MASTERARBEIT | MASTER'S THESIS

Titel | Title

Building Policy Analysis Models:

Disambiguation of Policy Reforms

verfasst von | submitted by

Nikoletta Jablonczay BA MA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Univ.-Prof. Dr. -Ing. Benjamin Roth B.Sc. M.Sc.

Abstract

This thesis aims to facilitate the transition from manual to automated disambiguation of policy reforms in the C3 Project's dataset, which consists of the Economist Intelligence Unit's Country Reports—an analysis of economic reforms in EU member states since the 1980s. The original C3 dataset contained only singleton mentions of policy reforms, limiting its use in automated data analysis. To address this limitation, the first significant contribution of this thesis was the manual expansion of the dataset to include multiple mentions of policy reforms. A baseline model using a simple clustering technique was developed as a reference point for evaluating a more advanced disambiguation system based on a pre-trained sentence embedding model. The system's performance was measured using the Link-Based Entity-Aware scorer and the Jaccard similarity index. Results showed an improvement in the F1 score from 0.395 in the validation set to 0.5507 in the test set, with the model correctly identifying 76 out of 138 predicted clusters. However, this improvement largely stemmed from accurately identifying singleton clusters, while the model struggled to handle multi-mention clusters. A Jaccard similarity score of 0.7346 indicated that while the model captured some thematic overlaps, it had difficulty with more nuanced distinctions. The study concludes that while NLP techniques can effectively extract policy mentions, the pre-trained model's effectiveness was limited without domain-specific training. Although the disambiguation system showed improvement over the all-singletons baseline, more advanced clustering strategies are needed to improve precision and recall in policy reform disambiguation.

Keywords: NLP, Disambiguation, Clustering, Embeddings, LEA Scorer, Policy Reforms

Kurzfassung

Diese Arbeit zielt darauf ab, den Übergang von der manuellen zur automatischen Disambiguierung politischer Reformen im Datensatz des C3-Projekts zu erleichtern, der aus den Länderberichten der Economist Intelligence Unit besteht. Es handelt sich um eine Analyse der Wirtschaftsreformen in den EU-Mitgliedstaaten seit den 1980er Jahren. Der ursprüngliche C3-Datensatz enthielt nur einzelne Erwähnungen von politischen Reformen, was seine Verwendung in der automatisierten Datenanalyse einschränkte. Um diese Einschränkung zu beheben, war der erste wichtige Beitrag dieser Arbeit die manuelle Erweiterung des Datensatzes, um Mehrfachnennungen von politischen Reformen aufzunehmen. Ein Basismodell, das eine einfache Clustering-Technik verwendet, wurde als Referenzpunkt für die Bewertung eines fortschrittlicheren Disambiguierungssystems entwickelt, das auf einem pre-trained Modell basiert. Die Leistung des Systems wurde mit dem Link-Based Entity-Aware Scorer und dem Jaccard Ähnlichkeitsindex gemessen. Die Ergebnisse zeigten eine Verbesserung des F1-Scores von 0,395 in der Validierungsmenge auf 0,5507 in der Testmenge, wobei das Modell 76 von 138 vorhergesagten Clustern korrekt identifizierte. Diese Verbesserung war jedoch größtenteils auf die genaue Identifizierung von Singleton-Clustern zurückzuführen, während das Modell Schwierigkeiten hatte, mit Clustern mit mehreren Erwähnungen umzugehen. Ein Jaccard-Ähnlichkeitswert von 0,7346 deutet darauf hin, dass das Modell zwar einige thematische Überschneidungen erfasste, aber Schwierigkeiten mit differenzierteren Unterscheidungen hatte. Die Studie kommt zu dem Schluss, dass NLP-Techniken zwar effektiv politische Erwähnungen extrahieren können, die Effektivität des vortrainierten Modells ohne domänenspezifisches Training jedoch begrenzt war. Obwohl das Disambiguierungssystem eine Verbesserung gegenüber der Baseline mit allen Singletons zeigte, sind fortschrittlichere Clustering-Strategien erforderlich, um die Genauigkeit und die Wiedererkennung bei der Disambiguierung von politischen Reformen zu verbessern.

Schlüsselwörter: NLP, Disambiguierung, Clustering, Embeddings, LEA Scorer, politische Reformen

Contents

List of Figures

List of Tables

List of Equations

1. Introduction

Text analyses in social and political science have a long history (Grimmer & Stewart, 2013). However, their application was limited due to the scarcity of data and the high costs associated with analysis (Glavaš et al., 2019). The advent of digital data sources and the integration of computation tools, particularly Natural Language Processing (NLP) techniques, marked a significant turning point in developing text-as-data approaches (Gigley, 1993; Gilardi & Wüest, 2020). NLP is a subfield of computer science where machines learn, understand, and produce human language content using computer technology and learning from large amounts of unstructured data (Hirschberg & Manning, 2015).

This master's thesis builds on previous work and data from the C3 Project. The C3 Project, part of the DFG-funded Collaborative Research Center 884, "The Political Economy of Reforms," initially based at the University of Mannheim and later relocated to the University of Vienna, undertook a comprehensive analysis of the political dimensions of economic reforms in EU member states since the 1980s. Its primary objective was to identify the conditions under which various governments could effectively implement economic reforms in a multiparty democratic context (Angelova et al., 2018; Bergman et al., 2023; Bläck et al., 2022; Strobl et al., 2021). The original dataset for the project was limited to coding only the initial mention of each reform (singleton mentions) through a manual coding process of economic and social reforms from the Economist Intelligence Unit's (EIU) Country Reports. Disregarding further mentions of the reforms, due to the project's overall goal of counting unique reforms, has now proven to constrain the potential for data analysis using automated processes (Lai et al., 2022; Bengtson & Roth, 2008).

One of the primary tasks of this thesis was to explore the transition from manual to automated disambiguation of policy reforms and their mentions. To facilitate this transition, we implemented a simple, unsupervised baseline model with two main objectives (Stolfo et al., 2022). First, the baseline model was a reference point for comparing it with the disambiguator system. Second, it assisted in developing a clustering technique required to utilize the Link-Based Entity-Aware (LEA) metric by Moosavi and Strube (2016) in evaluating the disambiguation model, which can be considered *'off-the-shelf pre-trained sentence embedding model.'* The baseline model was crucial for creating the ground truth and prediction datasets, thereby allowing for a systematic evaluation of the model's performance (Sennikova, 2020; Prakash, 2023).

Thus, a significant contribution of this thesis was augmenting the original dataset to facilitate the disambiguation process. This involved manually identifying further mentions of policy reforms within the EIU reports. The resulting expanded dataset, now including multiple mentions, provides the necessary benchmark *(ground truth)* for evaluating the disambiguation model using the LEA metric.

Based on the challenges and the tasks mentioned above, our **Research Questions** can be formulated as the following:

1) ***Can NLP techniques efficiently extract and disambiguate multiple mentions of policy reforms from the available text corpora of the EIU Country Reports?***
2) ***How effectively can a pre-trained sentence embedding model, combined with a simple clustering technique, identify and categorize multiple mentions of policy reforms?***
3) ***Can a simple unsupervised disambiguation system outperform an all-singletons baseline?***

RQ 1: Our results show that while a pre-trained sentence embedding model has the potential to identify policy reforms, a tailored approach is necessary for more complex clustering tasks. The baseline model laid the foundation for this exploration, revealing the intricate balance required between overgeneralization and over-segmentation in clustering techniques.

RQ 2: We had 206 unique IDs in the data frame, of which 59 had at least one or more references. We detected the best threshold to be 0.9167. The disambiguation results are as follows:

Table 1. Summary of Results at the Best Threshold

	Validation Set	**Test Set**
LEA Recall	0.3879	0.5507
LEA Precision	0.4023	0.5507
F1	0.3950	0.5507

The observed changes in LEA Recall, Precision, and F1 Score between the validation and test sets indicate that the machine learning model shows limited but noticeable potential in recognizing and categorizing policy reforms. However, these improvements were not consistent or balanced across thresholds, highlighting the inherent difficulty in fine-tuning the model for optimal performance.

After determining the optimal threshold through validation, the model's performance on the test set did not suggest that the chosen threshold generalized effectively. Instead, it revealed that the model’s behavior shifted drastically based on the threshold, from over-clustering to over-segmentation. It demonstrated the challenges of working with a pre-trained sentence embedding model without specific training on multiple reform mentions.

The Jaccard similarity also reconfirmed this result. Despite progressing from one large cluster to many smaller ones, these are predominantly singleton clusters. Thus, the model struggles to capture multi-element clusters accurately.

The oscillation between under and over-clustering indicates that further refinement is required, particularly in how the model handles complex contextual relationships between mentions (Grimmer & Stewart, Lin et al., 2022; 2013). A customized training process and a full training–validation–test cycle with a more extensive dataset may mitigate the current limitations. Despite these challenges, the study contributes to the project's broader objectives by offering a starting point for the automated analysis of policy reforms.

RQ 3: The baseline model was crucial for evaluating whether a simple unsupervised disambiguation system could outperform an all-singletons baseline. The initial evaluation with an all-singletons baseline resulted in a Precision of 0.106, a Recall of 1.0, and an F1 Score of 0.191. The challenge was to improve upon this baseline through our disambiguation system. Despite its success, the results must be handled carefully, as demonstrated by the above evaluation metric results.

The master's thesis is structured as follows: first, we present the problem statement and the thesis task. Next, we delve into the theoretical foundations of Machine Learning, Natural Language Processing techniques, and standard evaluation methods. We also provide an overview of coreference resolution and introduce the LEA Scorer, which will serve as an evaluation metric for our baseline model. Following the presentation of the original dataset, we explain the rationale behind creating an augmented dataset and discuss the annotation process in detail. The methods chapter elaborates on how the policy disambiguator model is configured and how the LEA Scorer and Jaccard index are used to evaluate our results. Subsequently, we summarize our analysis results and conclude with recommendations for further improvements.

2. Problem Statement

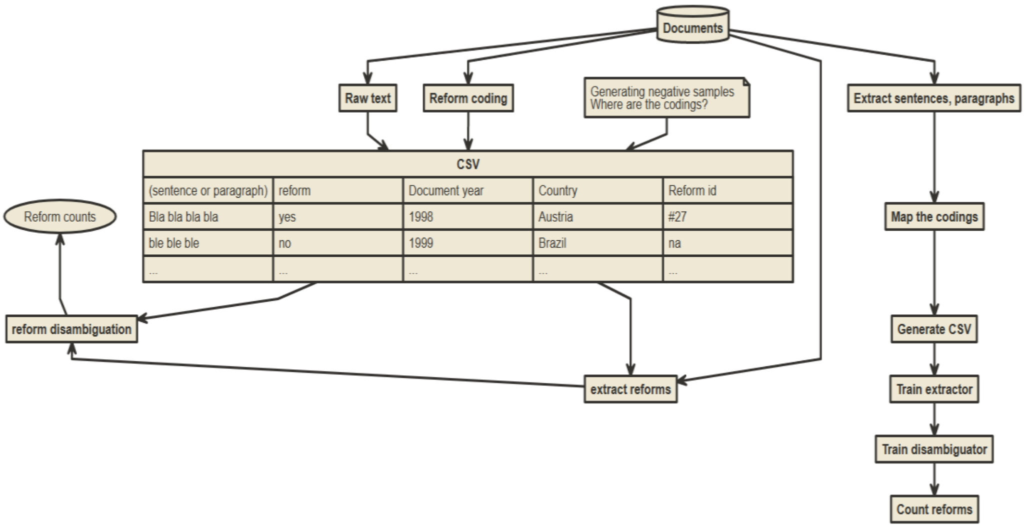
Developments in Natural Language Processing (NLP) and Machine Learning (ML) can aid political science by overcoming the traditional limitations of content analysis. Manual analysis of extensive text collections is time-consuming and costly, which can restrict the scope and depth of research. Automated text analytics reduce the need for extensive human resources and enable systematic analysis of large volumes of documents, uncovering new phenomena and correlations within political texts. This shift streamlines traditional analytical practices and uncovers new phenomena, concepts, and correlations previously obscured within political texts. (Grimmer & Stewart, 2013; Wilkerson & Casas, 2017)

The C3 project relied heavily on manual annotation to identify unique policy reforms, which, considering the project's overall aim, proves to be unsustainable for large-scale analysis due to the intensive time and economic resources required. Accordingly, the dataset provided by the C3 project presented several significant challenges, underscoring the complexity of the research due to the lack of multiple mentions:

- The small labeled dataset did not allow for a robust training phase. Training the machine learning model to recognize patterns and nuances in policy reform mentions was difficult without substantial annotated data.
- Only a single mention per policy reform limited the ability to conduct a complete evaluation. Without annotation for multiple mentions per policy reform, evaluating the model's performance in identifying subsequent mentions of the same policy reform is impossible.
- Manual annotation is susceptible to the annotators' biases and inconsistencies, particularly in interpreting what constitutes the 'first mention' and deciding how it should be coded. Different annotators might identify various aspects of the policy as primary, leading to variability that can affect the reliability of the data.
- Manual coding is labor-intensive, making scaling the analysis to a global level, as intended in the project's expansion goals, impractical. The time and cost involved in manually coding an increasing volume of documents would be prohibitive.

Therefore, integrating NLP in the analysis of the C3 dataset is highly motivated by overcoming these constraints. First, we aimed to enhance the methodology used in the C3 Project by developing a comprehensive data annotation process for policy reforms. The focus is on accurately identifying and categorizing unique policy reforms and their mentions within texts, moving beyond the traditional single-mention coding approach, where the annotators only coded the first reference of the reform, which was a considerable challenge to overcome. By creating a detailed and consistent dataset, the project aims to facilitate the identification of distinct mentions of policy reforms, which is essential for deploying the disambiguator model.

Given these challenges, this thesis seeks to bridge the gap between the original purpose of manual coding and preparing an evaluation dataset for policy reform disambiguation. Testing how the disambiguator performs contributes to the project's broader objective of counting unique reforms, as depicted in *Figure 1*—the project’s roadmap. It enables more efficient and comprehensive data collection and analysis on a global scale.



(sentence or paragraph)	reform	Document year	Country	Reform id
Bla bla bla bla	yes	1998	Austria	#27
ble ble ble	no	1999	Brazil	na
...	...	...	...	...

Figure 1. Roadmap of the Project

To contribute to achieving the overall project goal, this thesis intends to develop a baseline model for the policy disambiguator system to tackle the challenge presented by the original coding, which only allowed one mention per reform to be coded and did not record all mentions that referred to the same reform, thereby distinguishing unique reforms. Evaluating reform disambiguation systems necessitates multiple mentions per reform. Thus, a primary focus of this thesis was to augment the original dataset to include multiple mentions, allowing for the application of clustering techniques and the LEA metric for evaluation.

Accordingly, we implemented the following steps as our **Research Objectives and Tasks**:

1) **Augmentation of the Original Coding:** Given the absence of multiple-mentioned annotations of policy reforms, we augmented the original coding on selected EUI reports to overcome these shortcomings by including further mentions of the policy reform that were not coded, creating a usable test set for our disambiguation system.
2) **Application of Clustering Techniques:** We applied clustering techniques to prepare a golden and prediction dataset. The newly annotated dataset was considered ground truth because it now contained multiple mentions of policy reforms. The prediction dataset was prepared using "all-MiniLM-L6-v2" sentence transformer embeddings.
3) **Reduction of the Train-Validation-Test Cycle:** We reduced the train-validation-test cycle to validation and testing because we used a pre-trained sentence encoder. We calculated the ground truth and prediction datasets for each.
4) **Threshold Optimization Using LEA Scores:** In the validation cycle, we used a random, equally distant threshold to find the optimal one, checking the model's performance using the LEA-adjusted Precision and Recall method.
5) **Evaluation of Test Set:** Then, we checked whether the models improved by applying the best threshold on the test set.

The steps of the Research Objectives will be detailed in the Data and Methods Chapters, outlining each stage of the project from data preparation to the final analysis.

3. Theoretical Background

This section introduces the basic computational techniques that form the foundation of natural language processing techniques (NLP) and their application in text analysis. In the first part, we discuss the theoretical background of machine learning, highlighting the differences between supervised and unsupervised learning techniques. The second part describes deep learning methods and how the transformers renewed the field. Thirdly, we will discuss the development of coreference resolution systems, outlining related works.

3.1 Machine Learning and NLP Methods

Machine learning, a crucial subfield of artificial intelligence, enhances performance on various tasks by learning from experience and is considered the heart of data science (Jordan & Mitchell, 2015). According to Mitchell (1997), a computer program is said to learn from experience *E* for some task *T* and performance measure *P* if its performance on *T*, as measured by *P*, improves with experience *E* (Carpintero-Rentería et al., 2019; Das & Behera, 2017). In the context of NLP, machine learning algorithms, fueled by large datasets, learn to process and analyze language without being explicitly programmed (Ezugwu et al., 2022; Jin & Rinard, 2023). These algorithms are broadly categorized into supervised and unsupervised learning methods (Collobert et al., 2011; Goodfellow et al., 2016).

