

Detecting Topic Merges and Splits in Dynamic Discourse: A Graph-Based Framework for Political Text Corpora

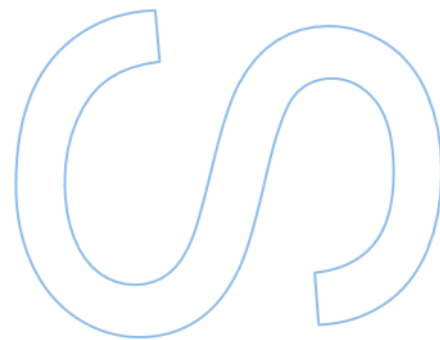
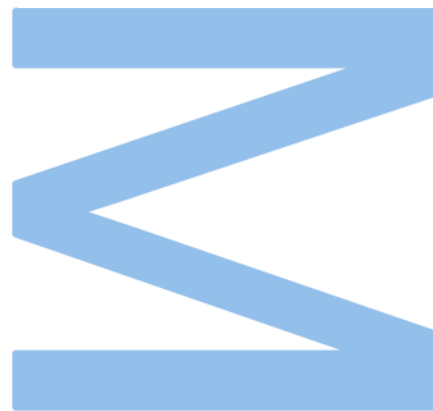
Cláudia Oliveira

Master in Data Science

Department of Computer Science

Faculty of Sciences of the University of Porto

2026



Detecting Topic Merges and Splits in Dynamic Discourse: A Graph-Based Framework for Political Text Corpora

Cláudia Oliveira

Dissertation carried out within the scope of the Master in
Data Science
Department of Computer Science
2026

Supervisor

Prof. Dr. Álvaro Figueira, Assistant Professor
Faculty of Sciences of the University of Porto

“ You can’t put a limit on anything. The more you dream, the farther you get. ”

Michael Phelps

Dedicated to my parents and my boyfriend

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Álvaro Figueira, for his guidance, invaluable advice, and continuous support throughout this research.

To my family, thank you for always believing in me, for giving me every opportunity to succeed, and for constantly encouraging me to become a better version of myself.

To my boyfriend, thank you for your endless patience, for believing in me even when I struggled to believe in myself, and for standing by my side every step of the way. I cannot begin to imagine what this journey would have been like without my safe place and my greatest supporter especially during the moments when everything seemed to be falling apart.

To my Plaza, thank you for teaching me the value of consistency. To my Mariana, thank you for showing me what persistence truly means. And to my Marta, thank you for sharing every emotion, every challenge, and every moment of uncertainty that this journey brought us.

Finally, I would like to thank everyone who, in one way or another, contributed to the successful completion of this dissertation.

Resumo

Os corpora textuais longitudinais evoluem continuamente à medida que o discurso público se transforma ao longo do tempo. Para além das mudanças graduais de vocabulário, os tópicos podem sofrer transformações estruturais discretas, incluindo o surgimento de novos temas, o desaparecimento de temas existentes, ou a divisão e fusão de discussões. Enquanto os modelos dinâmicos de tópicos assumem tipicamente uma evolução suave dos tópicos, as abordagens baseadas em janelas temporais (snapshots) identificam os tópicos de forma independente em cada janela temporal, deixando em aberto o problema de rastrear a sua evolução.

Esta tese apresenta o BERTilda (BERT-based Temporal Identification, Lifecycle Detection and Analysis), uma framework baseado em grafos para a deteção e caracterização da evolução de tópicos em corpora textuais temporais. A abordagem proposta descobre tópicos de forma independente em cada janela temporal através de um modelo de tópicos baseado em embeddings e reconstrói a sua evolução alinhando tópicos entre janelas consecutivas. O alinhamento de tópicos combina similaridade semântica num espaço de embeddings partilhado com um novo mecanismo bidirecional de cobertura de documentos, que capta tanto o fluxo de saída como o de entrada de documentos entre tópicos. Com base nestes sinais complementares, critérios transparentes baseados em regras identificam eventos do ciclo de vida dos tópicos, incluindo continuações, divisões, fusões, desaparecimentos e transições indefinidas, representados sob a forma de um grafo temporal de tópicos.

A framework é avaliada em múltiplos corpora de texto político, incluindo tweets do Congresso dos Estados Unidos, discursos do Debate Geral das Nações Unidas e discursos sobre o Estado da União. A avaliação experimental combina métricas de qualidade de tópicos, diagnósticos de estabilidade temporal, estudos de caso qualitativos, validação humana de eventos de ciclo de vida detetados e análises de robustez sob perturbações de limiares (thresholds). Os resultados mostram que o BERTilda produz representações de tópicos coerentes e estáveis, identificando simultaneamente transições estruturais significativas. Num subconjunto de referência anotado por três avaliadores independentes, a framework proposta atinge taxas de concordância maioritária até 87%, superando abordagens baseadas apenas em similaridade e baselines unidirecionais, particularmente na deteção de desaparecimentos de tópicos. Estes resultados demonstram que o BERTilda constitui uma framework interpretável e eficaz para a análise da evolução de tópicos em

corpora textuais longitudinais.

Palavras-chave: evolução de tópicos, grafos temporais, deteção de divisão e fusão, modelação de tópicos

Abstract

Longitudinal text corpora evolve continuously as public discourse changes over time. Beyond gradual shifts in vocabulary, topics may undergo discrete structural transformations, including the emergence of new themes, the disappearance of existing ones, or the splitting and merging of discussions. While dynamic topic models typically assume smooth topic evolution, snapshot-based approaches identify topics independently in each temporal window but leave the problem of tracking topic evolution largely unresolved.

This thesis presents BERTilda (BERT-based Temporal Identification, Lifecycle Detection and Analysis), a graph-based framework for detecting and characterizing topic evolution in temporal text corpora. The proposed approach discovers topics independently within each time window using an embedding-based topic model and reconstructs their evolution by aligning topics across consecutive windows. Topic alignment combines semantic similarity in a shared embedding space with a novel bidirectional document coverage mechanism that captures both document outflow and inflow between topics. Based on these complementary signals, transparent rule-based criteria identify topic lifecycle events, including continuations, splits, merges, disappearances, and unclear transitions, which are represented as a temporal topic graph.

The framework is evaluated on multiple political text corpora, including U.S. congressional tweets, United Nations General Debate speeches, and State of the Union addresses. The experimental evaluation combines topic quality metrics, temporal stability diagnostics, qualitative case studies, human validation of detected lifecycle events, and robustness analyses under threshold perturbations. Results show that BERTilda produces coherent and stable topic representations while identifying meaningful structural topic transitions. On a gold-standard subset annotated by three independent annotators, the proposed framework achieves majority agreement rates of up to 87% and outperforms similarity-only and unidirectional baselines, particularly in detecting topic disappearances. These findings demonstrate that BERTilda provides an interpretable and effective framework for analyzing topic evolution in longitudinal text corpora.

Keywords: topic evolution, temporal graphs, split and merge detection, explainable topic modeling

Table of Contents

List of Figures.....	x
List of Abbreviations.....	xi
1. Introduction.....	1
1.1. Motivation.....	1
1.2. Research Questions	5
1.3. Thesis Structure.....	6
2. State of the Art	7
2.1. Embedding-Based Topic Models.....	9
2.1.1. Top2Vec.....	9
2.1.2. BERTopic	10
2.2. Topic Modeling for Short and Noisy Texts	11
2.2.1. Aggregation-Based Approaches.....	11
2.2.2. Alternative Approaches.....	11
2.3. Dynamic Topic Models	12
2.4. Tracking Structural Change in Clusters and Topics.....	13
2.4.1. Community Detection in Static Graphs.....	13
2.4.2. Frameworks for Community Evolution.....	14
2.4.3. Cluster Evolution and Topic Tracking.....	14
2.4.3.1. Evolutionary Clustering Objectives	14
2.4.3.2. Snapshot Matching for Split and Merge Detection	14
2.5. Evaluation of Topic Models	15
2.5.1. Automatic Coherence Metrics.....	16
2.5.2. Topic Diversity.....	16
2.5.3. Temporal Stability Metrics.....	17
2.6. Political Discourse Analysis	17
2.6.1. Agenda-Setting and Framing.....	18
2.6.2. Political Polarization and Ideological Analysis.....	18
2.6.3. Computational Analysis of Congressional Communication.....	19
3. Problem Formulation	20
3.1. Temporal Text Corpora.....	20
3.2. Temporal Segmentation	21
3.3. Snapshot Topics	23
3.4. Inter-window Topic Alignment	24
3.5. Topic Lifecycle Events	25

3.6. Problem Statement	26
3.7. Challenges and Ambiguity	27
4. Methodology	28
4.1. Overview of the Proposed Framework	28
4.2. Topic Discovery per Temporal Window	29
4.3. Topic Representation and Similarity	30
4.4. Bidirectional Coverage Estimation	31
4.4.1. Forward Attribution (Outflow)	31
4.4.2. Backward Attribution (Inflow)	32
4.5. Event Labeling Rules	32
4.5.1. Continuation	33
4.5.2. Split	33
4.5.3. Disappearance	34
4.5.4. Merge	34
4.5.5. Unclear Transitions	34
4.6. Temporal Topic Graph Construction	34
4.7. BERTilda Algorithm	35
4.8. Implementation Considerations	36
5. Experimental Design	37
5.1. Research Questions	37
5.2. Datasets	38
5.2.1. Congressional Tweets	38
5.2.2. UN General Debates	39
5.2.3. State of the Union Addresses	39
5.3. Data Preprocessing	39
5.3.1. State of the Union Addresses	39
5.3.2. UN General Debates	40
5.3.3. Congressional Tweets	40
5.3.4. Overall Preprocessing Design	42
5.4. Experimental Configuration	42
5.4.1. Temporal Windowing	42
5.4.2. Threshold Parameters	43
5.5. Baselines and Comparison Methods	43
5.5.1. Topic Discovery Baselines	43
5.5.2. Temporal Modeling Baselines	44
5.5.3. Event Detection Baselines	44
5.6. Gold Set and Human Annotation Protocol	45

5.7. Evaluation Metrics	45
5.7.1. Topic Quality Metrics	45
5.7.2. Temporal Stability Metrics.....	46
5.7.3. Event Validation Metrics	47
5.8. Case Studies.....	48
6. Results.....	49
6.1. Static Topic Quality and Fragmentation (Congress)	49
6.2. Temporal Stability Diagnostics (Historical Corpora).....	50
6.2.1. UN General Debates (1970–2015)	50
6.2.2. State of the Union (1790–2018).....	52
6.3. Qualitative Case Studies (Congress)	54
6.3.1. Dashboard.....	54
6.3.1.1. Network Graph	54
6.3.1.2. Topics’ Information Table	55
6.3.1.3. Documents Over Time	56
6.3.2. Holocaust/Israel–Antisemitism.....	56
6.3.3. Town Halls and Constituent Services Assistance.	57
6.3.4. Disaster Management and Agricultural Impact.	58
6.3.5. Shooting Event to Political Polarization.	59
7. Human Validation	61
7.1. Evaluation Setup.....	61
7.1.1. Annotation Protocol.....	61
7.1.2. Annotated Evaluation Sample	61
7.2. Event Detection Baselines.....	62
7.2.1. Baseline Definitions	62
7.2.2. Protocol and Comparability Safeguards	62
7.3. Validation Results	63
7.3.1. Distribution of Predicted Events.....	63
7.3.2. Human Validation Rates	64
8. Robustness to threshold perturbations	66
9. Discussion	70
10. Conclusions	73
10.1. Summary of the Work	73
10.2. Research Questions	73
10.3. Limitations and Future Work.....	74
Bibliography	75

List of Figures

2.1. Evolution of topic modeling.	8
2.2. Topic Modeling Pipelines.	9
2.3. BERTopic Pipeline.	10
3.1. Documents distributed over time.	21
3.2. Temporal Window Segmentation.	22
3.3. Example of snapshot Topics per Window W_t . Each circle represents a topic.	24
3.4. Topic Alignment over time.	25
3.5. Illustration of the five topic lifecycle events.	26
4.1. BERTilda's Pipeline	29
4.2. Example of a temporal topic graph illustrating topic evolution across consecutive time windows.	35
5.1. Overview of the preprocessing workflow used to prepare textual tweet data for topic modeling.	41
6.1. Node color legend.	55
6.2. Example of node hover information.	55
6.3. Example of edge hover information.	55
6.4. Example of topic information table: window 1, topic 12. Because the topic did not persist beyond its initial window, no predecessor or successor topics are shown.	56
6.5. Documents over time: Example of topic 18 from window 63	56
6.6. Example temporal evolution of a topic (Holocaust/Israel–Antisemitism), illustrating split, continuations, and disappearance across consecutive windows.	57
6.7. Top words evolution of a topic (Holocaust/Israel–Antisemitism).	57
6.8. Example temporal evolution of a topic (Town Halls and Constituent Services Assistance), illustrating merge, split, continuations, and disappearance across consecutive windows.	58
6.9. Part of top words evolution of a topic (Town Halls and Constituent Services Assistance).	58
6.10 Example temporal evolution of a topic (Disaster Management and Agricultural Impact), illustrating merge and disappearance across consecutive windows.	59
6.11 Top words evolution of a topic (Disaster Management and Agricultural Impact).	59
6.12 Example temporal evolution of a topic (Shooting Event to Political Polarization), illustrating split, continuations, and disappearance across consecutive windows.	60
6.13 Top words evolution of a topic (Shooting Event to Political Polarization).	60
8.1. Local threshold-sensitivity analyses on the congressional gold set.	69
1. Word Cloud of the Top 50 words ranked by their TF-IDF scores.	85
2. Dashboard Layout.	85

List of Abbreviations

- BERT** Bidirectional Encoder Representations from Transformers. 4, 9
- BERTilda** BERT-based Temporal Identification, Lifecycle Detection and Analysis. 4
- BTM** Biterm Topic Model. 12
- c-TF-IDF** Class-based Term Frequency-Inverse Document Frequency. 30
- CTM** Correlated Topic Model. 8
- CV** Coherence Value. 45
- DETM** Dynamic Embedded Topic Model. 44
- DTM** Dynamic Topic Model. 19, 44
- ETM** Embedded Topic Model. 11
- FP** False Positive. 47
- HDBSCAN** Hierarchical Density-Based Spatial Clustering of Applications with Noise. 30
- HDP** Hierarchical Dirichlet Process. 8
- HTML** HyperText Markup Language. 41
- IRBO** Inverted Rank-Biased Overlap. 16
- LDA** Latent Dirichlet Allocation. 44
- LSA** Latent Semantic Analysis. 7
- NMF** Non-negative Matrix Factorization. 44
- NPMI** Normalized Pointwise Mutual Information. 45
- pLSA** Probabilistic Latent Semantic Analysis. 7
- RQ** Research Question. 5

SAGE Sparse Additive Generative Model.	12
SBERT Sentence-BERT.	9
SD Standard Deviation.	46
SVD Singular Value Decomposition.	7
TF-IDF Term Frequency-Inverse Document Frequency.	10
TOT Topics Over Time.	12, 19
TP True Positive.	47
TQ Topic Quality.	46
U.S. United States.	18
UMAP Uniform Manifold Approximation and Projection.	9
UN United Nations.	38
URL Uniform Resource Locator.	41

1. Introduction

In March 2020, when the first large-scale lockdowns were announced, public discourse around COVID-19 was, above all, a health discourse. Conversations revolved around infection curves, hospital capacity, and protective equipment. Within months, however, that single theme had unfolded into a series of parallel conversations: the economic consequences of restrictions, the tension between public health and individual freedoms, the logistics and global equity of vaccination campaigns, and, later, the political accountability for decisions taken under uncertainty. At times, these topics merged back together; at others, they disappeared from public attention as quickly as they had emerged.

A similar pattern can be observed in the discourse surrounding Brexit. What initially circulated as a relatively contained debate about the United Kingdom’s relationship with the European Union gradually fragmented into distinct narratives concerning sovereignty, immigration, trade, the status of Northern Ireland, and national identity. Some threads developed independently, others merged into broader political coalitions, and others still faded as the negotiation process advanced. Throughout the years that followed the 2016 referendum, the topical structure of the debate was anything but stable.

These two examples illustrate a broader phenomenon. In any sustained public debate, topics rarely remain fixed: they emerge, split, recombine, and dissolve as events unfold and actors reposition themselves. For researchers seeking to analyze such processes empirically, the relevant question is no longer simply *which topics are present in a given collection of texts?* It becomes *how did these topics emerge, fragment, recombine, and disappear over time?* This thesis is concerned with the second question and with the methodological challenges that arise when one attempts to answer it at scale.

1.1. Motivation

The exponential growth of digital text has fundamentally transformed the landscape of social and political research. From legislative speeches and party manifestos to social media posts and news articles, the volume of textual data produced daily far exceeds what any human analyst could process manually. This proliferation has intensified the need for automated methods capable of extracting meaningful structure from large-scale corpora, driving significant advances in natural language processing and computational social science.

Topic modeling has emerged as one of the most widely used approaches for uncovering latent semantic patterns in document collections. By representing a corpus as a mixture of latent topics (each characterized by a distribution over words), these methods enable researchers to identify recurring themes, summarize large archives, and analyze discourse at scale without requiring manual annotation. Since the introduction of Latent Dirichlet Allocation (LDA) [1], topic modeling has been applied across a broad range of domains, including political science, journalism, public health, and the digital humanities, to name a few.

However, the vast majority of this work has focused on static corpora, implicitly assuming that the thematic structure of the data remains stable over time. In practice, this assumption rarely holds. Many real-world text collections are inherently temporal, reflecting ongoing social processes such as political debate, legislative activity, news cycles, or the evolution of public opinion. Treating such collections as static not only discards valuable temporal information but may also produce misleading representations in which topics are defined by the average of their content across time rather than by their structure at any given moment.

In temporally dynamic corpora, the questions of interest change fundamentally. It is no longer sufficient to ask *what* topics exist; one must also ask *how* they emerge, evolve, and disappear. As the opening examples illustrate, a political issue may initially surface as a single, coherent topic and later split into distinct narratives as different groups push their own perspectives, or blend into other discussions as broader coalitions form around shared goals. Debates around climate legislation, for instance, may shift from a scientific to an economic or partisan framing depending on the electoral cycle, and issues that dominate one news cycle may resurface months later in an entirely different rhetorical register. Capturing such structural dynamics is essential for understanding how political discourse is organized and how it responds to external events.

This kind of analysis is particularly valuable in the context of political text. Legislative corpora and political social media offer a unique window into the strategic communication of elected representatives, the formation and dissolution of political coalitions, and the agenda-setting dynamics that shape public debate. At the same time, political texts pose specific methodological challenges, as they are often short and noisy, heavily reliant on implicit context, and organized around rapidly shifting events. These characteristics make the modeling of temporal topic evolution both especially important and especially difficult.

The dominant paradigm for temporal topic modeling relies on probabilistic models that incorporate predefined assumptions about how topics evolve over time. Latent Dirichlet Allocation [1], while powerful for static corpora, provides no mechanism for tracking how topics change over time. Dynamic Topic Models [2] addressed this by allowing topic distributions to evolve smoothly over time, capturing changes in word usage across different periods. Related approaches, such as Topics over Time [3] and continuous-time dynamic topic models [4], further extended this idea by modeling temporal dynamics under different probabilistic assumptions. More recently, neural approaches such as the Dynamic Embedded Topic Model [5] have replaced word distributions with latent embedding representations that evolve through time.

These models provide valuable temporal regularization and have demonstrated strong performance on a variety of corpora. However, they share a structural limitation that is central to the problem addressed in this thesis: by design, they assume that topics evolve *smoothly and continuously*. This assumption limits their ability to detect abrupt structural transformations. When a topic splits into two distinct subtopics, or when two previously separate topics merge into a unified discussion, these models will tend to smooth over the transition rather than flag it as a discrete event. The result is a representation that captures gradual drift but obscures the structural reorganizations that are often most informative from a social science perspective.

An alternative paradigm is to apply snapshot topic modeling (running a topic model independently on each temporal window) and subsequently align the resulting topics across windows. This approach avoids imposing continuity and allows the topic structure to change freely between periods. Embedding-based topic models are particularly suited for this setting, especially for short texts. Methods such as Top2Vec [6], BERTopic [7], and embedding-space topic models [8] leverage pre-trained language representations to produce semantically coherent topics without requiring large document-word co-occurrence statistics. However, snapshot approaches introduce a different problem: because topics are discovered independently in each window, there is no principled mechanism for establishing correspondences between them. A topic in window t and a semantically related topic in window $t + 1$ may or may not represent the same underlying discussion, and distinguishing continuation from coincidental similarity is non-trivial.

As a result, there is currently no unified framework that can both: 1) preserve the flexibility of snapshot topic modeling and, 2) explicitly detect and characterize structural events in

topic evolution.

To address the identified gap, we propose "BERTilda" (BERT-based Temporal Identification, Lifecycle Detection and Analysis), a framework that treats temporal topic modeling as an alignment-and-event-labeling problem on a graph. Rather than imposing temporal continuity through model priors, BERTilda separates the tasks of discovering and tracking topics. Topics are first identified independently (though in the same embedding space) in each time window using a snapshot model, after which their temporal evolution is reconstructed through a principled alignment step.

We instantiate the snapshot component with BERTopic [7], an embedding-based method well-suited to short political texts. Nevertheless, the proposed framework is model-agnostic and potentially domain-general, allowing the integration of any topic modeling method capable of producing document-topic assignments and interpretable topic representations. Topics across adjacent windows are then linked using two complementary signals. The first is semantic similarity between topic representations in a shared embedding space, capturing whether two topics discuss similar content. The second is *bidirectional coverage*, a novel interpretable flow signal that estimates: i) document *outflow* from a source topic, defined as the proportion of its documents that are attributed to a target topic in the next window, and ii) document *inflow* into a target topic, defined as the proportion of its documents that originate from a source topic in the previous window. Together, these two perspectives on document flow provide a richer and more interpretable basis for alignment than semantic similarity alone.

Given these signals, BERTilda applies a set of transparent, threshold-based rules to assign lifecycle labels to each topic link: *continuation* (a topic persists largely unchanged), *split* (a topic fragments into multiple successors), *merge* (multiple topics converge into a single successor), *disappearance* (a topic has no meaningful successor), and *unclear* (the transition does not fit any of the above patterns cleanly). These labeled links are assembled into a temporal topic graph, enabling a structured view of how each topic evolves over time.

This thesis makes three original contributions:

- **Bidirectional coverage as a flow-based alignment signal:** We introduce a novel interpretable mechanism that captures how documents transition between topics across time, complementing traditional embedding-based similarity measures.

- **A graph-based framework for lifecycle detection:** We propose a structured representation of topic evolution as a temporal graph, enabling the identification of continuation, split, merge, and disappearance events without assuming smooth evolution. The labeling rules are transparent and auditable, making the framework suitable for social science applications where interpretability is essential.
- **An evaluation protocol combining quantitative and qualitative analysis:** We design a comprehensive evaluation framework that integrates automatic topic quality metrics with human validation of detected events in political corpora. The protocol includes a gold-standard annotation study involving three independent annotators, providing a reliable basis for assessing the semantic validity of lifecycle labels.

This work focuses on temporal topic modeling in political text corpora. While the evaluation includes human validation, the interpretation of topic evolution remains inherently subjective, particularly in complex political contexts.

1.2. Research Questions

To guide this evaluation, we formulate the following research questions:

- **RQ1:** Does the snapshot topic discovery component yield coherent and diverse topics on short political texts, relative to common baselines? This question assesses whether the foundation of the pipeline (per-window topic modeling) produces topics of sufficient quality to support downstream alignment and event detection.
- **RQ2:** How stable are topic-quality diagnostics over time compared to dynamic topic models with explicit temporal priors? This question examines whether snapshot-based modeling can maintain consistent topic quality across time without relying on temporal smoothing.
- **RQ3:** Are the detected topic evolution events (continuations, splits, merges, and disappearances) semantically valid when evaluated by human annotators? This question directly assesses the interpretability and correctness of the lifecycle labels produced by the framework.
- **RQ4:** How does the proposed approach compare with alternative event detection strategies in terms of detection accuracy? This question benchmarks BERTilda against different pipeline variants to assess the contribution of each component to the method’s overall performance.

- **RQ5:** How sensitive is the event detection process to the thresholds used for topic alignment and event classification? This question investigates the robustness of the framework to parameter choices, which is important for understanding its reliability in practice.
- **RQ6:** Do qualitative case studies illustrate that detected events correspond to meaningful narrative reorganizations? This question complements the quantitative evaluation with in-depth analysis of specific events to assess whether the framework captures substantively interesting dynamics.

1.3. Thesis Structure

The remainder of this thesis is organized as follows. Chapter 2 reviews related work on topic modeling, temporal topic evolution, and methods for detecting structural changes such as splits and merges, as well as existing evaluation strategies. Chapter 3 formally defines the problem of temporal topic evolution and introduces the key concepts and notation used throughout the thesis. Chapter 4 presents BERTilda, the proposed approach for modeling and tracking topic evolution over time. Chapter 5 describes the datasets, baseline methods, and evaluation protocols used to assess performance. Chapter 6 presents and analyzes the experimental results. Chapter 7 reports a human evaluation study comparing BERTilda with its variants, focusing on interpretability and alignment quality. Chapter 8 investigates the robustness of BERTilda under parameter variations within controlled thresholds, comparing its stability against the strongest variant. Finally, Chapter 9 discusses the main findings and limitations of the proposed approach, and Chapter 10 summarizes the key contributions and insights, and outlines possible directions for future research and improvements.

2. State of the Art

Topic modeling is a family of unsupervised machine learning techniques designed to uncover latent thematic structures in large text collections. Rather than relying on predefined labels, topic models automatically identify groups of words that tend to occur together across documents, thereby revealing hidden semantic structures in text corpora. These methods have become fundamental tools in natural language processing and information retrieval, enabling the exploration of large text collections such as news archives, scientific publications, and social media streams.

Early approaches to discovering latent structure in text relied on matrix factorization techniques applied to document–term matrices. One of the earliest and most influential methods is *Latent Semantic Analysis* (LSA) [9], which uses singular value decomposition (SVD) to project documents and words into a lower-dimensional semantic space. In this framework, documents that share similar semantic content are represented by nearby vectors in the latent space. Although LSA demonstrated that linear algebra techniques can capture semantic relationships between words and documents, it lacks a probabilistic interpretation and does not define a generative model for text.

To address these limitations, probabilistic formulations of latent semantic structure were later proposed. One of the earliest models in this direction is *Probabilistic Latent Semantic Analysis* (pLSA) [10]. In pLSA, each document is modeled as a mixture of latent topics, and each topic corresponds to a probability distribution over words. Under this model, the generation of each observed word is explained by first sampling a latent topic from the document’s topic distribution and then selecting a word from the corresponding topic-specific distribution.

While pLSA introduced an important probabilistic interpretation of latent topics, it is prone to overfitting, which motivated the development of *Latent Dirichlet Allocation* (LDA) [1], which has become one of the most widely used topic modeling frameworks.

LDA extends the probabilistic framework of pLSA by introducing Dirichlet priors over the document-topic and topic-word distributions. This Bayesian formulation provides a complete generative model for documents and enables the inference of topic mixtures for previously unseen texts. