust (as we have more data for training) but we can still have a fair view of how the model performs on the Riksbank dataset.

Past policy rate change is included for each speech. Inspired by Baranowski, Bennani, and Doryń (2021), we also include policy rate changes over the last three months and the last six months from the time a speech is released as input.

3.2.2 Topic modelling

Although most speeches from central bank officials are related to monetary policy, there are still some speeches unconnected to the policy rate change decisions (for example, the importance of central bank balance sheets, cyber attacks, etc.). To remove these unexpected data, we use the topic modeling technique to filter the speeches before further processing.

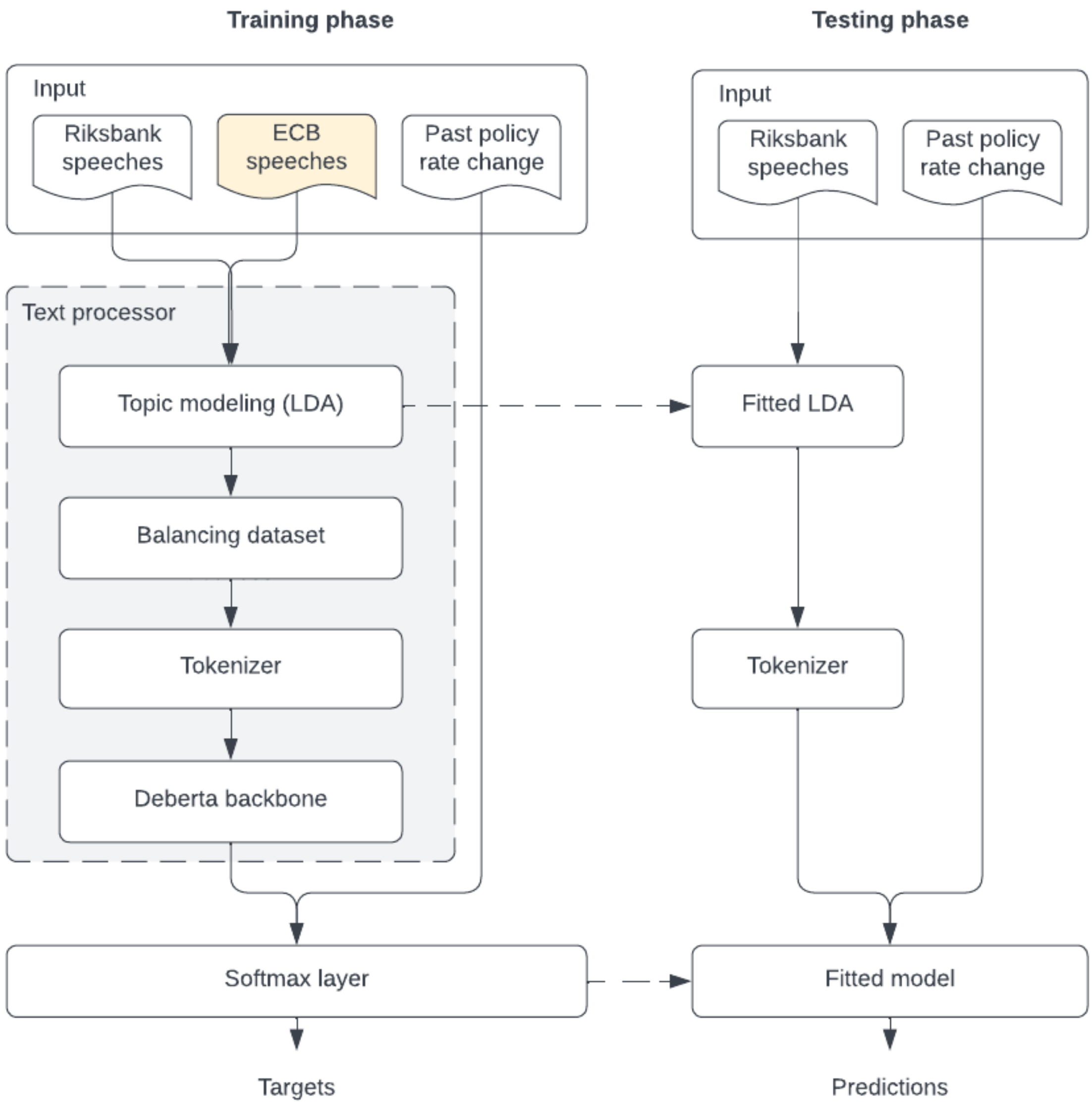

Figure 3.1: Model workflow

Topic modeling is a method to find some topics of documents even when we are not sure what these topics are. This paper uses the Latent Dirichlet Allocation (LDA) method proposed by Blei, Ng, and Jordan (2003), one of the most popular techniques in topic modeling. The basic idea of LDA is that documents are represented by a mixture of topics and each topic is defined as a distribution over words.

The input required for the algorithm are documents and a pre-defined number of clusters K. The output is a list of words for each cluster and their corresponding likelihood score.

For the choice of the number of topics, we make use of the result from Priola et al. (2022). Specifically, the authors find that around 80% of ECB speeches belong to one of six topics: Monetary Policy (36.7%), Financial Stability and Macroprudential Policy (16.7%), European Monetary Union Affairs (10.0%), Payments and Settlements (6.7%), Innovation (5.0%), and International Affairs (5.0%). We set K=6 to our model as the majority of training data are from ECB speeches.

3.2.3 Balancing dataset

Real-life datasets are usually exposed to the class imbalance problem. Highly imbalanced datasets can cause troubles for machine learning algorithms as these learners tend to be biased to major classes and in some extreme cases, minority classes can be totally ignored.

Japkowicz (2000) creates an artificial dataset to examine the effect of data imbalance on model performance. The author finds that highly complex problems see poor performance when dealing with highly imbalanced datasets, while simple and linear problems are unaffected. The problem in this paper is most likely complex and thus can be severely affected by the imbalance in the data.

According to a survey conducted by Johnson and Khoshgoftaar (2019), there are three main methods for dealing with imbalanced datasets:

- Data-level methods: modify the data distribution to lower the level of the imbalance. This can be done by under-sampling (reducing data in the majority classes) or over-sampling (increasing data in the minority classes).
- Algorithm-level methods: the learning process is adjusted to focus on the importance of the minority classes. This can be done by adding weights or penalties to the data for each class.
- Hybrid method: combine data-level and algorithm-level methods.

In this paper, because of the limitation in computational resources, we choose the Random Under-Sampling method. In this method, we randomly discard data from the majority classes to make the dataset less imbalanced.

3.2.4 DeBERTa tokenizer and backbone

In this paper, we use the DeBERTa model to process text features. The model belongs to the transformer-based family, which is first proposed by He, X. Liu, et al. (2021). DeBERTa is based on Google's BERT model released in 2018 and Facebook's RoBERTa model released in 2019.

The transformer model is suggested by Vaswani et al. (2017). The model is based mainly on the attention mechanism, which is designed to help the model learn the relationship between words, no matter where they appear in the sentence. Furthermore, multi-head attention and positional encoding (as can be seen in Figure 3.2) are also innovations at that time. More details on these concepts can be found in the Appendix.

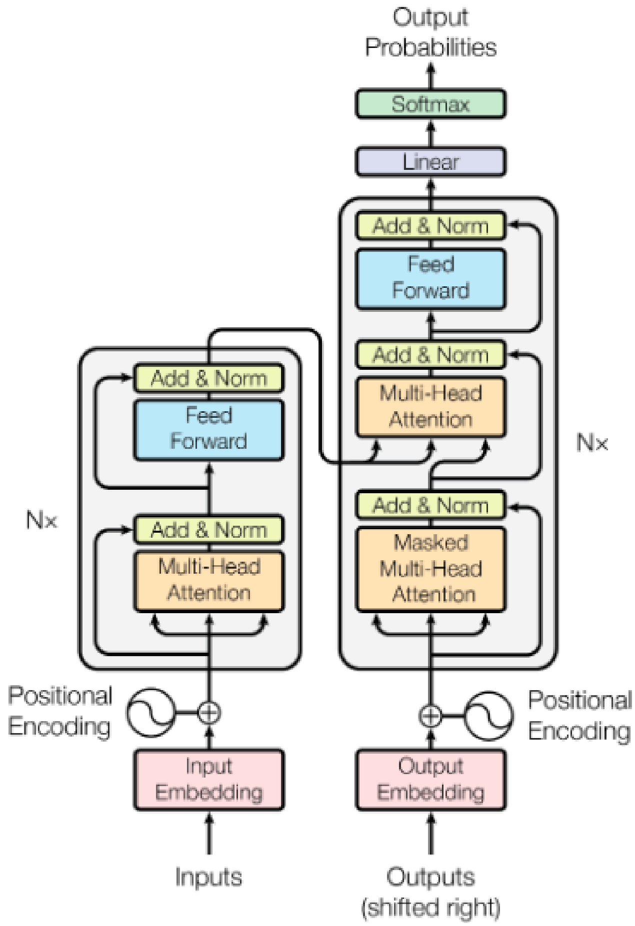

Figure 3.2: Transformer architecture (ibid.)

The key benefit when using a pre-trained model like DeBERTa is that we can get high performance even if we train the model on a small set of data. This benefit is achieved because the model is already trained on a much larger dataset. According to Goodfellow, Bengio, and Courville (2016), by using this transfer learning technique, we can improve the generalization of the models when we do not have much data. In this paper, we use the latest version of DeBERTa, proposed by He, Gao, and Chen (2023).

To deploy the DeBERTa model, we use tools developed by Hugging Face, a well-known platform for collaborating on machine learning projects. The platform supplies various tools and pre-trained models, which can be used for NLP tasks such as text classification, question-answering, and sentiment analysis, to name but a few.

There are three objects that we need to download from the platform: the tokenizer, the configuration, and the pre-trained weights.

- The tokenizer plays the role of an encoder. It takes sequences of words as input and converts them into an integer indices vector.

- The configuration contains the attribute of the model architecture such as the number of hidden layers, size of hidden layers, etc.
- The weights are parameter values that we can re-train or set as fixed.

After going through the DeBERTa model, the document will be vectorized. The attention mechanism (more details in Appendix) from the model helps to transform both the content and position information of words in the document into numerical-type vectors. These vectors then can be used for downstream tasks.

3.2.5 Softmax layer

In this step, we combine all the features together, then add the softmax layer on top of them to predict our dependent variable y.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

σ: softmax function

z: input vector

z_i: element ith of the input vector

The softmax layer converts a vector as K numbers into a probability distribution of K possible outputs. In this paper, we set K = 3 (as our dependent variable only has three possible values).

3.2.6 Result validation

The key performance metric that we use in this paper is accuracy, which calculates how often predictions equal ground truths.

$$Accuracy = \frac{\sum \text{true predictions}}{\sum \text{total predictions}}$$

To evaluate the model on unseen data, this paper uses the rolling cross-validation technique, which is one of the most common ways to split train and test sets when working with time series data.

The process is depicted in Figure 3.3. The horizontal axis represents the size of the data and the vertical axis represents the splits that we use to check the accuracy metric. We begin with splitting data into some splits containing a small subset of the data as a train

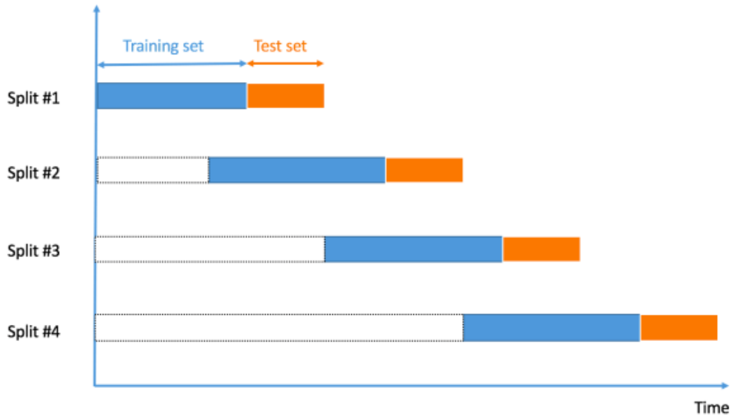

Image source: www.r-bloggers.com

Figure 3.3: Rolling time series cross-validation

set and some later data points as a test set. For each split, we calculate the accuracy for the test set. Our final performance metric is the accuracy average across splits.

Results

In this chapter, we first do some exploratory data analysis (EDA) to understand the patterns in the data, which can be used to confirm the hypothesis that we mention in Chapter Methodology. Next, we discuss our results and give our suggestions for future research. Our codes can be found here: `https://github.com/quang-vo-bi/central-bank-communication`.

4.1 Exploratory data analysis

4.1.1 Central Bank speeches

We start by describing some features of speeches from the Riksbank and ECB.

As we can see in Figure 4.1, there are not many speeches by the Riksbank (only 20-40 a year). That leads to the need to get more data to train our model. ECB speeches are a good candidate as the central bank releases around 80-120 speeches per year.

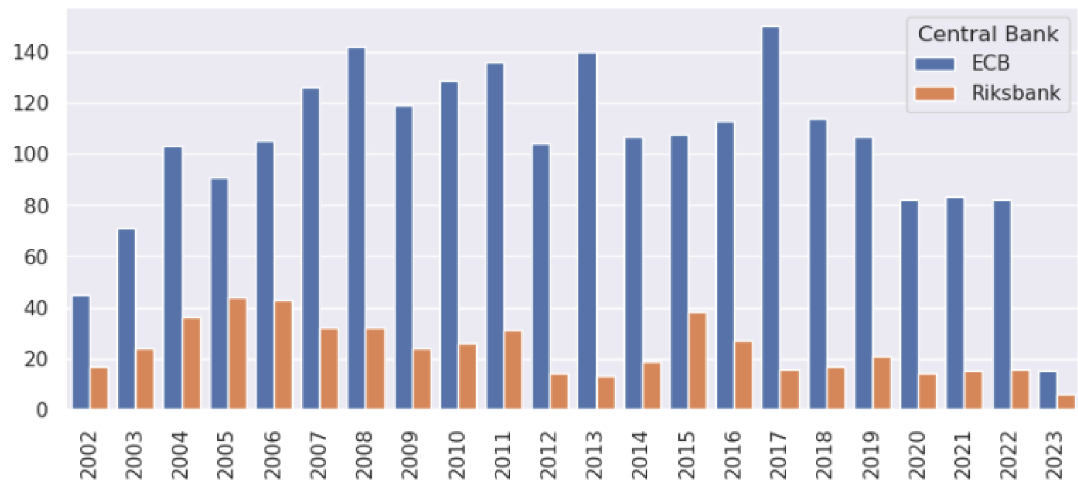

Figure 4.1: Number of speeches per year

Next, we want to see the distribution of speech length. The length of speeches is crucial because we need to set a suitable *max_length* hyperparameter for our DeBERTa model.

This hyperparameter determines the maximum length of the input that the model accepts. If an input has a length larger than *max_length*, it will be truncated.

As can be seen in Figure 4.2, the speeches are quite long. For the Riksbank, the second peak in the bimodal distribution indicates a significant portion of speeches with an average length of 4,000 words. As for ECB, the speeches' average length is also quite high at around 2,000 words.

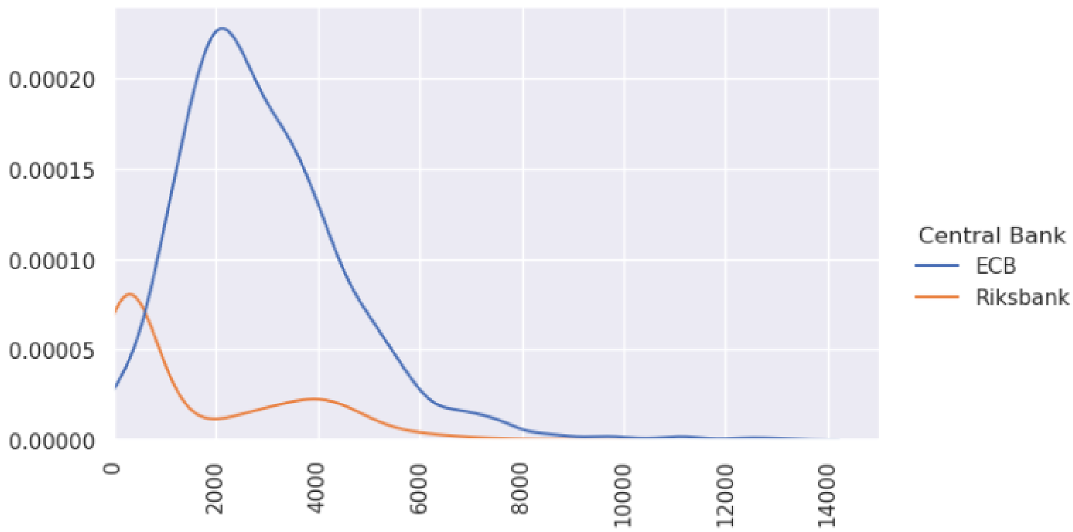

Figure 4.2: Distribution of speeches length

Due to the limitation in computational resources, we have to set the *max_length* parameter at 1024 or lower. However, too low *max_length* could lead to an underfitting problem as the model may not capture most of the important information.

To tackle this problem, we divide the speech into paragraphs. We then tokenize each paragraph. In this way, we can set the *max_length* hyperparameter as 128 but still can capture most of the important information as the average length of each paragraph is quite low as shown in Figure 4.3.

Another thing we consider when training our model is the variety of speech topics. We believe filtering out speeches that are not relevant to policy rate decisions can help to improve our model performance. The result from applying the LDA algorithm is shown in Figure 4.4.

As we can see, topics 2, 4, and 5 seem relevant to policy rate change decisions while topic 0 is too general and topics 1, and 3 can be seen as noise. In this paper, to pick relevant speeches, we only feed ones having topics containing one of these keywords **monetary, policy, rate, inflation, exchange, payment, stability, objective** to the DeBERTa model.

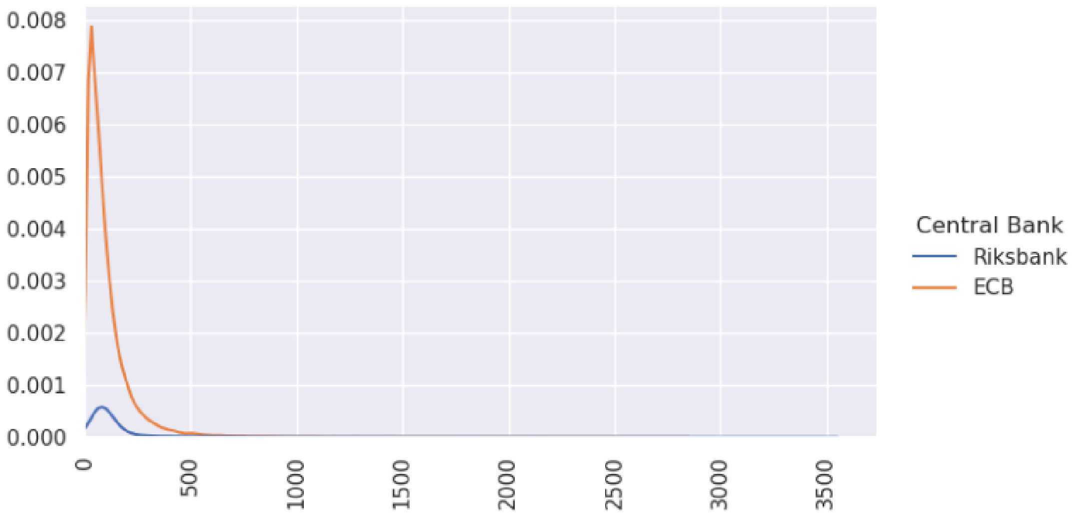

Figure 4.3: Distribution of paragraphs length

Figure 4.4: Topics of speeches using LDA

4.1.2 Policy rate

Next, we go through some characteristics of policy rates. Historical movements of the Riksbank and ECB policy rates are depicted in Figure 4.5 and their descriptive statistics are in Table 4.1

Central Bank	Min	Q25	Q50	Q75	Max	Earliest date	Latest date
ECB	-0.5	-0.4	0.25	1.0	3.25	2002-04-02	2023-04-01
Riksbank	-0.5	0.0	1.0	2.0	4.75	2002-04-02	2023-04-01

Table 4.1: Policy rate descriptive statistics

We can see that both rates have a high correlation in movement and similar quantile values, indicating highly correlated policy decisions between these two institutions. That confirms our reason for choosing ECB speeches as additional training data in Chapter Methodology.

Another characteristic of policy rates that needs to be considered is their cycle as we want both our training set and testing set to contain at least one cycle (each cycle is

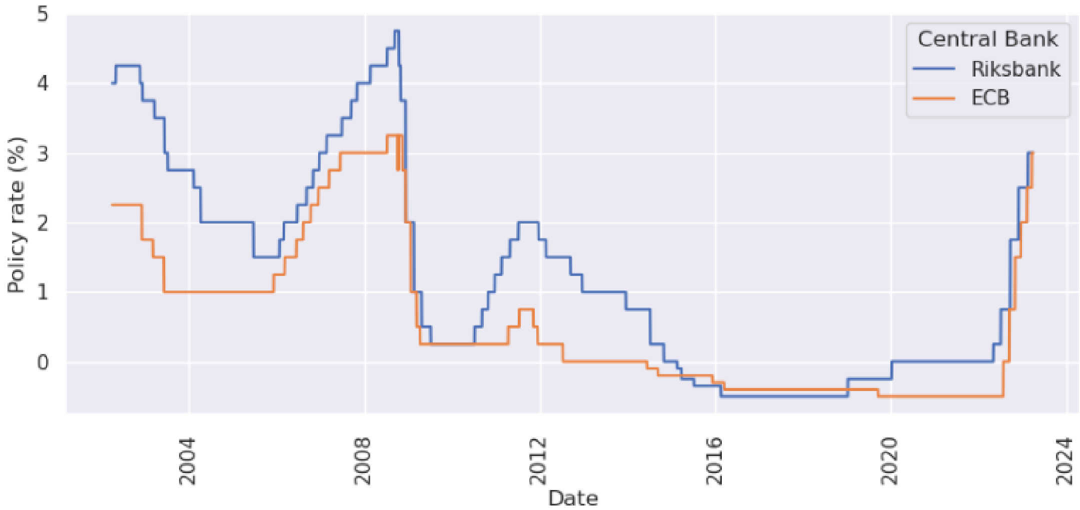

Figure 4.5: Historical policy rates

defined as seeing both a tightening and an easing period). Inspired by Cerqueira, Torgo, and Mozetic (2020), we deal with this challenge by using two splits to run our model. The first split has a training set from 2002 to 2014 and a testing set from 2015 to 2018, the second split has a training set from 2006 to 2018 and a testing set from 2019 to 2022. Using this rolling split technique helps us evaluate our model performance on a whole cycle.

The next thing that we want to check is the imbalance level of our dataset.

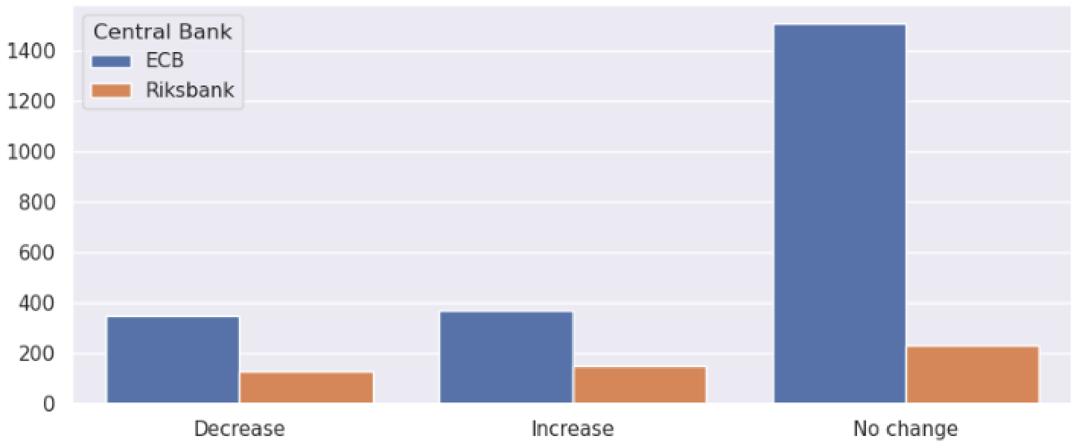

Figure 4.6: Class distribution

As expected, our dataset is highly imbalanced with *no change* decisions accounting for roughly 70%.(as seen in fig 4.6). This leads to the need for the undersampling technique that we depict in the last chapter.

4.2 Model performance and discussion

4.2.1 Base model

Figure 4.7 summarizes the accuracy of our base model (the model with configuration as depicted in Chapter Methodology) for each class. In this confusion matrix, the value at i-th row and j-th column means the proportion of the number of samples with the true label being i-th class and the number of samples with the predicted label being j-th class (0: Decrease policy rate, 1: No change, 2: Increase policy rate).

We achieve an overall accuracy of 71.4%. Specifically, our base model works pretty well in predicting an increase or no change in the policy rate, while the result is not good when predicting a decreasing decision.

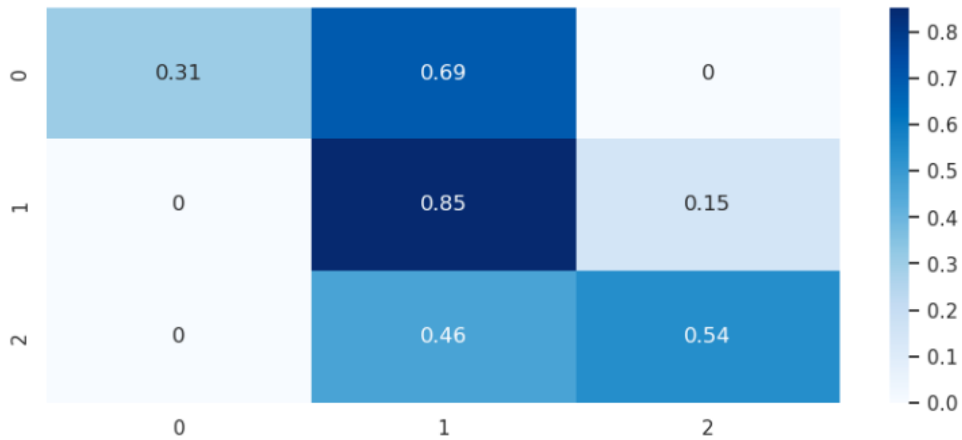

Figure 4.7: Accuracy results for the base model

For the poor performance of Class 0, we think of two possible reasons:

- Easing policy signals from ECB may differ from the Riksbank, and/or
- Speeches at the end of an easing period do not contain clear signals for the next policy actions.

4.2.2 Methodology Experiment

In this section, we do some experiments to see if we can remove any steps depicted in Chapter Methodology. The result of this section can give some suggestions for future research.

Table 4.2 summarizes our results for each experiment. The first column lists the name of the experiments. The next four columns are accuracy values for the whole dataset (Overall) and for each class (Decrease, No change, Increase).

There are three key findings:

Experiment	Overall	Decrease	No change	Increase
Without ECB corpora	30.6	51.4	0.0	100.0
Without topic modeling	69.1	13.5	85.4	50.5
Without under-sampling	72.4	0.0	100.0	28.1
Base model	71.4	30.6	85.3	53.7

Table 4.2: Overall accuracy and accuracy for each class (unit: %)

- Without ECB corpora, the model accuracy drops significantly to only 31%. This confirms our hypothesis that training models with more (relevant) data can improve model performance.
- Without the topic modeling step, our model still has a good performance of 69.1%. The base model is better in terms of predicting policy rate decrease decisions. However, the topic modeling step is time-consuming. If we do not need very high accuracy, we believe removing this step does not hurt the performance too much.
- Without the undersampling step, the model still gets an accuracy of 72.36%. However, the model simply predicts **no change** and totally ignores the **decrease** class. The high accuracy is only due to the high imbalance of the test set. If we compare the average accuracy of three classes (without considering the number of observations in each class), we can see that the accuracy of this experiment (42.7%) will be significantly lower than the base model (56.5%)

Conclusion

This paper addresses how to predict future policy actions of central banks using their speeches. Because the purpose of speeches is to inform audiences about the overall trend of the economy and possible future actions, we believe the question can be solved with appropriate methodology. The main challenge is how to quantify text data.

We suggest a model pipeline that starts with scraping data from the Riksbank and ECB. Then, we use modern machine learning techniques (LDA, DeBERTa) to process the text part in the data. Finally, we combine the outcome with past policy decisions to predict the final target. The result is promising with an overall accuracy of 71.4%.

Our paper has two main contributions. First, we focus solely on the Riksbank, which does not receive much attention from researchers. Second, this paper departs from earlier papers which mainly use the dictionary-based method and focus mainly on the tone of central banks rather than all possible information in their speeches. Promising results from this paper suggest that we can use DeBERTa instead.

There are two main limitations in this paper. First, because we do not own powerful computational resources, we may not train our model sufficiently (for instance, our neural network is only trained for two epochs due to the shortage of RAM). This may lead to the underfitting problem. Second, we cannot test the robustness of our methodology on other datasets (speeches from other central banks) as we do not have enough time and manpower.

In the future, it may be necessary to train with more corpora from other central banks. This could increase the generalization and the robustness of the model. It is also a promising idea to try other transformer-based models such as LongFormer. Finally, it is also important to extend our research by adding more features from other types of communications including meeting minutes and financial stability reports.

Bibliography

Anand, Abhinav et al. (2021). “Whose speeches impact European markets: ECB’s or the national central banks’?” In: *European Financial Management* 28.5, pp. 1413–1476.

Andersson, Malin, Hans Dillen, and Peter Sellin (2006). “Monetary policy signaling and movements in the term structure of interest rates”. In: *Journal of Monetary Economics* 53.8, pp. 1815–1855.

Apel, Mikael, Marianna Blix Grimaldi, and Isaiah Hull (2019). “How Much Information Do Monetary Policy Committees Disclose? Evidence from the FOMC’s Minutes and Transcripts”. In: *Sveriges Riksbank Working Paper Series* 381.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *A conference paper at ICLR 2015.*

Baranowski, Pawel, Hamza Bennani, and Wirginia Doryń (2021). “Do the ECB’s introductory statements help predict monetary policy? Evidence from a tone analysis”. In: *European Journal of Political Economy* 66.101964.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3, pp. 993–1002.

Blinder, Alan S. et al. (2008). “Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence”. In: *Journal of Economic Literature* 46.4, pp. 910–945.

Brubakk, Leif, Saskia ter Ellen, and Hong Xu (2021). “Central bank communication through interest rate projections”. In: *Journal of Banking and Finance* 124.106044.

Cerqueira, Vitor, Luis Torgo, and Nina Igor Mozetic (2020). “Evaluating time series forecasting models: An empirical study on performance estimation methods”. In: *Machine Language* 109.11, pp. 1997–2008.

Du, Xiuli et al. (2023). “Does central bank communication on financial stability work? ——An empirical study based on Chinese stock market”. In: *International Review of Economics Finance* 85, pp. 390–407.

Ellen, Saskia ter, Edvard Jansen, and Nina Larsson Midthjell (2020). “ECB Spillovers and domestic monetary policy effectiveness in small open economies”. In: *European Economic Review* 121.103338.

Geraats, Petra M. (2009). “Trends in Monetary Policy Transparency”. In: *International Finance* 12.2, pp. 235–268.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen (2023). “DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing”. In: *A conference paper at ICLR 2023*.

He, Pengcheng, Xiaodong Liu, et al. (2021). “DeBERTa: Decoding-enhanced BERT with Disentangled Attention”. In: *A conference paper at ICLR 2021*.

Hubert, Paul and Fabien Labondance (2021). “The signaling effects of central bank tone”. In: *European Economic Review* 133.103684.

Istrefi, Klodiana, Florens Odendahl, and Giulia Sestieri (2022). “ECB Communication and its Impact on Financial Markets”. In: *Banque de France Working Paper* 859.

— (2023). “Fed Communication on Financial Stability Concerns andMonetary Policy Decisions: Revelations from Speeches”. In: *Journal of Banking and Finance*. DOI: `https://doi.org/10.1016/j.jbankfin.2023.106820`.

Japkowicz, Nathalie (2000). “The Class Imbalance Problem: Significance and Strategies”. In: *In proceedings of the 2000 international conference on artificial intelligence (ICAI)*, pp. 111–117.

Johnson, Justin M. and Taghi M. Khoshgoftaar (2019). “Survey on deep learning with class imbalance”. In: *Journal of Big Data* 27.

Liu, Jianguo et al. (2022). “Can central bank communication effectively guide the monetary policy expectation of the public?” In: *International Review of Economics Finance* 75.101833.

Petropoulos, Anastasios and Vasilis Siakoulis (2021). “Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique”. In: *Central Bank Review* 21.4, pp. 141–153.

Priola, Maria Paola et al. (2022). “The informative value of central banks talks: a topic model application to sentiment analysis”. In: *Data Science in Finance and Economics* 2, pp. 181–204.

Söderström, Ulf (2008). “Re-Evaluating Swedish Membership in EMU: Evidence from an Estimated Model”. In: *Sveriges Riksbank Working Paper Series 227*.

Vaswani, Ashish et al. (2017). “Attention Is All You Need”. In: *Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)* 21.4, pp. 141–153.

Appendix

In this appendix, we summarize key ideas behind the DeBERTa model. Specifically, we review the definition of transformers and the evolution of the DeBERTa model.

A.1 Attention Mechanism

The attention mechanism was first proposed by Bahdanau, Cho, and Bengio (2015), which is used to help neural network models pay more attention to the most critical parts of the input when making a prediction.

In the paper, the authors illustrate this mechanism as a part of the Recurrent Neural Network (RNN) system. In the architecture, there are three key components: the alignment scores, the weights, and the context vector.

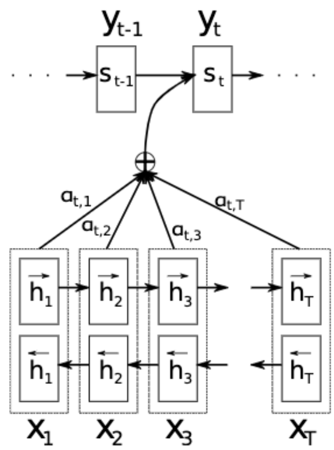

Figure A.1: Proposed model to illustrate attention mechanism

An alignment score e_{ij} measures how well inputs around position j match the output at position i.

$$e_{ij} = a(s_{i-1}, h_j)$$

a: a feedforward neural network that is trained at the same time as the RNN system
s_{i-1}: an RNN hidden state at time i - 1 h_j: the j-th annotation of the input sentence

The weight is the product when applying a softmax function to an alignment score.

This number reflects the importance of the annotation h_j to the hidden state s_{i-1}

$$\alpha_{ij} = softmax(e_{ij})$$

The context vector c_i is computed as a weighted sum of annotations h_i. c_i can be seen as *expected annotation* over all possible annotations.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

T_x: length of the input sequence

The next hidden state s_i then is computed as

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

As we can see, the hidden state s_i in the decoder that is built with this mechanism can decide which parts of the input sentence are important via c_i.

A.2 Transformer

The transformer architecture is proposed by Vaswani et al. (2017), which is based solely on the attention mechanism (as in Figure 3.2). There are three components in the model: scaled dot-product attention, multi-head attention, and positional encoding.

A.2.1 Scaled Dot-Product Attention

Here are some comparisons between the attention mechanism proposed by Bahdanau, Cho, and Bengio (2015) and the one proposed by Vaswani et al. (2017):

- There are three components in the scaled dot-product attention: the query Q (which is equivalent to s_{t-1} in the last section), the value V (equivalent to h_i) and the key K (in Bahdanau, Cho, and Bengio (2015), key and value are the same vector).
- The feedforward neural network a in Bahdanau, Cho, and Bengio (ibid.) is replaced by the dot product of Q and K scaled by $\sqrt{d_k}$ (d_k is the dimension of the key K)

Specifically, the attention (equivalent to context vector c_i) is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

A.2.2 Multi-head Attention

This mechanism is depicted in Figure A.2. When concatenating different scaled dot-product attention layers, the key benefit is that it allows the mechanism to create richer representations, which in turn can improve the model performance.

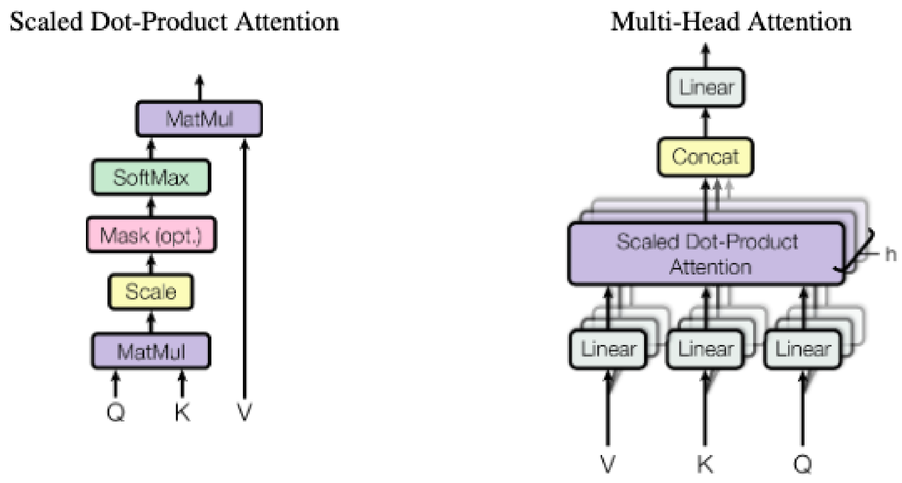

Figure A.2: Different versions of attention mechanisms

A.2.3 Postional Encoding

Because the model does not contain any recurrent procedures, the order information is ignored. To tackle this problem, Vaswani et al. (2017) suggests adding *positional encoding* to the input embeddings. These *positional encoding* are computed as follow:

$$PE_{pos,2i} = sin(\frac{pos}{10000^{2i/d_{model}}})$$
$$PE_{pos,2i+1} = cos(\frac{pos}{10000^{2i/d_{model}}})$$

where *pos* is the position of the word in the input, d_{model} is the size of the input embedding, and i refers to each of the individual dimensions of the input embedding.