Supervised Learning

Supervised learning algorithms use labeled datasets to train models, which enhances their precision and accuracy. Based on the input features, in other words, labels, the algorithms predict the output. These models predict outcomes or classify data based on learned relationships between input and output data. For example, supervised learning is crucial in applications such as spam detection or voice recognition systems. Standard algorithms include linear classifiers, decision trees for classification tasks, Support Vector Machines, and linear and logistic regression models for predicting numerical outcomes (Nasteski, 2017; Castelli et al., 2018). An illustrative example is a model predicting commute times by analyzing variables such as time of day and weather conditions, adjusting predictions based on factors like rain, which can extend driving times (Saravanan & Sujatha, 2018).

Unsupervised Learning

Unsupervised learning, in contrast, involves algorithms that identify patterns and structures in unlabeled data without explicit outcomes to predict. These models autonomously discover inherent structures in the data, which is crucial for tasks like clustering, association, and dimensionality reduction. For example, K-means clustering might be used to segment markets or group data points based on similarities without prior knowledge of the groups. Association rule learning, commonly utilized in recommendation systems, identifies items frequently purchased together, such as bundling baby clothes with diapers and applesauce. Despite their independence from labeled data, these models often require human intervention to ensure the outputs' relevance and accuracy (Ghahramani, 2003).

Using machine learning faces a crucial problem, overfitting, as Goodfellow et al. (2016) and Ying (2019) highlighted. The ultimate goal of an algorithm is to perform well not only on the training data but also on unseen input. Various techniques can be employed to prevent overfitting to achieve this goal, including regularization, dropout, and early stopping. Regularization can take the form of L1 or L2, where the former adds the absolute value of the weight to the cost function, and the latter adds the squared value of the weight to steer the weight toward a moderate direction. Dropout can help reduce interdependency between units by ignoring the output of some network units. Finally, early stopping can be triggered to halt training when the validation loss stops decreasing (Ying, 2019).

Supervised and unsupervised learning have unique advantages and are selected based on the specific needs of the task. While supervised learning models benefit from the clarity of learning signals provided by labeled data, unsupervised learning is invaluable for exploring data where labels are unavailable or obtaining them is impractical.

3.2 Deep Learning

Deep learning, a subfield of machine learning, uses deep neural networks that learn layered representations of data. These networks of multiple layers of interconnected nodes or neurons mimic the neural structures of the human brain to detect and interpret complex patterns in large datasets (Goodfellow et al., 2016). Deep learning is considered "deep" because it uses successive layers in neural networks that perform mathematical operations to map non-linear relationships between inputs and outputs. This multi-layered approach enables the model to learn complex structures hierarchically (Hodnett, 2019; Kavlakoglu, 2020).

Each layer in a deep neural network transforms its input data into a more abstract and composite representation (Bengio, 2016). The network requires input data to be vectorized, allowing it to process information through a series of function chains that generate layers of representations (Goodfellow et al., 2016). The model thus evolves its understanding, layer by layer, learning intricate details from simple to progressively more complex patterns (Goldberg, 2016).

Among the most basic yet powerful architectures in deep learning is the feedforward neural network (FNN), a multilayer Perceptron. The original perceptron model is an online classic learning algorithm: instead of considering the entire dataset, it only looks at one example at a time. It is also error-driven, meaning it does not update its parameters as long as the model runs. It has only one hyperparameter, which is the maximum number of iterations needed to avoid overfitting (Daumé, 2017). Initially, this model was used as a binary classification.

Multilayer perceptrons are a more advanced type of neural network. FNN processes input data forward through its layers without any loops, making predictions based on integrating information from previous layers (Goodfellow et al., 2016). It employs forward propagation to make initial predictions. It uses backpropagation—refining neural network predictions by adjusting weights through gradient descent—to minimize errors in output layers, learning from the data incrementally (Duong, 2023).

Advanced deep learning methodologies have revolutionized our ability to work with diverse data types in NLP (Bein et al., A. 2023). Convolutional Neural Networks (CNNs) are feedforward neural networks that use filters and pooling layers and are particularly adept at identifying patterns in visual content such as images and videos (LeCun & Bengio, 1995). On the other hand, Recurrent Neural Networks (RNNs) apply backpropagation through time and solve machine learning problems involving sequential data like speech recognition, machine translation, and other language modeling tasks (Lipton et al., 2015).

While each type of network specializes in different tasks, they can sometimes work together effectively, such as in projects that involve adding captions to videos. However, RNNs sometimes face challenges with data that have long sequences because they can lose track of earlier information due to problems known as vanishing or exploding gradients (Goldberg, 2016; Hochreiter & Schmidhuber, 1997). Newer solutions like attention mechanisms or transformers have been developed to address these issues.

Bidirectional LSTMs, or Bi-LSTMs, are neural networks that process sequential data in forward and backward directions. This makes them particularly valuable for language translation and sentiment analysis (Wibawa et al., 2024).

Deep learning has transformed the field of NLP, allowing computational models to accomplish various tasks. These tasks include primary text classification and more complex tasks such as semantic parsing, machine translation, and automated summarization.

3.2.1 Transformers

The transformer model, introduced by Vaswani et al. in 2017, revolutionized machine learning with attention mechanisms for processing sequential data such as natural language, and unlike traditional approaches like RNNs and LSTMs, transformers process input data simultaneously in parallel, leading to greater efficiency and faster training times for large datasets (Javaloy & García-Mateos, 2020; Smys & Raj, 2021). At the heart of the transformer lies its attention mechanism using queries (Q), keys (K), and values (V), which dynamically assign importance to different parts of the input data. By using queries, keys, and values, this mechanism can determine the significance and relevance of each word in a sequence, allowing the model to focus on the most relevant information and optimize its understanding and processing of language (Vaswani et al., 2017; Niu et al., 2021).

The architecture of transformers consists of two primary components: the encoder and the decoder. Both components include multiple layers that carry out specific functions. The encoder transforms the input text into vectors representing the embedded words. Positional encoding is added to these vectors to provide context about the word positions within the sentence because transformers do not process data sequentially by nature (Lin et al., 2022).

Each encoder layer comprises multi-head attention mechanisms and feed-forward neural networks, which allow the model to focus on different positions within the sequence simultaneously (Niu et al., 2021). To ensure that the input data remains intact throughout processing, residual connections are used around each sub-layer, followed by layer normalization. This technique helps mitigate the vanishing gradient problem commonly encountered in deep neural networks (Lin et al., 2022; Vaswani et al., 2017).

Like the encoder, the decoder architecture incorporates additional elements, such as masked multi-head attention layers that restrict its view to prior tokens in the sequence. This ensures that predictions for a token rely solely on the preceding known tokens, thus preserving the autoregressive nature of the model. The decoder's output is then translated into a probabilistic format using a softmax layer, calculating the likelihood of each subsequent word in the sequence (Chorowski & Jaitly, 2016).

In general, the transformer model represents a significant breakthrough in NLP, offering a more efficient and effective method for handling sequential tasks than traditional models. Its ability to process inputs in parallel while focusing on the contextual relationships within data makes it useful for various applications, including translation, text summarization, and content generation.

BERT

Masked Language Modeling (MLM) is used in natural language processing, where specific tokens within a sequence are randomly masked or hidden from the model during training. The primary task for the model is to predict these masked tokens, based not only on the tokens immediately surrounding the mask but also on the entire context of the sentence. This approach enables the model to effectively process and understand data bidirectionally, considering the preceding and subsequent tokens in making predictions. Such a capability is crucial for tasks that demand a nuanced understanding of language context within sequences (Devlin et al., 2018).

BERT (Bidirectional Encoder Representations from Transformers) exemplifies this type of model. Developed to enhance context comprehension, the attention mechanism allows BERT to process words about each other across a sentence simultaneously. This architecture allows BERT to capture a deeper understanding of language nuances, vastly improving performance on a wide range of NLP tasks, from sentiment analysis to coreference resolution (Briggs, 2021; Salazar et al., 2020).

In addition to MLM, BERT incorporates a technique known as next-sentence prediction (NSP). NSP involves predicting whether a second sentence in a pair logically follows the first, fostering a model's ability to understand relationships and connectivity between sentences and enhancing its predictive capabilities in multi-sentence contexts (Devlin et al., 2018; Shi & Demberg, 2019).

These innovative training strategies—MLM and NSP—enable BERT and similar models to predict missing words and refine their predictions based on the broader context, leading to a richer and more accurate language understanding. This bidirectional approach marks a significant advancement in NLP, propelling the capabilities of language models to new heights and broadening their applicability across various complex linguistic tasks.

Sentence Transformers

Vector embeddings, often called "embeddings," are numerical representations of words, phrases, or entire documents. These vectors capture the semantics or meaning of the original data in a high-dimensional space. Unlike traditional count-based representations like Bag of Words or TF-IDF, embeddings capture semantic relationships between words, enabling machines to understand and process language more effectively.[1]

The *all-MiniLM-L6-v2*[2] sentence-transformers model is designed to map sentences and paragraphs into a 384-dimensional dense vector space. This model, fine-tuned on a large dataset of over 1 billion sentence pairs using a contrastive learning objective, can be used for tasks like clustering or semantic search. It encodes text into dense vectors that capture semantic information, facilitating tasks such as semantic search and text clustering (Martinez, 2023).

3.3 Natural Language Processing

Deep learning has enhanced natural language processing capabilities, enabling advanced applications such as topic classification, emotion analysis, question answering, and language translation (LeCun et al., 2015; Smith, 2019). These tasks can be categorized into those that analyze texts at the linguistic level and those that delve into semantic understanding (Collobert et al., 2011; Wankmüller, 2022).

Linguistic analysis involves several tasks that help break down and structure text data. Sequence segmentation, for instance, divides text into manageable segments, like tokenization, where texts are split into tokens or words (Smith, 2019). Language modeling predicts the next word in a sequence based on preceding tokens, laying the foundation for many predictive text

[1] https://www.aimodels.fyi/models/huggingFace/all-minilm-l6-v2-sentence-transformers
[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

applications (Bengio et al., 2003). Lemmatization and stemming reduce words to their base or root forms, standardizing variations in text (Manning et al., 2008). Part-of-speech tagging identifies each word's grammatical category, helping with sentence parsing and grammatical analysis. Syntactic parsing, another critical task, analyzes the syntactic structure of sentences to clarify the grammatical relationships between words. (Das & Smith, 2011)

On the other hand, semantic analysis tasks aim to grasp the deeper meaning of text. Word sense disambiguation determines the context-dependent meaning of words, such as distinguishing between different senses of the word 'party' (Raganato et al., 2017). Coreference resolution links pronouns and references to the same entity within a text (Pradhan et al., 2012). Information extraction pulls structured information from unstructured text, including tasks like named entity recognition and relationship extraction (Sarawagi, 2008). Semantic parsing transforms text into a structured, machine-readable format like semantic graphs (Cai & Knight, 2013). Text summarization condenses longer texts into concise summaries, using extractive methods to select key sentences or abstractive methods to generate new sentences (Rush, 2015; Nallapati et al., 2016; Ngoko, 2018).

Topic modeling uncovers themes or topics within large bodies of text, helping reveal latent content in a corpus (Chang et al., 2009). Sentiment analysis identifies the sentiment expressed in a text, determines who holds the sentiment, and finds the target of that sentiment (Liu, 2015). Natural language inference, another task, determines whether a statement is true, false, or undetermined based on the text (Bowman et al., 2015). Dialogue systems engage in conversations by tracking user preferences and generating appropriate responses (Henderson et al., 2014; 2019). Finally, question-answering systems respond to user-posed questions by pulling relevant information from texts (Saeidi et al., 2018; Wankmüller, 2022).

Beyond these tasks, natural language generation is key in automated report writing and content creation applications. This includes text-to-text and data-to-text generation, showcasing how deep learning technologies enhance NLP systems' ability to understand and generate text (Gatt & Krahmer, 2018; Wankmüller, 2022). Natural Language Generation tasks like text-to-text and data-to-text generation also play critical roles in applications such as automated report writing and content creation (Gatt & Krahmer, 2018; Wankmüller, 2022). These tasks highlight integrating deep learning technologies to enhance NLP systems' understanding and generation capabilities.

3.3.1 Evaluation Metrics in NLP

Only appropriate evaluation metrics can help train models effectively and assess their performance in machine learning and natural language processing. Accuracy, precision, recall, and F1-score are standard metrics for evaluating text classification and information retrieval models. Other metrics, such as BLEU and ROUGE, are used in machine translation evaluation, and Perplexity and WER metrics are used for Automatic speech recognition and text generation (Vickers et al., 2024). We concentrate on the Accuracy, precision, recall, and F1-score measures (Mungalpara, 2023):

- True Positive (TP): Instances *correctly* identified as positive.
- True Negative (TN): Instances *correctly* identified as negative.
- False Positive (FP): Instances *incorrectly* identified as positive.
- False Negative (FN): Instances *incorrectly* identified as negative.

In machine learning, accuracy is a crucial metric that denotes the proportion of correctly classified instances compared to the total number of instances. Regarding natural language processing tasks, accuracy is often a critical indicator of a model's overall performance. However, it may not be enough in cases where there are imbalanced classes or different error costs. To fill this gap, precision and recall come into play. Precision measures the accuracy of predictions for the positive class, while recall measures how well the model identifies all positive instances in the dataset. It is essential to comprehend these metrics to make informed decisions about improving and utilizing a model. This is particularly relevant in binary classification problems. In multi-class problems, these metrics might be adjusted to understand how well a model performs thoroughly (Hossin & Sulaiman, 2015; Leevy et al., 2018).

Precision is the proportion of true positives to the total number of instances identified as positive. It is defined as:

Equation 1. Precision

$$Precision = TP / (TP + FP)$$

Recall, or sensitivity, measures the proportion of actual positives correctly identified by the model, reflecting its ability to detect positive instances (Al-Saadawi, 2024). It is defined as:

Equation 2. Recall

$$\text{Recall} = TP / (TP + FN)$$

F1-score is the harmonic mean of precision and recall. It is beneficial when the classes are imbalanced. The F1-score ranges from 0 to 1, where 1 represents perfect precision and recall (ibid.). It is defined as:

Equation 3. F1 Score

$$F1 = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

When data is imbalanced, precision and recall carry equal significance. However, there are instances where one is prioritized over the other (Gupta, 2021). In the case of text data analysis, recall may be more significant when missing a relevant instance (false negative) is more harmful than incorrectly identifying an irrelevant instance (false positive). This is because recall helps address the potential for selection bias, which can occur if the likelihood of a document being selected as relevant varies with its actual relevance to the topic under study.

While precision provides critical insights into the model's classification accuracy effectiveness by focusing exclusively on true and false positives, many false positives can undermine the trustworthiness of the results and, if not correctly managed, could skew the analysis. Thus, high recall rates help minimize such biases' impact and ensure that fewer relevant documents are overlooked (Wankmüller, 2022).

Ultimately, the F1 Score balances the trade-off between precision and recall, providing a single measure to evaluate a model's accuracy without disproportionately favoring one metric. This evaluation is essential for developing robust NLP systems that reliably interpret and analyze text data. (Manning et al., 2008; Wankmüller, 2022)

3.4 Coreference Resolution

Coreference resolution is a critical process in natural language processing that helps identify all phrases or mentions within a text that refer to the same entity. Doing so facilitates a coherent understanding of narrative threads and relationships across a document, making it an essential tool in various applications such as information extraction, question answering, text understanding, and machine translation (Pradhan et al., 2012; Pradhan et al., 2014).

The significance of coreference resolution has been widely recognized in the academic community. It enables the disambiguation of pronouns, nouns, and other referring expressions, thereby enhancing the overall comprehension of text. It also helps identify the relationships between entities in text and provides a better understanding of the structure and semantics of language.

Coreference resolution continues to be an active area of research and development in computational linguistics. Its crucial role in interpreting textual data makes it a critical element in natural language processing and information retrieval. Historically, coreference resolution techniques developed in the mid-90s with researchers like McCarthy and Lehnert, who experimented with decision trees and rule-based systems. These early efforts were supported by foundational corpora from the Message Understanding Conferences (MUC) and later the ACE1 datasets, which have been pivotal in advancing this field (Hirschman & Chinchor, 1997; Chinchor, 2001; Doddington et al., 2004).

Pradhan et al. (2012) define coreference resolution as a task that not only detects entities and events but also identifies all mentions of these entities and clusters them into equivalence classes. The historical backdrop of coreference resolution includes the work of McCarthy and Lehnert in the mid-90s, who utilized decision trees and hand-written rules. The MUC and ACE1 corpora, established by Doddington et al. (2004), have since become standard datasets for coreference studies and have supported significant achievements in automatic coreference entity recognition. These include the MUC's coverage of smaller-sized corpora, the development of inter-annotator agreement (ITA) metrics by Hirschman et al. (1998) and subsequent enhancements by Poesio et al. (2004), Poesio & Artstein (2005), and Passonneau (2004), the implementation of complex evaluation metrics, and addressing the knowledge bottleneck.

Pradhan et al. (2012) suggested that improvements in coreference resolution metrics require a large corpus with high inter-annotator agreement covering an unrestricted set of entities and events, a standard evaluation scenario, and enhanced learning algorithms.

The process of coreference resolution generally unfolds in two main stages: first, identifying the mentions in the text that could potentially refer to the same entity, and second, determining whether these mentions correspond to the same referent. This task extends beyond mere mention detection, as it involves clustering mentions into equivalence classes representing single entities or events (Cai & Strube, 2010).

Coreference resolution tests a system's linguistic capabilities and artificial intelligence, challenging models to interpret text with a human-like understanding of nuanced relationships (Morgenstern et al., 2016). The development of neural network models has significantly advanced this domain. Earlier models relied heavily on manually crafted rules and parsing algorithms (Wiseman et al., 2016; Clark and Manning, 2015, 2016), while more recent approaches employ neural networks to handle the task in an end-to-end manner, simultaneously detecting and resolving coreferences (Lee et al., 2017, 2018; Kantor & Globerson, 2019).

Modern methodologies in coreference resolution can be categorized into entity-level models that construct and leverage holistic representations of entities and mention-ranking models that prioritize linking anaphoric mentions to their antecedents. An example of this innovative approach is the CorefQA model, which integrates question-answering techniques to enhance coreference identification, showcasing the integration of contextual depth in model training (Wu et al., 2020).

Example Methods for Coreference Resolution[3]:

- Rule-based systems utilize predefined linguistic rules to detect coreferences, considering syntactic structure, gender agreement, and proximity. While straightforward to implement, rule-based methods may encounter challenges with more intricate linguistic patterns.
- Supervised machine learning models can predict coreference relationships using syntactic and semantic similarity features and word embeddings. These models include decision trees, support vector machines, and logistic regression.
- Deep learning, specifically neural networks, has transformed coreference resolution. Recurrent neural networks (RNNs) and transformer-based models capture complex text patterns and dependencies. Transformers like BERT have delivered impressive results in coreference resolution.

[3] https://medium.com/p/5ba4f570bffe

- Hybrid Approaches combine rule-based and machine learning or deep learning methods. They can, for example, establish initial coreference clusters, which can then be refined using machine learning or deep learning models.

As coreference resolution techniques evolve, they increasingly incorporate the contextual precision and nuanced understanding of transformer-based models, setting new standards for accuracy and efficacy in NLP.

3.4.1 Evaluation Metrics for Coreference Resolution

Evaluation metrics are critical for assessing the performance of models in NLP, especially for complex tasks like coreference resolution. Historically, traditional metrics such as MUC, B3, CEAF, and BLANC have provided a structured framework to evaluate coreference systems.

These traditional metrics, however, have faced criticism regarding their comprehensiveness in evaluating coreference resolution systems. Early research by Luo et al. (2005), Denis & Baldridge (2007), Culotta et al. (2007), and Haghighi & Klein (2009) simplified the resolution task by providing systems with key mentions, thus limiting evaluations to key mention-based systems. This approach was pointed out as a significant limitation by critics such as Ponzetto and Strube (2006), Bengtson and Roth (2008), and Stoyanov et al. (2009). They argued that these metrics do not provide an end-to-end assessment of coreference resolution systems, highlighting a critical gap in the traditional evaluation methodology (Cai & Strube, 2010).

This gap led to a search for metrics that could offer more interpretability and discriminative power, which is essential for a comprehensive evaluation. In response, the Link-based Entity-Aware (LEA) metric was developed by Moosavi and Strube (2016). The LEA metric addresses these concerns by considering the importance of each link between mentions within an entity chain, offering a more nuanced and detailed evaluation of how well a system performs coreference resolution. This metric evaluates not just the presence of correct links but also their relevance and accuracy in the context of the entire document, providing a deeper insight into the effectiveness of coreference resolution systems.

LEA represents a significant advancement in the field by providing a more transparent, more detailed understanding of a model's performance in real-world tasks. This is particularly useful for developing systems that require a high degree of linguistic understanding and are sensitive to the contextual nuances in text.

3.5 Application of NLP in Political Science

Natural Language Processing has transformed political science research by enabling the automatic analysis and extraction of valuable insights from textual data. This shift is particularly significant given the historical limitations posed by the high costs and complexity of analyzing text data. With the advent of digital data sources and computational tools, the field has embraced text-as-data approaches, significantly enhancing the scope and depth of research (Gigley, 1993; Gilardi & Wüest, 2020; Grimmer & Stewart, 2013; Glavaš et al., 2019).

The increasing availability of NLP techniques has decreased the costs of analyzing extensive text collections, leading to widespread adoption in political and social sciences (Mayer & Pfeffer; Shah et al., 2015; Engel et al., 2021; Luz, 2022). This shift has expanded the scope of inquiry for social scientists, underscoring the power of NLP as a computational tool for textual data analysis (Jurafsky & Martin, 2000; Jin & Mihalcea, 2022). The growing popularity of unstructured data, coupled with the computational power to analyze it, has opened up new research opportunities and holds great promise for the future of political science.

Integration and Impact of NLP in Political Science

The adoption of NLP in political science marks a paradigm shift. It is increasingly becoming mainstream due to its ability to structure large volumes of textual data into formats that reveal political trends and dynamics. This abundance of data has revolutionized the ability of researchers to ask and answer empirical questions, analyze political actors, and uncover nuanced insights into political behavior (Benoit, 2019).

Machine learning, including NLP, contrasts with traditional scientific inquiry, often focusing on identifying and estimating causal effects. Machine learning improves conceptualization, operationalization, and measurement of concepts in political science. This typically involves a deductive process where a vague theoretical concept is clearly defined, dimensionality determined, and then operationalized into measurable indicators. This is crucial because political science frequently deals with latent characteristics within texts that are produced across various political processes, such as election manifestos or legislative speeches (Diekmann, 2007; Ahlquist & Breunig, 2012; Wankmüller, 2022).

Applications and Methodologies

NLP facilitates complex analyses such as sentiment analysis, topic modeling, and document similarity estimation. Techniques like unsupervised learning, including k-means clustering, help organize documents by maximizing differences between groups, which must be interpreted post-analysis. Such capabilities are invaluable in studying political manifestos, understanding shifts in political strategy, and exploring policy areas and topics (Grimmer & Stewart, 2013; Benoit, 2019; Orellana & Bisgin, 2023).

Furthermore, NLP addresses the limitations of traditional quantitative methods, enabling more timely and cost-effective insights into political opinions and sentiments. It allows, for example, studying political polarization, providing a more profound understanding and enabling more effective communication strategies (Ahmed, 2021; Németh, 2023; Wankmüller, 2022).

NLP's adoption in political science has been driven by its ability to perform complex analyses such as sentiment analysis, topic modeling, and the estimation of document similarity, which are invaluable for content analysis (Orellana & Bisgin, 2023; Jin & Mihalcea, 2022). For instance, unsupervised learning techniques like k-means clustering are employed to organize documents into groups that maximize inter-group differences and minimize intra-group differences, although these groups must be interpreted post-analysis (Grimmer & Stewart, 2013; Benoit, 2019).

One critical application of NLP is the study of political party manifestos, as illustrated by the research of Orellana and Bisgin (2023), which employs NLP techniques to dissect and understand the ideological underpinnings and strategic communications within these documents. Furthermore, NLP aids in examining how political parties evolve, focusing on the success and visibility of their policy agendas. This analysis extends to understanding the changing attitudes of political parties towards various policy areas and topics, revealing shifts in political strategy and priorities.

Moreover, Ahmed (2021) demonstrates how NLP can address the limitations inherent in traditional quantitative methods, such as surveys or regression analysis used to gauge political opinions and sentiments. By applying NLP, researchers can drastically reduce the costs and time required for evaluation, providing more timely and cost-effective insights into public opinion. Additionally, NLP is instrumental in studying political polarization, a growing area of interest within political discourse, as explored by Németh (2023). This application is crucial in today's political environment, where understanding the degrees and nature of polarization can inform more effective political communication strategies.

As Jin and Mihalcea noted (2022), NLP is a vital policymaking tool. By extracting actionable information from text and performing sentiment analysis, NLP helps policymakers capture the nuances of public opinion and sentiment, facilitating more informed and responsive policy decisions. Through these varied applications, NLP not only enhances our understanding of political processes but also enriches the methodological arsenal of political scientists, enabling a deeper and more nuanced exploration of political phenomena.

Challenges and Opportunities

Working with NLP and digital data presents significant challenges and advantages for political scientists, especially in supervised learning and data integrity. One of the foremost challenges involves the inherent selection biases within the data generation and collection processes. Biases in the dataset can significantly impact the outcomes of any analysis conducted, as they may be associated with the values of dependent or explanatory variables. This can affect the accuracy and validity of the conclusions drawn, making it essential for political scientists to carefully identify and address these biases to ensure the reliability of their inferences (Wankmüller, 2022).

When dealing with text data, humans are often viewed as the ultimate arbiters of validity, equipped with the nuanced understanding necessary to make informed conceptual judgments—such as determining the sentiment expressed toward a specific entity or whether a text aligns with a particular conceptual category (Krippendorff & Craggs, 2016; Wankmüller, 2022). The validity of text-based supervised learning approaches is traditionally assessed by comparing the model's predictions with human coding, where a lower expected generalization error signifies higher validity (Grimmer & Stewart, 2013, p. 279). Despite evidence suggesting that human coding is not infallible (Mikhaylov et al., 2012; Ennser-Jedenastik & Meyer, 2018), it remains a widely accepted method for validating text-based measurement instruments in conjunction with comparisons to human judgments (Hayes & Krippendorff, 2007; Xu et al., 2019).

This reliance on human judgment introduces inherent complexities in dataset creation. The development and utilization of these datasets are subject to various selection mechanisms that can introduce biases—specifically, selection biases that occur if the process of selecting observational units is correlated with the units' values on dependent variables or if the assignment of units to explanatory variables is correlated with these values (King et al., 2001). Such biases are not only introduced by human annotation but can also be immanent features in the transformer models (Yang et al., 2023). Understanding these mechanisms is critical as they may influence the representativeness and reliability of the dataset, ultimately impacting the outcomes of machine learning models trained on such data.

Despite these challenges, integrating NLP and machine learning techniques offers substantial advantages by achieving a low expected generalization error to ensure accurate predictions, valid results, cost-effectiveness, and time efficiency (ibid.).

Aligning with that, this project aims to strengthen the application of NLP in political science as computational methods continue to advance and become more accessible. These tools enhance understanding of complex political phenomena and democratize data analysis by reducing costs and technical barriers, allowing researchers to tackle previously unanswerable questions. The future of NLP in political science looks promising, with potential for further methodological innovations and broader applications in policy-making and political analysis.

4. Data

This chapter provides insights into the C3 project’s data sources and explains how these data sources were used to create an annotated dataset for disambiguation.

4.1 Data Source

The dataset we use in this thesis is a product of the DFG-funded Collaborative Research Center 884 research project 'The Political Economy of Reforms'. Initially based at the University of Mannheim from 2014 to 2017 and later relocated to the University of Vienna, the project analyzed the political dimensions of economic reforms in EU member states since the 1980s (Angelova et al., 2018; Bergman et al., 2023; Bläck et al., 2022; Strobl et al, 2021). The primary goal of the C3 project was to discern the conditions under which various governments could effectively implement economic reforms in a multiparty democratic context.[4] This research's methodology is based on the Economist Intelligence Unit’s Country Reports, formerly the Quarterly Economic Review until Spring 1986. These reports have been instrumental in our research, providing insights into entire reform processes and detailed background information on each reform measure's conditions and circumstances.

The available dataset was small, considering the project's purpose. It consisted of a selected dataset prepared manually by the C3 project’s annotators. First, we deployed a clustering technique to create the ground truth clusters, and we implemented a disambiguator system using the all-MiniLM-L6-v2 Sentence Transformer for predicted clusters. Then, we evaluated the model using Moosavi and Strube’s LEA Scorer (2016).

The ground truth dataset consists of manually annotated clusters of policy mentions that refer to the same reform (Basile et al., 2021). These clusters serve as a reliable benchmark because they are created based on human judgment and domain knowledge. In contrast, the predicted clusters are generated by the machine learning model. These clusters represent the model's attempt to group policy, which mentions that it believes it refers to the same reform based on the patterns it has learned from the data. Comparing the predicted clusters against the ground truth clusters allows us to evaluate the model's effectiveness in accurately disambiguating and clustering policy mentions.

[4] "Strong" vs. "Weak" Governments and the Challenge of Economic Reforms | Mannheimer Zentrum für Europäische Sozialforschung (uni-mannheim.de)

4.2 C3 Project Dataset

The objectives of the C3 project were to collect information on policy measures at all stages of the political process, from a policy's announcement to the date it comes into force. Various information was extracted from the project’s sources, such as political actors’ documents, EIU Reports, and OECD documents. In our analysis, we only include the EIU reports, all in English language. Specifically, the three types of information of interest are 1) Reform Ambitions, 2) Actual Reforms, and 3) Failed Reforms. In this thesis, we worked only with Actual Reforms and their mentions.

The C3 project involves identifying three distinct types of policy measures extracted from various primary sources. These sources comprise two categories of documents. The first category pertains to primary sources produced by political actors of the respective countries under study, such as government declarations, programs, or coalition agreements. The government declarations used in this study are transcripts of speeches typically held when a new government takes office or at other regular intervals. These documents mainly contain information about the reform ambitions of the respective government, but the specificity of these reform ambitions varies considerably. This project's second category of sources includes periodically published reports from the OECD and the Economist Intelligence Unit (EIU). This firm is part of the Economist Group. The OECD Economic Surveys are published annually in earlier years and once every 18 months in later years, while the EIU Country Reports are issued monthly with fewer pages. The relevant information in the OECD and EIU reports is distributed all over the text and must be searched carefully and checked for relevance. The EIU reports were chosen to code everything that is included for the whole period under study. In contrast, the OECD Economic Surveys were analyzed in a second step to fill the data gaps that might exist or to add additional information about reform processes to the information from the EIU reports.

We created our annotated dataset based on the comprehensive EIU country reports, which we want to explore further. These reports offer insights into ongoing reform processes and the background and context of each measure. They cover everything from reform ambitions announced during the legislative period to descriptions of both successful and unsuccessful reforms. Although the structure of these reports has mainly remained consistent over the years, only a tiny portion of the 20-30 pages in each issue is typically relevant to our purposes. Nonetheless, scanning the headings and other sections for relevant keywords or information is essential. It is worth noting that the same reform measure may be mentioned multiple times in the same or subsequent issues.

The coding of the EUI reports instructed by the C3 Project included that even though there were various annotators in the project and to include all policy mentions without duplicates and excluding further mentions of the reforms, they often did double coding to supervise each other's work; they were instructed only to code the first mention of the reforms. This not only affected within country identified mentions but also further country reports. Accordingly, the aim was to code repeated measures only once within one country report or additional reports. Because of this characteristic of the dataset containing only singleton mentions, we could not train or evaluate our disambiguator in this setting. However, this original database allowed us to test the necessary format to create our baseline and test the LEA Scorer.

Extracted Information

Reform Ambitions refer to identifying governments' general reform attitudes and programs. General announcements or broad aims a government wants to achieve are only extracted from government declarations.

Actual Reforms pertain to information about concluded reform measures mentioned in the country reports. The project did not code information about achievements that might retrospectively be mentioned in a government declaration. The notion of an Actual Reform is broad in scope. Everything that has left the stage of a plan or draft is essentially considered an Actual Reform. This includes measures that the cabinet has decided on, measures in the legislative process, and measures reported as coming into force.

Failed Reforms, on the other hand, refer to cases where legislation is never implemented. A government may scrap a plan or withdraw a bill for various reasons, or a bill may not win a majority in parliament or be rejected as unconstitutional by a court. Information about such events is also relevant to learn about the "strength" of governments. Failed Reforms are, therefore, added to the database. This allows us to see whether reform processes were completed successfully or stopped elsewhere. It is worth noting that Failed Reforms make up only about one percent of all coding.

Although it is important to distinguish the different types of reforms, we must highlight that considering the project's overall goal to count unique reforms, we only utilized Actual Reforms when implementing the disambiguator model.

4.3 Dataset Preparation

We needed to shape the available dataset by the C3 project for the following three purposes:

1) To help develop a clustering function similar to other coreference systems.
2) To develop an identification system for policy reforms and mentions.
3) Test how the adapted LEA Scorer performs on our dataset.

In the coreference system, the task is to find all expressions that refer to the same entity in a text. We wanted to find all policy mentions that refer to the same reform. Also, this format aligns with the requirements of the coreference resolution metric by Moosavi and Strube (2016) to evaluate the disambiguation system. The coreference resolution system builds on ground truth and prediction datasets.

As previously highlighted, one of the significant challenges we faced was that the available dataset only consisted of singleton mentions. Nevertheless, we elaborated an identification system for the policy reforms and mentions. The *'ID Number'* column identifies the policy reforms. If there were further mentions of the same policy reform, we would utilize the *'Row ID'* to detect the different policy mentions referring to the same policy reform. So, in case of multiple mentions belonging to the same reform, e.g., five mentions to one reform, we would have from 1 to 5 different Row ID values, but the ID Numbers were the same for all five mentions.

For our analysis, we worked with selected columns of the dataset: Country Names and corresponding country codes, the issue of the EIU report, and year/month/day. Then, the *'Description of Measure'* depicts those reforms annotated initially by the project's coder.

To apply the LEA evaluation metric, we structured the disambiguator system to align with the format required for coreference resolution evaluations. However, it is important to note that we did not implement a full coreference resolution system. Instead, we adapted our data to fit the necessary format for LEA. Specifically, we extracted a ground truth dataset (also called golden mention clusters) composed of mentions that refer to the same entity—in our case, policy reforms. This golden truth dataset was derived from the original *'Aim/Measure'* column, considered Actual Reforms. We also eliminated 5 Failed Reforms from the Dataset because we only concentrated on Actual Reforms or Reform Ambitions.

Table 2 depicts the original dataset before annotation.

Table 2. Dataset before Annotation, example

Reform ID	Country	Country Code	Issue	Description of Measure
0	Austria	1	198605	The minister finance, Dr Vranitsky, has called for a reduction in the deficit in future years [...]
6	Belgium	2	199901	[[...] general government deficit/ GDP ratio [...] To achieve this result the government both increased taxes and reduced expenditure growth.] The increase in the tax burden was underpinned by higher indirect rather than direct taxes. [...]
111	Netherlands	10	199506	Notwithstanding earlier objections from both the VVD and D66, the PvdA minister of social affairs and employment, Ad Melkert, achieved a measure of success with the application of the G4.Sbn ($2.9bn) fiscal job-stimulation package that had been set aside for 1996 at the time of the coalition agreement.
130	Portugal	11	199505	The acceleration in the rate of decline was partly due to an increase in automotive taxes at the beginning of this year, [but also illustrates an underlying sense of uncertainty on the part of the public.]
144	Spain	12	198701	Fuel prices were lowered in early November as part of government measures to reduce the rate of inflation. Automotive gasoline prices were reduced [...] by 5.3 per cent for ordinary petrol.

4.4 Dataset Annotation

Given that the available dataset consisted exclusively of singletons, it was necessary to augment it to develop our disambiguation model and test it using the LEA Coreference Metric. The C3 project's annotators created this database, and we utilized a sample of 206 reforms. As previously noted, these reforms were singletons because the C3 project only coded the first mention of each reform.

To expand this baseline dataset with multiple mentions, we revised the corresponding EIU reports and searched for additional mentions, sentence by sentence, related to the reforms identified in the database. We added these uncoded mentions and linked them to their respective reforms. For example, from the first Austrian report, EIU Issue 198605:

- **Original code:** *"The minister of finance, Dr. Vranitsky, has hinted at the need for an increase in tax rates to help achieve this goal."* (Row ID: 5; Reform ID: 1)

- **New mention added:** *"Raising taxes in what will probably be the last budget, for 1987, before a general election will not be popular, but for the first time, it now appears that the budget deficit issue is to be addressed seriously."* (Row ID: 6; Reform ID: 1)
- **Another new mention:** *"In calling for action on the deficit, Dr. Vranitsky focused his remarks on receipts rather than expenditures, specifically raising the question of increasing tax rates."* (Row ID: 7; Reform ID: 1)

Similarly, in the second Austrian report, EIU Issue 199408:

- **Original code:** *"Austria will join the European Union (EU) in 1995. This was the outcome of the referendum on EU membership held on June 12."* (Row ID: 10; Reform ID: 203)
- **New mention added:** "*The referendum result in favor of EU entry has strengthened the positions of the foreign minister, Alois Mock (People's Party-ÖVP), the chancellor, Franz Vranitzky (Social Democratic Party-SPÖ), and the president, Thomas Klestil."* (Row ID: 11; Reform ID: 203)

We needed sentence-level extraction rather than paragraph-level annotations to adapt the dataset for our model. Each coded policy from the EUI reports was represented by the sentence containing the action that identified the reform. This was stored in a new column titled *'Actual Reference'*, as demonstrated in *Table 3*, how the dataset looked after annotation. This decision was made in consultation with a representative of the C3 project, agreeing that sentence-level representation would provide more reliable inputs for the model, given the potential for multiple mentions within a single paragraph.

For example, from the Danish EIU Report, Issue 199308:

- **Original annotation ('Description of Measure'):** *"Before the currency crisis, the Social Democrats had already laid the two main planks of their economic strategy. Major changes to the tax system and labour market reforms have been passed by the Folketing, making full use of the coalition's majority of one (No 2-1993, page 14). The tax reforms, the most ambitious for years, are laid out in a program entitled 'A New Course Towards Better Times.'"*

- **Actual Reference (sentence-level extraction):** *"Major changes to the tax system and labour market reforms have been passed by the Folketing, making full use of the coalition's majority of one."*

This process of sentence extraction allowed us to build a more focused dataset for applying the LEA metric, providing clearer boundaries for the disambiguation model to work with.

Table 3. Creating 'Actual Reference', example

Row ID	Reform ID	Country	Description of Measure	Actual Reference
10	203	Austria	Austria will join the European Union (EU) in 1995. This was the outcome of the referendum on EU membership held on June 12. The result was an overwhelmingly decisive one; 66.6% of those who turned out voted in favour of joining and the participation rate was as high as 81.3%.	Austria will join the European Union in 1995.
85	15	Denmark	Before the currency crisis, the Social Democrats had already laid the two main planks of their economic strategy. Major changes to the tax system and labour market reforms have been passed by the Folketing, making tull use of the coalition's majority of one (No 2-1993, page 14). The tax reforms, the most ambitious for years, are laid out in a programme entitled "A New Course Towards Better Times"	Major changes to the tax system and labour market reforms have been passed by the Folketing, making tull use of the coalition's mjority of one.
120	52	France	"In December 1988 Pierre Beregovoy agreed in principle to a state owned insurance company's acquiring a majority holding in astate owned bank. Thus Gan, France's fourth largest insurance company (after UAP, Axa-Midi and AGF), was authorised to raise from 34 per cent to 51 per cent its holding in Credit Industriel et Commercial (Cie) - with the French state's holding being correspondingly reduced from 66 per cent to 49 per cent. [...]	In December 1988 Pierre Beregovoy agreed in principle to a state owned insurance company's acquiring a majority holding in astate owned bank.
229	123	Portugal	Despite the PSD government's stated commitment to the Maastricht treaty convergence criteria for EMU, little progress is expected to be made in reducing the budget deficit in 1995. The target agreed by the government with the EU is 5.8% of GDP, which is in line with last year's outturn of 5.9%.	The target agreed by the government with the EU is 5.8% of GDP, which is in line with last year's outturn of 5.9%.
265	140	Spain	The government has now set up an export promotion programme and this should have some impact in 1987. A weakening of the peseta against EC currencies is probable.	The government has now set up an export promotion programme and this should have some impact in 1987.

The original paragraph structure, which consisted of the reforms and their context, allowed us to extract additional mention pairs from the same paragraph, detect additional mentions from the EIU reports, and prepare a multiple-mention dataset for our disambiguator model.

For example, when we found that specific reforms were addressed to different actors, we split the reforms to generate multiple mentions of the reform, as depicted in *Table 4*.

Table 4. Annotated Dataset, example

Row ID	Reform ID	Country	Description of Measure	Actual Reference
13	212	Austria	Austria will apply EU customs rates in trade with non-EU countries as of 1995. [This means that the average for customs duties on goods from non-EU countries, currently 10.7%, will be lowered to the EU average of 7.3%. Austria's average customs duty of 9.8% for manufactured goods will be lowered to the new EU average of 3.9%. This reduction should not be too significant, however, as the exemptions applied will bring the actual duty for manufactured goods down to only 6.2%.] [...]	This means that the average for customs duties on goods from non-EU countries, currently 10.7%, will be lowered to the EU average of 7.3%
14	212	Austria	Austria will apply EU customs rates in trade with non-EU countries as of 1995. [This means that the average for customs duties on goods from non-EU countries, currently 10.7%, will be lowered to the EU average of 7.3%. Austria's average customs duty of 9.8% for manufactured goods will be lowered to the new EU average of 3.9%. This reduction should not be too significant, however, as the exemptions applied will bring the actual duty for manufactured goods down to only 6.2%.] [...]	Austria's average customs duty of 9.8% for manufactured goods will be lowered to the new EU average of 3.9%
184	129	Portugal	A target of Esc190bn in privatisation revenue has been set for 1995 including revenue from the sale of minority stakes in [Portugal Telecom and] Electricidade de Portugal.	A target of Esc190bn in privatisation revenue has been set for 1995 including revenue from the sale of minority stakes in Electricidade de Portugal.
185	129	Portugal	A target of Esc190bn in privatisation revenue has been set for 1995 including revenue from the sale of minority stakes in [Portugal Telecom and] Electricidade de Portugal.	A target of Esc190bn in privatisation revenue has been set for 1995 including revenue from the sale of minority stakes in Portugal Telecom.
193	137	Spain	In November the minister of public works, Mr Coscullela, announced that during 1987-91 Pta800 bn ($6.1 bn) is to be spent on modernising and extending the road network. Expenditure of Pta105-100 bn is planned for 1987. An additional Pta25 bn ($190 mn) is to be spent on improvements in ports and coastal areas.	In November the minister of public works, Mr Coscullela, announced that during 1987-91 Pta800 bn ($6.1 bn) is to be spent on modernising and extending the road network.
194	137	Spain	In November the minister of public works, Mr Coscullela, announced that during 1987-91 Pta800 bn ($6.1 bn) is to be spent on modernising and extending the road network. Expenditure of Pta105-100 bn is planned for 1987. An additional Pta25 bn ($190 mn) is to be spent on improvements in ports and coastal areas.	An additional Pta25 bn ($190 mn) is to be spent on improvements in ports and coastal areas

We augmented the annotation by including mentions not coded originally because these were considered duplicates in the projects' original data collection. As shown in *Table 5*, we detected 109 additional mentions to the original 206 policy reforms of the Baseline Database incorporating 14 countries' EIU reports.

Table 5. Distribution of Reforms and their Mentions

Country ID	Country	Issue	Number of Original Reforms	Number of Added Mentions
1	Austria	198605	3	7
2	Belgium	199901	7	9
3	Denmark	199308	21	17
4	Finland	199112	12	5
5	France	198902	11	5
6	Germany	199012	6	14
7	Ireland	199404	17	1
8	Italy	198712	12	2
9	Luxemburg	200109	17	0
10	Netherlands	199506	11	2
11	Portugal	199505	9	7
12	Spain	198701	15	9
13	Sweden	198902	12	1
14	UK	198903	16	2
1	Austria	199408	13	12
6	Germany	200104	17	7
12	Spain	199908	8	6

As earlier highlighted, the final goal of the C3 research project is to count unique policy reforms. In total, we had 313 policy reforms with their corresponding mentions.

According to the analysis, 206 original reforms and 109 mentions were added either by proofreading the EIU reports or by splitting paragraphs referencing multiple reforms. Out of the 206 original reforms, 147 are singletons, meaning these reforms have no mention.

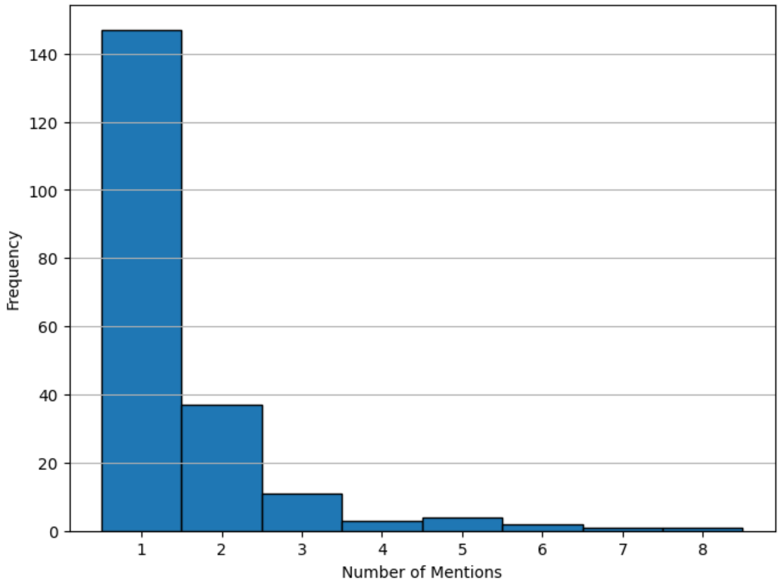


Figure 2. Distribution of Mentions per Policy Reform

The histogram *(Figure 2)* above illustrates the distribution of mentions per policy reform. Most policy reforms (147) have only one mention, reflecting the original dataset's focus on coding only the first mention of each reform. A smaller number of reforms have multiple mentions, with the frequency decreasing as the number of mentions increases. For example, around 40 reforms have two mentions, while a few have as many as five or more. The maximum number was 8 (including the reform and its mentions). This skewed distribution highlights the challenge of developing a robust disambiguation model due to the limited number of multiple mentions, which is crucial for effectively training machine learning models.

Given the high number of singletons, the average number of mentions per reform is approximately 0.53. This indicates that while most reforms are singletons (i.e., they do not have any additional mentions), a smaller number of reforms have multiple mentions, which raises the average slightly above zero. If the total number of mentions (109) were evenly distributed across all reforms (206), each reform would have an average of slightly more than half a mention. Most reforms (about 71%) are singletons, so there are relatively few reforms with multiple mentions, which keeps the average number of mentions low.

The unique reforms also created our analysis's validation and test sets. Since we had 206 unique reforms, we equally divided them into validation and test sets, 50-50%, using the *'Reform ID'*. This is to evaluate the coreference resolution model accurately based on balanced validation and test sets representative of the entire dataset. For the split, we used the train_test_split function from the sklearn.model_selection[5] package, ensuring each entity (policy reform) was assigned exclusively to the validation or the test set:

- test_size=0.5: This parameter indicates that 50% of the data should be allocated to the test set.
- random_state=42: Setting a random state ensures the reproducibility of the split, allowing consistent evaluation across different runs.

As a result, we have obtained the following data for the analysis:

- Validation Set Size: 103 instances, out of which 37 reforms and 66 are mentions
- Test Set Size: 103 instances, out of which 69 are reforms and 34 are mentions

Before splitting the datasets, we performed basic text-cleaning operations to ensure the consistency and quality of the data. This included:

1) Text Lowering: Converting all text to lowercase to avoid discrepancies between mentions that differ only in capitalization.
2) Extra Space Removal: Eliminating unnecessary spaces that could interfere with text analysis and model performance.

This dataset, which forms the foundation of this thesis, was again developed through manual annotation following the methodology of the C3 project. We opted for manual revision and coding over automated natural language processing techniques—such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, dependency parsing, or sentiment analysis (Röder et al., 2014; Schmitt et al., 2019; Trask et al., 2015)—due to the small size of the dataset and the need to create a high-quality ground truth that captures the nuanced distinctions between different reforms. The manual annotation process focused on accurately identifying and coding policy reform mentioned in EIU country reports, resulting in a dataset that can serve as a benchmark for evaluating disambiguation models.

While this manual approach ensures a high degree of accuracy and attention to detail, it is both time-consuming and resource-intensive, limiting the scale of the dataset. Given the project's goal of global analysis, this limited scope represents a significant challenge. Furthermore,

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

manual coding introduces the risk of human bias, where individual interpretations of what constitutes a ‘reform’ or a ‘mention’ may vary between coders, potentially affecting consistency and objectivity.

Despite the manual nature of this dataset, it underscores the potential for integrating supervised machine learning and NLP techniques in future work to automate and scale the analysis. By leveraging these technologies, political and social science research could be conducted more efficiently and at a larger scale, addressing some of the limitations currently posed by manual coding (Wankmüller, 2022). This would align with the broader objectives of the C3 project, which aims to create a global database of policy reforms and with machine learning researchers working to enhance model validity and reliability.

In conclusion, while our dataset provides a solid foundation, it has limitations. The number of policy reforms with multiple mentions is notably low, restricting the model's ability to learn from repeated reference patterns effectively. Expanding the dataset through further manual revision of additional national reports or identifying policy reforms mentioned across multiple country reports would be crucial in enriching the data and improving model performance.

5. Methods

Coreference resolution is applied in NLP for entity detection, which involves identifying all mentions of entities in a text and clustering them into equivalence classes (Pradhan et al., 2012). Although our method is not a traditional coreference resolution, it aims to address a similar challenge of identifying and clustering references—specifically, mentions of the same entity, in this case, policy reforms. Unlike coreference resolution models that eliminate singleton clusters from their dataset, we retained them in our dataset due to their small size. One task to evaluate the policy disambiguator systems was manually detecting mentions by reviewing a selected number of EUI reports and adding additional mentions to our dataset, as described in the Data Chapter.

Next, we implemented a baseline model to develop an appropriate clustering technique necessary for the subsequent analysis. This baseline model played a dual role: firstly, it established a point of comparison for model performance. Secondly, it was instrumental in devising the clustering technique to create ground truth and predicted clusters. In contrast to the annotated dataset, which aimed to identify multiple mentions of policy reforms, the baseline model initially treated all mentions as singletons. This simplification was essential for developing the clustering method that would later be applied to the more complex annotated dataset and to enable the use of the Link-Based Entity-Aware (LEA) metric for evaluation. Therefore, the baseline's clustering technique allowed us to align the data format with the requirements of the LEA metric.

Aligned with the C3 project's overarching goal, the disambiguator system is developed as a baseline model to guide future modeling attempts (Lai et al., 2022; Bengtson & Roth, 2008). Using this clustering technique, we applied an 'all-MiniLM-L6-v2' pre-trained sentence encoder to disambiguate mentions without manual labels automatically. This process began by generating embeddings for each mention using the pre-trained sentence encoder and calculating the cosine similarity between these embeddings. A threshold was then applied to cluster the mentions based on their similarity scores. Consequently, this resulted in the creation of predicted clusters that could be evaluated against the ground-truth clusters.

Finally, the effectiveness of the disambiguation model was evaluated using the LEA scorer. We determined the best threshold using the validation set and then applied this threshold to the test set to assess whether the model improved. The model's precision, recall, and F1 scores were evaluated by comparing the predicted and ground-truth clusters, and a qualitative assessment was conducted to understand the nuances of the model's performance. Additionally, we calculated the Jaccard index on the test set to provide a more holistic evaluation of the clustering performance. This additional metric helped to understand the degree of overlap between the predicted and golden clusters, complementing the insights provided by the LEA metric. This comprehensive approach provided insights into the disambiguation model's performance and highlighted the complexities of policy reform disambiguation in text analysis.

An extended analysis was conducted to simulate a real-world scenario where manual annotation becomes impractical due to the volume of data collected globally from various countries and languages. In such a scenario, the system should be able to extract policy reforms while disambiguating them from further mentions, facilitating the counting of unique policy reforms. Thus, this study's disambiguator system demonstrates how effectively pre-trained models can be applied compared to manual annotation and whether automated methods could be applied to a large dataset, reducing the need for extensive human resources (Pilny et al., 2024). Given the small dataset used in this study, it is sensible to forego a training phase in the baseline model. Training on small datasets often leads to overfitting and poor generalization, resulting in unreliable outcomes. Establishing a baseline ensures that future work with larger datasets can involve proper training and fine-tuning of models, providing a foundation for ongoing research.

5.1 Baseline Model

The baseline model established a foundational clustering methodology for the ground truth and prediction datasets. This model used a simple unsupervised approach to group mentions, which was essential for evaluating the model's performance using the Link-Based Entity-Aware (LEA) metric—a metric typically utilized in coreference resolution tasks. We could assess how well the model differentiated between singleton mentions and multi-mention clusters within policy reform texts by implementing the baseline clustering technique.

To gauge the performance of this baseline approach, an experiment was conducted using a dataset that contained only singleton clusters. The results from this experiment were as follows *(Table 6)*:

Table 6. Baseline Model Performance

True Positives (TP)	False Positives (FP)	False Negatives (FN)	Precision	Recall	F1 Score
4,155	35,194	0	0.1056	1.0	0.1910

These results are expected, given that the dataset was composed entirely of singleton clusters, where each mention was treated as a separate entity. The perfect recall (1.0) indicates that the baseline correctly identified every mention as a unique cluster. However, the low precision (0.1056) arises not from the model's inability to group related mentions but from the dataset's inherent design, leading to many false positives. In this scenario, every mention is a false positive in the context of clustering since they all represent unique clusters. Consequently, the low F1 score of 0.1910 reflects this approach's limitations when applied to more complex clustering tasks.

This outcome serves as a reference point for evaluating more sophisticated models. We specifically assess whether our disambiguation system, which creates multi-mention clusters using pre-trained sentence embeddings, can surpass the performance of the all-singletons approach. This comparison is crucial for the analysis section, where we reflect on the model's effectiveness relative to this baseline.

5.2 Ground Truth / Golden Mention Clusters

We used the Dataset Row ID and Reform ID to identify belonging reforms and their mentions. Each row served as a unique identifier, simplifying the tracking and counting of mentions. To cluster these mentions, we implemented a for loop. The clustering of golden mentions happened because of the reforms' manual annotation, as presented in the Data chapter.

We developed a clustering technique. By matching the IDs, we simulated coreference resolution. Specifically, mentions identified by the *'Row ID'* corresponding to the same policy *'Reform ID'* were clustered together. The format of a list of sets allowed us to cluster mentions referring to the same policy reform effectively. However, as previously mentioned, because only the first mention of each reform was coded initially, the initial dataset consisted solely of singleton entities. Given our dataset's structure, where each mention is uniquely associated with a policy reform without duplication, we anticipated forming clusters where each cluster corresponds to a single entity.

This technique was then used to extract the golden mention clusters, which serve as the benchmark for evaluating the model's performance. This involved grouping the list of mentions that refer to the same entity with their assigned policy reform IDs, which were provided through further annotations.

We created both golden and prediction clusters to assess our model's performance. The golden clusters served as the single source of truth, essential for calculating accuracy metrics such as precision, recall, and F1 score. They also facilitated the assessment of the model's performance on validation and test datasets (Wiebe et al., 1999).

The key objective was to cluster mentions referring to the same policy reform, thereby organizing them to facilitate practical data analysis. For instance, in `x = [{1, 2}, {3, 4, 6}] `, we identify two unique policy reforms: the first referenced by mentions 1 and 2, and the other referenced by mentions 3, 4, and 6. This requirement dictated how we organized our transformed data, ensuring the coreference evaluation functions could effectively process it.

Equation 4. Clustering of Reforms and Mentions

```
• Initialize $G$ as an empty dictionary.
• For each row in DataFrame $D$:
    o Assign y to $D$['ID'] value of the current row.
    o Assign x to $D$['Row ID'] value of the current row.
    o If y is not a key in $G$:
        ▪ Set $G$[y] to a set containing x.
    o Else:
        ▪ Add x to the set $G$[y].
• Output the dictionary $G$.
```

First, we initialized an empty dictionary G to iterate over the DataFrame. The dictionary G maps each unique *'Reform ID'* (y) to its associated *'Row ID'* (x) values. This way, G categorizes all *'Row ID'* values into unique sets based on each unique *'Reform ID'*. This structure is fundamental for subsequent evaluations using the LEA metric.

5.3 Prediction Dataset

Given our project's constraints, we opted for a text-based similarity approach without an explicit training phase due to the dataset's small size. We implemented the following steps:

1) Pre-Trained Sentence Encoder for Text Embeddings:

We used a pre-trained sentence encoder (all-MiniLM-L6-v2) to generate text embeddings for each mention, representing mentions in a high-dimensional vector space without additional training. The bidirectional nature of the model enables the capture of contextual information from both directions (left-to-right and right-to-left), ensuring a comprehensive understanding of each mention's context.

2) Cosine Similarity Calculation:

After creating the embeddings, we calculated the cosine similarity between all pairs of mentions. Cosine similarity measures the cosine of the angle between two vectors, providing a similarity score from -1 (exactly opposite) to 1 (identical) (Reimers & Gurevych, 2019). Higher scores indicate more significant similarity.

3) Threshold-Based Clustering:

We applied a threshold-based clustering method to form clusters of mentions referring to the same policy reform. Mentions with similarity scores equal to or exceeding the predefined threshold were clustered, while mentions below the threshold remained singletons. We implemented a merge process to ensure that a mention did not appear in more than one cluster. This involved merging clusters with a non-empty intersection to unify mentions into one cluster.

The clustering process unfolds as follows:

1. **Initial Clustering**: Mentions are initially grouped based on the threshold criterion. If a mention has similarity scores with other mentions that exceed the threshold, it can belong to multiple clusters.
2. **Cluster Merging**: We utilize the merge_clusters function to merge any clusters that share common mentions. This function operates by checking for intersections between clusters and merging those with any overlap. Merging continues iteratively until no further merges are possible, ensuring that each mention belongs uniquely to one cluster.

This method ensures robustness in our clustering approach by refining the groupings and effectively handling overlaps, forming accurate clusters of policy reform mentions. The threshold was tuned using the validation set to ensure optimal clustering performance.

5.4 Evaluation

Given Dataset's small size, we condensed the typical train-validate-test cycle into a validate-test cycle. This approach allowed us to focus on validating and testing the model without implementing a training phase, which would be impractical with the limited data available.

To identify the optimal threshold for clustering, we employed the Link-Based Entity-Aware (LEA) metric on the validation set and implemented it using Python. The LEA metric provides a comprehensive evaluation by considering both precision and recall in the context of coreference resolution. After determining the best threshold, we applied this threshold to the test set to evaluate the final performance of our model.

All proposed evaluation metrics for coreference resolution use recall, precision, and F1 to report the performance of a coreference resolver. Recall measures the fraction of correct coreference information (e.g., coreference links or entities) that is resolved. Precision measures the fraction of resolved coreference information that is correct. F1 is the weighted harmonic mean of recall and precision (Moosavi & Strube, 2016). While F1 is typically used to compare coreference resolution systems, it is also essential that the corresponding recall and precision values are interpretable and discriminative.

The LEA metric takes a list of predicted clusters of mentions and a list of ground truth clusters of mentions, outputting a score indicating how well the predicted clusters correspond to the ground truth clusters. This metric is helpful in our evaluation as it allows us to quantify the accuracy and effectiveness of our pre-trained disambiguation model compared to the manually annotated ground truth.

In addition to the LEA scorer, the Jaccard similarity index was included to provide a more comprehensive assessment of the model's performance. While the LEA metric offers detailed insights into precision and recall, the Jaccard index measures the overall overlap between predicted and golden clusters, offering a complementary perspective on the model's efficacy.

Overview of the LEA Metric

The LEA metric is designed to address the limitations of traditional evaluation metrics in coreference resolution like MUC and B3, CEAF, and BLANC by considering the importance and resolution accuracy of entities within a text. This metric aims to provide a more discriminative and accurate assessment of coreference resolution systems.

Definition of Importance and Resolution Score

Importance of an Entity: One of the new features of the LEA metric compared to the traditional metric was that LEA computes for each entity how vital the entity is and how well it is resolved (Moosavi & Strube, 2016).

Equation 5. LEA Importance

$$\frac{\sum_{e_i \in E}(\text{importance}(e_i) \times \text{resolution-score}(e_i))}{\sum_{e_k \in E} \text{importance}(e_k)}$$

(source: Moosavi & Strube 2016, p. 636)

- e_i and e_k represent elements in the set E.
- importance(e_i) and importance(e_k) represent the importance of e_i and e_k, respectively.
- resolution-score(e_i) represents the resolution score of e_i.

The importance of an entity is defined primarily by its size, i.e., importance(e)=|e|, where |e| represents the number of mentions within the entity. This gives higher importance to larger entities, e.g., reforms with more mentions. However, the importance metric can be adapted to include other factors, such as entity type or mention types, based on the specific requirements of the domain or task (ibid.).

Resolution Score: For a given key entity ki, the resolution score is calculated as the ratio of the number of correctly resolved coreference links to the total coreference links possible within ki:

Equation 6. LEA Resolution

$$resolution\text{-}score(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}$$

(source: Moosavi & Strube 2016, p. 636)

Where:

- k_i represents a specific element or keyword, in our case mentions.
- r_j represents an element from the set R.
- link(k_i ∩ r_j) represents the link between k_i and the intersection of k_i and r_j.
- link(k_i) represents the link of k_i.

Here, rj represents a response entity, and link($ki\cap rj$) denotes the number of coreference links between ki and rjr that are correctly identified.

1) LEA calculates an entity's "importance" by analyzing the number of mentions within a given text. This approach assumes that the more mentions an entity has, the more significant it is.

2) We determine the number of unique coreference links within an entity by computing the number of potential pairs (links) from n mentions as n(n-1)/2. Singletons (entities with only one mention) are corrected to ensure they are accurately counted.

3) To identify how well a response entity matches the coreference pattern of a key entity, we count the coreference links they have in common using an iterative function that accounts for singletons and multi-mention pairs.

4) The resolution score for a response entity against a set of key entities is calculated as the fraction of correctly resolved coreference links relative to the total links in the response entity. This score is crucial for assessing the accuracy of coreference resolution and evaluating the model's performance.

LEA Equations

LEA Recall:

Equation 7. LEA Recall

$$\text{Recall} = \frac{\sum_{k_i \in K}\left(|k_i| \times \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}\right)}{\sum_{k_z \in K} |k_z|}$$

(source: Moosavi & Strube 2016, p. 636)

Where:

- k_i represents a specific element or keyword in the set K.
- r_j represents an element from the set R.
- link(k_i ∩ r_j) represents the link between k_i and the intersection of k_i and r_j.
- link(k_i) represents the link of k_i.
- |k_i| represents the size or count of k_i.
- |k_z| represents the size or count of k_z.

This formula calculates the LEA recall by considering the weighted sum of correctly resolved links across all key entities, normalized by their total size. Unlike traditional recall, which equally weights each entity, LEA recall assigns weight based on the entity's importance (number of mentions). This approach recognizes that some entities may have greater complexity or more mentions, making them more significant than others. (Bengston & Roth, 2008; Moosavi & Strube, 2016).

LEA Precision:

Equation 8. LEA Precision

$$\text{Precision} = \frac{\sum_{r_i \in R}\left(|r_i| \times \sum_{k_j \in K} \frac{link(r_i \cap k_j)}{link(r_i)}\right)}{\sum_{r_z \in R} |r_z|}$$

(source: Moosavi & Strube 2016, p. 636)

Where:

- r_i represents a specific element or resource in the set R.
- k_j represents an element from the set K.
- link(r_i ∩ k_j) represents the link between r_i and the intersection of r_i and k_j.
- link(r_i) represents the link of r_i.
- |r_i| represents the size or count of r_i.
- |r_z| represents the size or count of r_z.

LEA precision involves computing importance-weighted scores to ensure that all entities are appropriately considered. This calculation considers the number of accurately identified coreference links for each predicted entity while considering its importance. The final precision score is normalized to ensure accuracy and reliability by dividing it by the predicted links.

Handling of Singletons

- LEA accounts for singletons using self-links, where a self-link indicates a mention is coreferent only with itself. This means that if a mention does not refer to any other entity, it is still counted correctly in the evaluation. This approach avoids unfairly lowering the score for mentions that are not part of any larger group.

Advantages of LEA

- **Comprehensive Assessment**: Unlike other metrics that may only consider extra or missing links, LEA evaluates all coreference links, providing a more thorough and reliable measure of system performance.
- **Flexibility in Mapping**: LEA permits one-to-many mappings of entities and rewards all correct coreference relations, making it robust against different types of entity structures.
- **Importance Weighted Evaluation**: LEA differentiates between missing or extra entities based on their importance, providing a nuanced evaluation that recognizes the significance of larger entities.

We computed the Link-based Entity-Aware (LEA) recall and precision scores using the lea_recall and lea_precision functions for each generated cluster set at different thresholds. These functions assess how well the model's predictions (formed clusters) matched against the gold standard dataset (gold_val_res) in accurately identifying and linking mentions within and across documents. Additionally, we calculated the F1 score for each threshold to evaluate the overall model performance.

Using the optimal threshold identified in the previous step, we again utilized our clustering functions but applied them to the test data. This step again involves evaluating each pair of sentences in the dataset and grouping those that meet or exceed the threshold similarity score. The resulting set of clusters (pred_test_res) represents groups of sentences considered similar enough to be related or refer to the same underlying subject.

6. Analysis

Given the small size of the annotated dataset, the baseline model was designed to align with the project's context and data constraints. Although incorporating a training phase was considered, it was omitted to streamline the process, focusing instead on validation and testing. The Link-Based Entity-Aware (LEA) metric evaluated the model's performance, providing a comprehensive measure of the disambiguation process. The evaluation involved comparing the golden clusters from the manually annotated dataset with the predicted clusters generated using pre-trained sentence embeddings, incorporating precision and recall.

The evaluation process focused on fine-tuning the clustering threshold on the validation set to identify an optimal value, which was then applied to the test set to gauge the model's generalization. We evaluated the model using 25 thresholds, from -1.0 to 1.0. As depicted in *Figure 3*, the threshold that yielded the highest F1 score, 0.9167, was considered the optimal threshold for the test set.

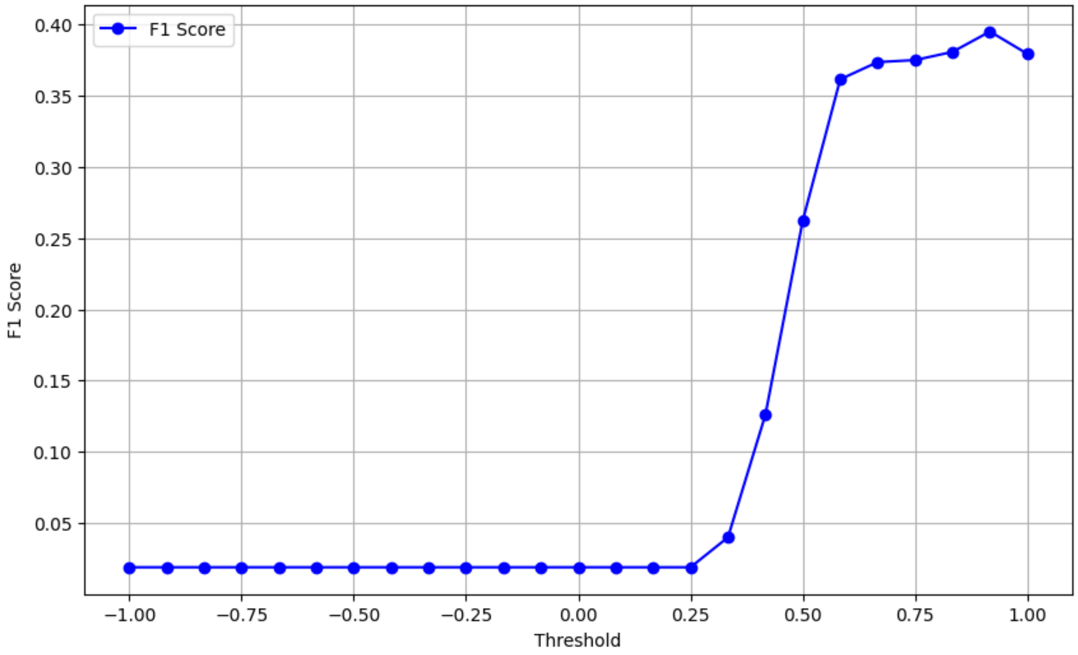


Figure 3. Best Threshold

The results of the validation and test sets, at the 0.9167 best threshold, were the following:

Table 7. Summary of the Results at the Best Threshold

	Validation Set	Test Set
LEA Recall	0.3879	0.5507
LEA Precision	0.4023	0.5507
F1	0.3950	0.5507

The observed improvements in the LEA Recall, Precision, and F1 Score from the validation to the test sets indicate that the model generalizes somewhat effectively. As the threshold increases, the model better captures singleton clusters but struggles with clusters containing multiple mentions. This pattern remains consistent when comparing the validation and test sets.

A notable trend is that lower thresholds result in larger single clusters, while higher thresholds lead to increased singletons, as observed previously. Interestingly, despite lower F1 scores between thresholds of 0.5 and 0.583, the model's clustering aligns better with the reference clusters at these thresholds, successfully capturing some clusters with multiple mentions. However, this capability diminishes at higher thresholds, including the best-considered threshold.

Given the complex nature of clustering policy reforms, relying solely on quantitative metrics may not fully reveal how well the model aligns with the underlying structure of the data. Therefore, it is crucial to complement our quantitative evaluations with qualitative analysis. This holistic approach allows for a deeper exploration of how the model's predictions reflect actual policy reforms and highlights specific cases where the model successfully or unsuccessfully captures the intended relationships within the data. Accordingly, a qualitative review could help us surpass the discrepancies between the golden and predicted clusters.

To illustrate why the model fails to cluster the policy reforms and their mentions correctly, we will demonstrate some correctly and incorrectly identified clusters, as shown in *Tables 8, 10*, etc. We also provide the corresponding actual references with a qualitative analysis in *Tables 9, 11*, etc.

6.1 Validation Set Results

Table 8. Validation Set Results

Threshold	LEA Recall	LEA Precision	F1 Score
-1.0000	0.6207	0.0096	0.0190
-0.9167	0.6207	0.0096	0.0190
-0.8333	0.6207	0.0096	0.0190
-0.7500	0.6207	0.0096	0.0190
-0.6667	0.6207	0.0096	0.0190
-0.5833	0.6207	0.0096	0.0190
-0.5000	0.6207	0.0096	0.0190
-0.4167	0.6207	0.0096	0.0190
-0.3333	0.6207	0.0096	0.0190
-0.2500	0.6207	0.0096	0.0190
-0.1667	0.6207	0.0096	0.0190
-0.0833	0.6207	0.0096	0.0190
0.0000	0.6207	0.0096	0.0190
0.0833	0.6207	0.0096	0.0190
0.1667	0.6207	0.0096	0.0190
0.2500	0.6207	0.0096	0.0190
0.3333	0.5747	0.0207	0.0400
0.4167	0.5142	0.0718	0.1260
0.5000	0.4168	0.1916	0.2625
0.5833	0.3737	0.3500	0.3615
0.6667	0.3707	0.3764	0.3735
0.7500	0.3707	0.3793	0.3750
0.8333	0.3764	0.3851	0.3807
0.9167	0.3879	0.4023	0.3950
1.0000	0.3793	0.3793	0.3793

Threshold: -1.0 to 0.25

In the thresholds from -1.0 to 0.25, all the mentions were clustered into one large cluster. This means the model considered all policy mentions to be part of the same reform; thus, everything is considered one big reform for these thresholds. This over-clustering reflects a typical issue in clustering models when thresholds are too low, causing the model to conflate separate entities due to a lack of sufficient distinction between them.

Thresholds from 0.3333 to 0.5

Table 9. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster[6]
{16, 17}		{16, 17}
	{55}	{51, 52, 53, 54, 55, 88}
	{120}	{120, 121}
	{267, 253}	{251, 252, 253, 254} & {267, 255}

Fewer mentions are grouped into one big cluster, and more smaller clusters emerge. However, the maximum number of elements in a cluster at this threshold was 3, far from the 5-8 element clusters seen in the golden clusters. While the clusters are becoming more refined, they still do not match the golden clusters. For example, the disambiguator begins separating policy mentions but often overcompensates, segmenting mentions that should be grouped. Clusters like {106, 107, 108, 109} remained fragmented. The cluster {16, 17} was correctly identified as a small, isolated cluster, but clusters like {55} and {120} were isolated wrongly. Partially formed clusters like {267, 253} show the beginning of meaningful separation but still do not align with the golden clusters.

[6] We refer with correctly identified clusters to the case where the predicted clusters match with those in the golden clusters. Consequently, we name incorrectly identified cluster when the model made up new clusters or separated clusters that do not occur in the golden clusters. This depiction of the results follows also in the following tables.

Table 10. Qualitative Analysis of Validation Set Results

Row ID	Actual Reference	Qualitative Analysis
{16, 17} /predicted and golden/	The report claims that in 1986 Mr Vranitzky, then minister of finance, issued a state guarantee for the purchase of a passenger ship, Mozart by the DDSG the national Danube ship company. {16} "It was based on the public auditor's report, but Mr Vranitzky was not found to be at fault, nor was it established that any money flowed to hirn or the SPO in connection with the purchase of Mozart." {17}	Both sentences prominently mention "Mr. Vranitzky" and the purchase of the "Mozart" ship, creating a clear entity-based link. The sentences present a natural flow of events, from the state guarantee issuance to the public auditor's findings, maintaining a coherent narrative. Both sentences revolve around the same thematic elements, like the "state guarantee" and the "public auditor's report."
{55} /predicted/ vs. {51, 52, 53, 54, 55, 88} /golden/	"The leader of the Christian People's Party, the energy minister, Jan Sjursen, says the whole project should be reconsidered." {55} "It has criticised the manner in which funds have been distributed to alleviate the environmental impact of building the road and rail link across the Oresund from Copenhagen to Sweden." {51} "A special meeting has been called between the prime minister, Poul Nyrup Rasmussen, and his Swedish counterpart, Carl Bildt, to discuss the road and rail link planned between Copenhagen and Sweden." {52} "Excavation for the new route has started, but several scientists have expressed the fear that laying a stretch of tunnel directly on the sea bed might adversely affect the salinity of the Baltic Sea." {53}	The other mentions ({51}, {52}, {53}) focus specifically on the Oresund infrastructure project, discussing its environmental impact, political discussions, and scientific concerns. They explicitly mention the project's key figures, environmental consequences, and specific political actions. In contrast, {55} speaks generally about "the whole project" without directly referencing the Oresund project or its related issues. It lacks the explicit thematic connection in the other mentions, such as direct references to government figures, scientific fears, or environmental impact. Its broader, more financial tone and lack of direct connection could have led the model to isolate it from the other mentions. By focusing on the more general nature of {55} and its lack of explicit references, this explanation more accurately reflects why the model struggled to cluster it with the other, more thematically consistent mentions.
{120} /predicted/ vs. {120, 121} /golden/	"In December 1988 Pierre Beregovoy agreed in principle to a state owned insurance company's acquiring a majority holding in a state owned bank." {120} "Thus Gan, France's fourth largest insurance company (after UAP, Axa-Midi and AGF), was authorised to raise from 34 per cent to 51 per cent its holding in Credit Industriel et Commercial (Cie) - with the French state's holding being correspondingly reduced from 66 per cent to 49 per cent." {121}	Both mention state-owned insurance companies increasing their holdings in financial institutions, involving government approval for changes in ownership. However, they differ in the level of specificity provided. Mention {120} gives a broad statement about an acquisition agreement in principle, without naming specific companies or detailing the changes in holdings. In contrast, mention {121} offers a specific account, identifying "Gan" and "Credit Industriel et Commercial (Cie)" and detailing the exact changes in percentages of holdings.

We could interpret these results to suggest that the model detects semantic shifts or structural cues (e.g., changes in topic or entity) but cannot consistently distinguish related but not identical mentions. For example, the case of mention {55} showcases how the model misses subtle connections when small shifts in narrative or emphasis (e.g., scientific vs financial focus) could mislead the model into incorrect clustering.

Thresholds 0.5

Table 11. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster
	{160, 161, 99, 100}	{99}, {100}, {160}, {161}
	{214, 201, 202}	{201}, {202}, {214}

This threshold represents a significant jump in forming individual clusters. Many entities start forming clusters or small groups that resemble the golden clusters. LEA Recall decreases to 0.4087 as the threshold increases, while LEA Precision improves, reaching up to 0.1905. This trade-off suggests that the clusters formed at this threshold are more meaningful, albeit at the cost of excluding relevant mentions (lower recall). New clusters like {160, 161, 99, 100} or {214, 201, 202} do not exist in the golden clusters.

Table 12. Qualitative Analysis of Validation Set Results

Row ID	Actual Reference	Qualitative Analysis
{160, 161, 99, 100} /predicted/	"There will, however, be a temporary increase in employees' social security contributions in order to maintain unemployment benefit at current levels, and cuts in sickness insurance benefits. "{99} “Changes are to be made to the employment act, legally binds the government to prevent long-term unemployment.” {100} “The employers, for their part, had been incensed by the government’s bill for the tightening of fixed-term contract provisions.” {160} "With regard to the contentious bill to amend the 1972 Works Council Act and extend the role of works councils, an earlier argument between the employment interests of the Mittelstand (medium-sized companies), 11 were adopted in the revised bill has now been placed before parliament. " {161}	99 and 100 mentions are from the Finnish reports, and 160 and 161 are from the German reports. Even though the four mentions all discuss employment-related issues, this clustering could also be an example of over-generalization by the model since it cannot capture the specificities between legal topics, such as social security contributions vs. employment contracts. Furthermore, this result also points toward considering how a later model will overcome the issue of counting unique reforms in the same country report without grouping mentions from different country reports.
{214, 201, 202} /predicted/	"Earlier this year both the International Labour Organisation and the OECD produced reports arguing that the Luxembourg pension system is not viable in the long term. However, on July 16th a government roundtable reached agreement to raise all state-financed pensions in the private sector by 3.9%." {201} "Earlier this year both the International Labour Organisation and the OECD produced reports arguing that the Luxembourg pension system is not viable in the long term. However, on July 16th a government roundtable reached agreement to raise the basic pension by 4.9%, with additional bonuses for each year worked." {202} "On July 16th agreement was reached to increase private-sector pensions with as from 2002, to bring them into line with state-sector pensions larger increases for minimum and widows’ pensions." {214}	All three mentions reference the International Labour Organisation, the OECD, and the Luxembourg pension system. These shared entities provide a strong link, which the model likely used to connect the sentences. Additionally, all three focus on July 16th, representing a specific event that ties the mentions together, making it logical for the algorithm to identify them as part of the same topic. Accordingly, it could be well derived why the model could have decided to cluster these together, which makes sense from a thematic standpoint, as all sentences discuss Luxembourg pension reform within the same time frame and policy context. There are only very slight nuances, such as the differences between the state vs. private sector, that might have been too specific for the model.

Accordingly, this threshold demonstrates a more rigid clustering approach, forming smaller and more precise clusters. However, this comes at the cost of excluding some related mentions, indicated by a decrease in recall.

Thresholds from 0.5833 to 0.75

Table 13. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster
{42, 43, 44}		{42, 43, 44}
	{106}, {107}, {108}, {109}	{106, 107, 108, 109}

This conservative approach prevents the formation of larger, more cohesive clusters that align with the golden clusters, such as {106, 107, 108, 109}.

Table 14. Qualitative Analysis of Validation Set Results

Row ID	Actual Reference	Qualitative Analysis
{106}, {107}, {108}, {109} /predicted/ vs. {106, 107, 108, 109} /golden/	"Michel Rocard announced in early February that some Fr10 bn ($1.6 bn) of budgeted government expenditure would be temporarily frozen." {106} "What appears from the new draft is the importance of fiscal reform in the government's overall designs." {107} "West Germany is the only exception, admittedly a most important one for France, with a sharp deceleration in 1989 (as a result of a fiscal squeeze) giving way to some recovery in 1990 as scheduled tax concessions are made." {108} "If the government feels that it cannot get the growth of private consumption under proper control with monetary policy, then it will act by fiscal policy instead - as witness Mr Rocard's February announcement." {109}	While the researcher manually added mentions {107, 108, 109} to form a broader cluster around fiscal policy, the model did not cluster them together, which can be considered a reasonable decision. The model detected superficial differences among monetary policy, tax concessions, and country-specific references, thus missing the overarching connection. This indicates a limitation in the model's ability to capture more nuanced and thematic links between policy discussions.
{42, 43, 44} /predicted and golden/	"During the same period there was an even larger improvement in the primary surplus (deficit excluding interest payments) which increased from 2.3% of GDP in 1993 to 6.1% in 1998. To achieve this result the government increased taxes." {42} "During the same period there was an even larger improvement in the primary surplus (deficit excluding interest payments) which increased from 2.3% of GDP in 1993 to 6.1% in 1998. To achieve this result the government reduced expenditure growth." {43} "During the same period there was an even larger improvement in the primary surplus (deficit excluding interest payments) which increased from 2.3% of GDP in 1993 to 6.1% in 1998. The increase in the tax burden was underpinned by higher indirect rather than direct taxes." {44}	The model correctly clustered these mentions together but not consistently across all thresholds. At lower thresholds, the shared theme of improvements in the primary surplus kept them grouped. At higher thresholds, the model splits them due to subtle differences, like focusing on taxes rather than expenditure cuts. This illustrates how higher thresholds can lead to over-segmentation, separating mentions based on minor details rather than the overall theme.

At these thresholds, the entities {42, 43, 44}, previously treated as standalone singletons at threshold 0.4167, merged at threshold 0.75. However, despite this merging, the model still predominantly treats most entities as singletons or very small clusters, indicating an overly conservative clustering approach. Thus, the model had mixed success in clustering related mentions across different thresholds. It correctly grouped broader themes at lower thresholds, but at higher thresholds, it tended to over-segment due to subtle differences. This highlights the challenge of balancing overarching themes and nuanced variations. While some clusters were identified correctly, consistency across thresholds remained a challenge, indicating the need for further refinement.

From Threshold 0.8334 to Threshold 1.0:

(Best Threshold: 0.9166666666666665)

At higher thresholds, specifically from 0.8334 to 1.0, the LEA Recall and Precision metrics converge. At the extreme threshold of 1.0, both LEA Recall and Precision reach 0.3771, along with the F1 score. This convergence indicates that the model becomes increasingly conservative in forming clusters, often erring on the side of caution. However, this conservatism leads to a significant issue: it results in many singleton clusters, missing out on forming valid, multi-mention groups. At the maximum threshold of 1.0, the model outputs only singleton clusters.

These findings emphasize the necessity of choosing a threshold that aligns well with the model's capabilities and the inherent data structure to optimize clustering outcomes. Lower thresholds tend to compromise precision by over-clustering mentions into overly broad groups. In contrast, higher thresholds lead to excessive fragmentation, separating mentions that should be grouped.

The golden clusters serve as a benchmark for evaluating the model's performance. Close inspection of thresholds around 0.4167 to 0.5 reveals a more balanced approach. Despite identifying an optimal threshold through quantitative analysis, the qualitative review suggests that the model still struggles to balance overgeneralization and over-segmentation effectively. This indicates a need for more sophisticated clustering techniques.

Additionally, a pre-trained sentence encoder may not capture the nuanced relationships between mentions without additional tuning due to context variations. This limitation highlights the potential benefit of more advanced methods, such as training on task-specific data or employing more context-aware models. Such improvements could help the model better grasp the complexity of relationships between mentions, leading to more accurate clustering outcomes.

6.2 Test Set Results

Table 15. Test Set Results

Threshold	LEA Recall	LEA Precision	F1 Score
-1	0.413	0.0061	0.0121
-0.917	0.413	0.0061	0.0121
-0.833	0.413	0.0061	0.0121
-0.75	0.413	0.0061	0.0121
-0.667	0.413	0.0061	0.0121
-0.583	0.413	0.0061	0.0121
-0.5	0.413	0.0061	0.0121
-0.417	0.413	0.0061	0.0121
-0.333	0.413	0.0061	0.0121
-0.25	0.413	0.0061	0.0121
-0.167	0.413	0.0061	0.0121
-0.083	0.413	0.0061	0.0121
0	0.413	0.0061	0.0121
0.0833	0.413	0.0061	0.0121
0.1667	0.413	0.0061	0.0121
0.25	0.413	0.0061	0.0121
0.3333	0.3986	0.0061	0.012
0.4167	0.3478	0.085	0.1366
0.5	0.3302	0.2353	0.2748
0.5833	0.4498	0.3904	0.418
0.6667	0.4425	0.4565	0.4494
0.75	0.4928	0.4855	0.4891
0.8333	0.4928	0.4855	0.4891
0.9167	0.5507	0.5507	0.5507
1	0.587	0.587	0.587

Thresholds from -1.0 to 0.25

The model exhibited severe over-clustering at these lower thresholds, similar to the patterns observed in the validation set, grouping all mentions into a single large cluster.

In this range, both LEA Recall and Precision remained low, with a recall of 0.413 and a precision of 0.0061, resulting in an F1 score of 0.0121 across these thresholds. This tendency to over-cluster at lower thresholds suggests that the model struggles to identify the fine-grained differences necessary for accurate disambiguation when the similarity threshold is too lenient. As a result, the model lumped distinct policy reforms into overly broad clusters, leading to poor evaluation metrics.

Threshold from 0.3333 to 0.5

Table 16. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster
	{160, 161, 99, 100}	{99}, {100}, {160}, {161}
	{214, 201, 202}	{201}, {202}, {214}
	{224, 225, 219, 220}	{219}, {220}, {224}, {225}
{142, 143}		{142, 143}

In this threshold range, the model begins to show some segmentation in the test set, similar to the behavior observed in the validation set. However, these clusters are still improperly segmented, indicating that the model relies heavily on shallow semantic features, such as recurring phrases or structural patterns, rather than a deeper understanding of the context. For instance, clusters like {224, 225, 219, 220} are incorrectly merged due to phrases like "government compromised," leading to the incorrect assumption that these mentions are part of the same reform despite different contextual backgrounds. While the cluster {142, 143} is correctly identified from threshold 0.4167 to 0.75, it eventually becomes segmented at higher thresholds, demonstrating the model's inconsistency in capturing nuanced relationships, and it is a somewhat random phenomenon that the model captures the thematic and contextual connection among the mentions.

Table 17. Qualitative Analysis of Test Set Results

Row ID	Actual Reference	Qualitative Analysis
{224, 225, 219, 220} /predicted/ vs. {219}, {220}, {224}, {225} /golden/	“The plan provided for a loosening of the government's tight grip on the housing market by cancelling some G30bn ($18bn) of government loans to public housing corporations.” {219} “The plan provided for a loosening of the government's tight grip on the housing market by allowing them to operate more freely on the market.” {220} “In April the government compromised: it will devote G2.25bn to subsidising labour costs at the lowest wage levels.” {224} “In April the government compromised: while G2.25bn is to be allocated for financing a cut in the siphoning premium.” {225}	The first two sentences discuss loosening the government's control over the housing market but refer to different actions (canceling loans vs. allowing more market freedom). The last two sentences mention financial compromises in April—one about subsidizing labor costs and the other about financing a cut in the siphoning premium. The recurring expressions "government compromised" and "plan provided for" might have led to superficial semantic clustering due to their similar syntactic structures.
{142, 143} /golden/	“Mr Eichel has proposed the creation of a new Federal Agency for Financial Market Supervision, which would amalgamate the three current supervisory agencies responsible for banking, insurance and securities markets.” {142} “In February the finance minister, Hans Eichel, presented a proposal for an ambitious reform of the supervision of financial markets, which would merge responsibility for the supervision of banking and insurance companies into a single organisation." {143}	The model correctly identified these mentions as belonging to the same reform. Both sentences discuss the creation of a new supervisory agency for financial markets in Germany. They revolve around the same thematic elements, such as amalgamating current supervisory agencies into one organization. This indicates that the model can accurately capture thematic and entity-based links with solid and consistent contextual cues.

In summary, while the model still struggles with nuanced distinctions at this threshold, as seen in the incorrect merging of clusters like {224, 225, 219, 220}, it shows the capability to correctly identify and cluster thematically related mentions in some cases, such as {142, 143}. This suggests the model can identify clusters accurately when the context and semantic connections are clearer and more explicit. However, it still tends to rely on recurring phrases and shallow semantics, which can lead to over-segmentation and incorrect merging in more complex or subtle cases.

Threshold 0. 5834 to 0.75

Table 18. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster
	{181, 183, 63}	{181}, {183}, {84, 85, 87, 62, 63}
	{145, 66, 311}	{66}, {145}, {311}
{245, 246}		{245, 246}

As the threshold moves into this range, the model starts to break up large clusters and form smaller groups. However, similar to the validation set, these clusters still do not fully align with the golden clusters. For example, mentions like {181, 183, 63} are grouped due to shared references to vehicle-related financial regulations, despite differences in their national contexts (Denmark vs. Netherlands). Conversely, the cluster {245, 246} was incorrectly clustered with {257} at threshold 0.5 but is now correctly clustered at thresholds 0.5834 and 0.6667.

Table 19. Qualitative Analysis of Test Set Results

Row ID	Actual Reference	Qualitative Analysis
{181, 183, 63} /predicted/ vs. {181}, {183}, {84, 85, 87, 62, 63} /golden/	"Registration tax on trucks will be more than doubled, and taxes increased on other vehicles." {63} “In addition to greater confidence in the economy, SIMI points to the reduction in vehicle registration tax in the budget.” {181} “In addition to greater confidence in the economy, SIMI points to allowable expenses for company cars.” {183}	There are overlapping references to vehicles and taxation policies. The semantic similarity around the concept of vehicle-related financial regulations might lead the model to group them, even though they address different national reports in Denmark and the Netherlands or tax directions. The reason for this clustering, therefore, might lie in the similarities of economic concepts.
{245, 246} /golden/	“Although a further 12.5 per cent tariff reduction with the EC took place on January 1, 1987, the impact on imports should be slightly less than in 1986, because of the probable weakening of the peseta against EC currencies.” {245} “The three most important influences on the trade account in 1986 were Spain's entry into the EC (and the subsequent reduction of trade tariffs), the fall in the price of energy imports and many raw materials, and strong appreciation of the peseta against the dollar (more than 60 per cent of Spainfs imports are paid for in dollars)." {246}	The model correctly identified this cluster. Both sentences discuss the economic implications of Spain's trade policy, specifically the influence of EC tariff reductions on Spain's trade account and imports. This correct clustering indicates that when the contextual connections are more explicit and direct, the model can identify and group them correctly.

In summary, the model still struggles to accurately separate mentions when subtle distinctions, such as different national contexts, are involved, as demonstrated by the incorrect clustering of {181, 183, 63}. However, the model can correctly identify clusters when the thematic elements are more straightforward and directly connected, as seen with the correct clustering of {245, 246}. This suggests that the model's performance depends on the clarity and specificity of the relationships between mentions.

Threshold from 0.75

Table 20. Examples of Correctly vs. Incorrectly Identified Clusters

Correct Clusters	Incorrect Clusters	Golden Cluster
	{261, 262, 263, 264}	{268, 261}, {262}, {263}, {264}
	{308, 309, 310}	{308, 309}, {310}

The test set results show improved precision at higher thresholds, but over-segmentation remains a significant issue. The model's recall and precision values converge at these thresholds, indicating a more conservative clustering approach. However, it results in missed opportunities for grouping mentions that should be clustered together. For instance, clusters like {261, 262, 263, 264} (similar to the {106, 107, 108, 109} cluster in the validation set) illustrate the model's difficulty in distinguishing finer nuances within policy topics, even though the overall clustering structure shows some improvement. The F1 score at this threshold (0.489) suggests that, while the clusters formed are structurally closer to the golden set, the model still struggles to balance overgeneralization and over-segmentation.

Table 21. Qualitative Analysis of Test Set Results

Row ID	Actual Reference	Qualitative Analysis
{261, 262, 263, 264} /predicted/ vs. {268, 261}, {262}, {263}, {264} /golden/	"Fuel prices were lowered in early November as part of government measures to reduce the rate of inflation. Automotive gasoline prices were reduced by 5.3 per cent for ordinary petro." {261} "After a sharp rise in the consumer price index of. 1.1 per cent in September, which brought the rise in prices from a year earlier up to 9.5 per cent, the government lowered the price of petrol and Introduced special import programmes to lower food prices." {261} "Fuel prices were lowered in early November as part of government measures to reduce the rate of inflation. Automotive gas oil fell by 3.3 per cent to Pta58 per litre." {262} "Fuel prices were lowered in early November as part of government measures to reduce the rate of inflation. The price of industrial fuel was reduced by 15 per cent." {263} "Fuel prices were lowered in early November as part of government measures to reduce the rate of inflation. the price of fuel oil used in thermal stations by 15.3 per cent." {264}	The project's original annotators consider 261, 262, 263, and 264 as separate reforms. We identified 268 as a mention of 261 during the manual annotating process. Despite the shared theme of fuel price reduction in response to inflation, the original annotation treated {261, 262, 263, 264} as separate reforms, reflecting the specific policy measures for different fuel types. The model incorrectly clustered them, overgeneralizing them due to the overarching theme of fuel price adjustment. This example highlights the model's challenge in differentiating between specific mentions within a broader policy context.
{308, 309, 310} /predicted/ vs. {308, 309}, {310} /golden/	"The budget did, however, tinker with legislation covering the taxation of savings. Personal savings and investment changes in the taxation of life assurance ". {308} "The budget did, however, tinker with legislation covering the taxation of savings. Personal savings and investment changes in the taxation of pensions." {309} "The budget did, however, tinker with legislation covering the taxation of savings. Personal savings and investment only minor changes to capital gains tax." {310}	308 and 310 were originally coded as standalone reforms, and 309 was identified as an additional mention of 208 reform. The model incorrectly merged {308, 309} while isolating {310}. Although the references focus on taxation, the original annotation differentiated these mentions due to their specific context (life assurance vs. pensions vs. capital gains tax). The model's failure to capture these distinctions underscores its difficulty in recognizing finer nuances in the policy text.

The analysis shows that while the model can sometimes group mentions based on overarching themes, it needs to work on maintaining precision when distinguishing between closely related yet distinct policy reforms. This limitation indicates an inherent challenge in using a pre-trained sentence encoder without additional training for domain-specific nuances.

Moreover, over-clustering remains problematic despite improved precision at this threshold, as seen in the above examples. The model does not effectively capture subtle policy differences, such as vehicle tax variations or fuel price adjustments. Instead, it tends to merge these nuanced mentions into larger clusters, missing the specific distinctions identified in the golden clusters.

The transition from overgeneralization at lower thresholds to over-segmentation at higher thresholds underscores the complexity of fine-tuning clustering models for policy reform analysis. Middle-range thresholds, like 0.75, offer a better balance between these extremes but still require further refinement to capture the subtleties within the dataset. Implementing a tailored training phase, rather than solely relying on pre-trained models, could enhance clustering accuracy and lead to more precise grouping without excessive fragmentation.

In conclusion, the discrepancies observed in the test set, including the splitting of clusters that should be combined and the merging of unrelated mentions, reveal a systematic issue in how the model handles specific types of relationships. This highlights the need for more advanced approaches to accurately capture the complexity of policy reform discussions in the text.

6.3 Summary of the Results

Table 22. Summary of the Results at the Best Threshold

	Validation Set	**Test Set**
LEA Recall	0.3879	0.5507
LEA Precision	0.4023	0.5507
F1	0.3950	0.5507

The results of the analysis of the baseline model show both strengths and limitations in its ability to disambiguate policy reform mentions. The F1 score of 0.5507 on the test set, though an improvement from the validation set (F1 score of 0.3950), requires careful interpretation. This improvement primarily reflects the model's ability to identify singleton clusters rather than effectively capturing the complexity of clusters with multiple mentions. Thus, while the model can handle unseen data somewhat, it still exhibits significant gaps in precision and recall, particularly for more complex clustering tasks.

Table 23. Changes in Jaccard Similarity and Number of Matches

Threshold	Matches[7]	Jaccard Similarity	Predicted Clusters[8]
-1.0	0	0.0435	1
0.25	0	0.0435	1
0.3334	0	0.2719	2
0.4167	8	0.6410	21
0.5	26	0.6748	56
0.5834	46	0.7007	92
0.6667	55	0.6949	114
0.75	65	0.7087	125
0.8334	66	0.7093	127
0.9167	76	0.7346	135
1.0	81	0.7464	138

The evaluation included the Jaccard similarity score and the LEA metric to provide a comprehensive view of the model's performance. While the LEA metric offers a detailed assessment of precision and recall by focusing on coreference links, the Jaccard index measures the overlap between predicted and golden clusters. Together, these metrics provide a nuanced evaluation, highlighting different aspects of the model's effectiveness.

As shown in *Table 22*, the Jaccard similarity score of 0.7346 at the optimal threshold suggests that the model captures some thematic overlap. Nevertheless, only 76 out of 138 predicted clusters match the golden clusters. The Jaccard similarity and LEA metrics indicate the model's limitations in achieving precise clustering, confirming the model's oscillation between overgeneralization and over-segmentation. At lower thresholds (e.g., -1.0 to 0.25), the model tends to form a single large cluster, while at higher thresholds (e.g., 1.0), it predominantly outputs singleton clusters. Despite the increasing Jaccard index at higher thresholds, many of these matches correspond to singletons rather than accurately capturing multi-element clusters. This tendency to isolate mentions rather than group-related ones reflects the model's struggle to identify the nuances of more complex cluster structures. Interestingly, the Jaccard index at the 0.5834 threshold is not much lower than the best threshold despite having 92 predicted clusters—closer to the number of golden clusters—compared to 135 at the best threshold. This finding suggests that a higher number of predicted clusters does not necessarily result in better thematic alignment.

[7] Matches between the ground truth clusters and the predicted clusters at the threshold.

[8] Number of predicted clusters at the threshold against the number of golden clusters (103).

The disambiguation model shows a noticeable improvement compared to the baseline model using only singleton clusters, which resulted in an F1 score of approximately 0.191. The F1 score of 0.5507 on the test set indicates that our approach surpasses the singleton baseline, demonstrating some capability in forming clusters beyond single mentions. However, this improvement largely stems from the model's accuracy in identifying singletons rather than effectively resolving more complex, multi-mention clusters.

While comparing our model with the baseline, it is important to acknowledge the compositional differences between the datasets used. The baseline model treats all mentions as singletons and sets a conservative reference point, demonstrated by its low precision but perfect recall. By design, the baseline's focus on singletons simplifies the task, whereas our annotated dataset encompasses both singletons and multi-mention clusters, adding complexity. Our model's performance, though not perfect, represents an attempt to bridge this complexity by moving beyond the baseline's simplistic clustering approach.

The results suggest that while the baseline model is a useful starting point for automated policy reform disambiguation, its limitations underscore the need for refinement. Specifically, the model's reliance on pre-trained sentence embeddings without domain-specific training hindered its capture of nuanced relationships between policy reform mentions. Incorporating task-specific data and training could enable the model to recognize better and categorize complex policy reforms, moving beyond the baseline's simplistic clustering. Additionally, the model was evaluated solely on English-language reports, restricting its generalization ability across multilingual policy environments. Future research should explore cross-lingual and domain adaptation techniques to enhance the model's ability to handle diverse textual nuances in policy reform discussions.

Despite its limitations, the study shows potential for automated annotation in policy analysis. However, the model's shortcomings suggest that integrating more sophisticated automated techniques and human oversight is necessary. Semi-supervised learning models may help reduce the need for extensive manual annotation while accelerating data collection and analysis. Refining clustering strategies is crucial to avoid over-segmentation and overgeneralization, particularly for multi-mention entities, to ensure more meaningful and accurate groupings.

7. Conclusion

The primary goal of this thesis was to establish a baseline model for disambiguating policy reforms within text corpora, focusing on identifying and categorizing multiple mentions of policy reforms. A significant contribution of this work was the augmentation of the original dataset. A more comprehensive dataset was created by manually reviewing and annotating additional mentions, serving as a crucial benchmark for future policy reform disambiguation studies. The study aimed to answer the following research questions:

1) ***Can NLP techniques efficiently extract and disambiguate multiple mentions of policy reforms from the available text corpora of the Economist Intelligence Unit's Country Reports?***
2) ***How effectively can a pre-trained sentence embedding model, combined with a simple clustering technique, identify and categorize multiple mentions of policy reforms?***
3) ***Can a simple unsupervised disambiguation system outperform an all-singletons baseline?***

To address these questions, a baseline disambiguation system was developed using NLP techniques through a pre-trained sentence embedding model. While the model demonstrated some effectiveness in policy reform disambiguation, its success was largely confined to identifying singleton clusters, struggling to group related mentions into the same reform, especially at lower thresholds. This limitation in handling multi-mention clusters highlights the model's limitations and the need for more sophisticated approaches to effectively capture the nuances of policy reform texts.

The model's performance was evaluated using the LEA scorer and the Jaccard similarity index. Although the F1 score improved from 0.395 on the validation set to 0.5507 on the test set, this improvement was primarily due to the model's ability to identify singletons rather than effectively clustering multiple mentions of the same reform. The Jaccard similarity score, which reached 0.7346 at the optimal threshold, indicated some thematic overlap between predicted and golden clusters. However, this overlap resulted from accurately identifying singletons rather than multi-element clusters.

Threshold tuning revealed the model's tendency to over-cluster at lower thresholds (e.g., -1.0 to 0.25), grouping unrelated mentions and resulting in poor precision. Conversely, at higher thresholds (e.g., 0.75 and above), the model exhibited over-segmentation, correctly identifying singletons at the expense of forming cohesive clusters of related mentions. Although the Jaccard similarity score improved at the optimal threshold of 0.9167, the exact matches were predominantly singletons.

The disambiguation system demonstrated improved performance compared to the all-singletons baseline, which yielded a precision of 0.106 and an F1 score of 0.191. However, this improvement must be interpreted cautiously, as it primarily stemmed from the model's ability to identify singletons rather than resolve more complex clustering tasks correctly. The baseline model's simplistic approach emphasized the inherent difficulty in disambiguating policy reform mentions, offering a benchmark for assessing more advanced techniques.

The absence of a tailored training phase further limited the model's ability to capture domain-specific nuances—the reliance on pre-trained sentence embeddings without fine-tuning led to difficulties in accurately clustering multi-mention reforms. Future work should incorporate domain-specific training to enhance the model's capacity for effective disambiguation. Additionally, the model's over-segmentation at higher thresholds suggests the need for refined clustering strategies that balance overgeneralization and over-segmentation.

The model was only evaluated on English-language reports, restricting its broader applicability. Future research should expand its capabilities to encompass multiple languages and diverse policy environments. The model's difficulty differentiating between national reports underscores the importance of integrating country-specific information. The results indicate that the model struggled to distinguish between reports from different countries without explicit contextual cues. Addressing these challenges will enhance the model's accuracy and generalizability in a broader, multilingual policy analysis context.

References

Ahmed, N. (2022). Natural Language Processing Techniques for Political Opinion and Sentiment (Doctoral dissertation, Harvard University).

Ahlquist, J. S. & Breunig, C. (2012). Model-based clustering and typologies in the social sciences. Political Analysis, 20(1), 92–112. https://doi.org/10.1093/pan/mpr039

Al-Saadawi, H. F. T., & Das, R. (2024). TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network. Applied Sciences, 14(6), 2252.

Angelova, M., Bäck, H., Müller, W. C., & Strobl, D. (2018). Veto player theory and reform making in Western Europe. European Journal of Political Research, 57(2), 282-307.

Bäck, H., Müller, W. C., Angelova, M., & Strobl, D. (2022). Ministerial autonomy, parliamentary scrutiny and government reform output in parliamentary democracies. Comparative political studies, 55(2), 254-286.

Basile, V., Cabitza, F., Campagner, A., & Fell, M. (2021). Toward a perspectivist turn in ground truthing for predictive computing. arXiv preprint arXiv:2109.04270.

Bein, A. S., & Williams, A. (2023). Development of Deep Learning Algorithms for Improved Facial Recognition in Security Applications. IAIC Transactions on Sustainable Digital Innovation (ITSDI), 5(1), 19-23.

Bengtson, E., & Roth, D. (2008). Understanding the value of features for coreference resolution. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 294-303).

Bengio, Yoshua (2016). Springtime for AI: The Rise of Deep Learning | Scientific American. https://www.scientificamerican.com/article/springtime-for-ai-the-rise-of-deep-learning/

Benoît, C. (2019). The new political economy of regulation. French Politics, 17(4), 482-499.

Bergman, M. E., Angelova, M., Bäck, H., & Müller, W. C. (2023). Coalition agreements and governments' policy-making productivity. West European Politics.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Briggs, J. (2021). Masked-language modeling with bert Masked-Language Modeling With BERT | by James Briggs | Towards Data Science

Cai, S., & Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 748-752).

Cai, J., & Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In Proceedings of the SIGDIAL 2010 Conference (pp. 28-36).

Carpintero-Rentería, M., Santos-Martín, D., Chinchilla, M., & Rebollal, D. (2019). Microgrid Infrastructure Compendium Analysis with a Model Creation Tool and Guideline Based on Machine Learning Techniques. Energies, 12(23), 4509.

Castelli, M., Vanneschi, L., & Largo, Á. R. (2018). Supervised learning: classification. por Ranganathan, S., M. Grisbskov, K. Nakai y C. Schönbach, 1, 342-349.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

Chorowski, J., & Jaitly, N. (2016). Towards better decoding and language model integration in sequence to sequence models. arXiv preprint arXiv:1612.02695.

Clark, K., & Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1405-1415).

Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:1606.01323.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12, 2493-2537.

Culotta, A., Wick, M., & McCallum, A. (2007). First-order probabilistic models for coreference resolution. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (pp. 81-88).

Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering, 5(2), 1301-1309.

Das, D., & Smith, N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 1435-1444).

Daumé, H. (2017). A course in machine learning (pp. 149-155). Hal Daumé III.

Denis, P., & Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (pp. 236-243).

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

Diekmann, A. (2007). Empirische Sozialforschung (11th ed.). Rowohlt.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In Lrec (Vol. 2, No. 1, pp. 837-840).

Duong, S. (2023). Machine Learning: An Introduction to Propagation and Gradient Descent in Feed Forward Neural Networks https://www.linkedin.com/pulse/machine-learning-introduction-propagation-gradient-descent-duong/

Engel, U., Quan-Haase, A., Liu, S. X., & Lyberg, L. (2021). Introduction to the handbook of computational social science. In Handbook of Computational Social Science, Volume 2 (pp. 1-13). Routledge.

Ennser-Jedenastik, L., & Meyer, T. M. (2018). The impact of party cues on manual coding of political texts. Political Science Research and Methods, 6(3), 625-633.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, 104743.

Gatt, A. & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61, 65–170. https://doi.org/10.1613/jair.5477

Ghahramani, Z. (2003). Unsupervised learning. In Summer school on machine learning (pp. 72-112). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gigley, H. M. (1993). Projected government needs in human language technology and the role of researchers in meeting them. In Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993.

Gilardi, F., & Wüest, B. (2020). Using text-as-data methods in comparative policy analysis. In Handbook of research methods and applications in comparative policy analysis (pp. 203-217). Edward Elgar Publishing.

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2019). Computational analysis of political texts: bridging research efforts across communities. In Proceedings of the 57th annual meeting of the association for computational linguistics: Tutorial abstracts (pp. 18-23).

Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.

Goodfellow, I. (2016). Deep learning.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.

Gupta, L. (2021). Precision-Recall Tradeoff in Real-World Use Cases. https://medium.com/analytics-vidhya/precision-recall-tradeoff-for-real-world-use-cases-c6de4fabbcd0

Haghighi, A., & Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 385-393).

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication methods and measures, 1(1), 77-89.

Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., ... & Wen, T. H. (2019). A repository of conversational datasets. arXiv preprint arXiv:1904.06472.

Henderson, M., Thomson, B., & Williams, J. D. (2014). The second dialog state tracking challenge. In Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL) (pp. 263-272).

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Hirschman, L., & Chinchor, N. (1997). MUC-7 coreference task definition.

Hochreiter, S. (1997). Long Short-term Memory. Neural Computation MIT-Press.

Hodnett, M., Wiley, J. F., Liu, Y. H., & Maldonado, P. (2019). Deep Learning with R for Beginners: Design neural network models in R 3.5 using TensorFlow, Keras, and MXNet. Packt Publishing Ltd.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluation Hodnett, M., Wiley, J. F., Liu, Y. H., & Maldonado, P. (2019). Deep Learning with R for Beginners: Design neural network models in R 3.5 using TensorFlow, Keras, and MXNet. Packt Publishing Ltd.s. International journal of data mining & knowledge management process, 5(2), 1.

Javaloy, A., & García-Mateos, G. (2020). Preliminary Results on Different Text Processing Tasks Using Encoder-Decoder Networks and the Causal Feature Extractor. Applied Sciences, 10(17), 5772.

Jin, C., & Rinard, M. (2023). Evidence of Meaning in Language Models Trained on Programs. arXiv preprint arXiv:2305.11169.

Jin, Z., & Mihalcea, R. (2022). Natural language processing for policymaking. In Handbook of Computational Social Science for Policy (pp. 141-162). Cham: Springer International Publishing.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Kantor, B., & Globerson, A. (2019). Coreference resolution with entity equalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 673-677).

Kavlakoglu, E. (2020). AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the Difference, 27, 2020.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American political science review, 95(1), 49-69.

Krippendorff, K., & Craggs, R. (2016). The reliability of multi-valued coding of data. Communication Methods and Measures, 10(4), 181-198.

Lai, T. M., Bui, T., & Kim, D. S. (2022, May). End-to-end neural coreference resolution revisited: A simple yet effective baseline. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8147-8151). IEEE.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), 1-30.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI open, 3, 111-132.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions.

Luo, X. (2005). On coreference resolution performance metrics. In Proceedings of human language technology conference and conference on empirical methods in natural language processing (pp. 25-32).

Luz Tortorella, G., Cauchick-Miguel, P. A., Li, W., Staines, J., & McFarlane, D. (2022). What does operational excellence mean in the Fourth Industrial Revolution era?. International Journal of Production Research, 60(9), 2901-2917.

Manning, C. D. (2008). Introduction to information retrieval.

Martinez, N. A. (2023). Understanding Vector Embeddings in NLP: An Introduction with the ALL-MINILM-L6-V2 Model. https://www.linkedin.com/pulse/understanding-vector-embeddings-nlp-introduction-model-martinez/

Mayer, K., & Pfeffer, J. (2019). Lazer et al.(2009): Computational Social Science. Schlüsselwerke der Netzwerkforschung, 335-337.

Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. Political Analysis, 20(1), 78-91.

Mitchell, T. M. (1997). Does machine learning really work?. AI magazine, 18(3), 11-11.

Moosavi, N. S., & Strube, M. (2016). Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 632-642).

Mungalpara, J. (2023). https://jaimin-ml2001.medium.com/evaluation-methods-in-natural-language-processing-nlp-part-1-ffd39c90c04f

Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4(51-62), 56.

Ngoko, I., Mukherjee, A., & Kabaso, B. (2018). Abstractive Text Summarization Using Recurrent Neural Networks: Systematic Literature Review. International Conference on Intellectual Capital and Knowledge Management and Organisational Learning, (), 435-439, XIII.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. Journal of computational social science, 6(1), 289-313.

Orellana, S., & Bisgin, H. (2023). Using natural language processing to analyze political party manifestos from New Zealand. Information, 14(3), 152.

Passonneau, R. (2004). Computing reliability for coreference annotation.

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. Communication Methods and Measures, 13(4), 287-304.

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2024). From manual to machine: assessing the efficacy of large language models in content analysis. Communication Research Reports, 1-10.

Poesio, M., & Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky (pp. 76-83).

Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J. (2004). Learning to resolve bridging references. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (pp. 143-150).

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference (pp. 192-199).

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., & Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2014, p. 30). NIH Public Access.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Joint conference on EMNLP and CoNLL-shared task (pp. 1-40).

Prakash, P. (2023). Understanding Baseline Models in Machine Learning Importance, Strategies, and Application to Imbalanced Classes. https://medium.com/@preethi_prakash/understanding-baseline-models-in-machine-learning-3ed94f03d645

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word sense disambiguation: a uinified evaluation framework and empirical comparison. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (Vol. 1, pp. 99-110).

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., & Both, A. (2014). N^3-A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In LREC (pp. 3529-3533).

Rush, A. (2015). A neural attention model for abstractive sentence summarization. arXiv Preprint, CoRR, abs/1509.00685.

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., ... & Riedel, S. (2018). Interpretation of natural language rules in conversational machine reading. arXiv preprint arXiv:1809.01494.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2019). Masked language model scoring. arXiv preprint arXiv:1910.14659.

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In 2018 Second international conference on intelligent computing and control systems (ICICCS) (pp. 945-949). IEEE.

Sarawagi, S. (2008). Information extraction. Foundations and Trends® in Databases, 1(3), 261-377.

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 338-343). IEEE.

Sennikova, T. (2020). How to Build a Baseline Model. A pragmatic approach for building a baseline model to understand your data. https://towardsdatascience.com/how-to-build-a-baseline-model

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. The ANNALS of the American Academy of Political and Social Science, 659(1), 6-13.

Shi, W., & Demberg, V. (2019). Next sentence prediction helps implicit discourse relation classification within and across domains. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 5790-5796).

Smith, N. A. (2019). Contextual word representations: A contextual introduction. arXiv preprint arXiv:1902.06006.

Smys, S., & Raj, J. S. (2021). Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. Journal of trends in Computer Science and Smart technology (TCSST), 3(01), 24-39.

Stolfo, A., Tanner, C., Gupta, V., & Sachan, M. (2022). A Simple Unsupervised Approach for Coreference Resolution using Rule-based Weak Supervision. In Proceedings of the 11th Joint Conference on Lexical and Computational Semantics (pp. 79–88).

Stoyanov, V., Gilbert, N., Cardie, C., & Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 656-664).

Strobl, D., Bäck, H., Müller, W. C., & Angelova, M. (2021). Electoral cycles in government policy-making: Strategic timing of austerity reform measures in Western Europe. British Journal of Political Science, 51(1), 331-352.

Trask, A., Michalak, P., & Liu, J. (2015). sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. arXiv preprint arXiv:1511.06388.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vickers, P., Barrault, L., Monti, E., & Aletras, N. (2024). We Need to Talk About Classification Evaluation Metrics in NLP. arXiv preprint arXiv:2401.03831.

Wankmüller, S. (2022). Introduction to neural transfer learning with transformers for social science text analysis. Sociological Methods & Research, 00491241221134527.

Wiebe, J., Bruce, R., & O'Hara, T. P. (1999, June). Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics (pp. 246-253).

Wibawa, A. P., Fadhilla, A. F., Paramarta, A. K. A. I., Triono, A. P. P., Setyaputri, F. U., Akbari, A. K. G., & Utama, A. B. P. (2024). Bidirectional Long Short-Term Memory (Bi-LSTM) Hourly Energy Forecasting. In E3S Web of Conferences (Vol. 501, p. 01023). EDP Sciences.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. Annual Review of Political Science, 20(1), 529-544.

Wiseman, S., Rush, A. M., & Shieber, S. M. (2016). Learning global features for coreference resolution. arXiv preprint arXiv:1604.03035.

Wu, W., Wang, F., Yuan, A., Wu, F., & Li, J. (2020). CorefQA: Coreference resolution as query-based span prediction. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 6953-6963).

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2, 165–193.

Yang, Y., Duan, H., Abbasi, A., Lalor, J. P., & Tam, K. Y. (2023). Bias a-head? analyzing bias in transformer-based language model attention heads. arXiv preprint arXiv:2311.10395.

Ying, X. (2019). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.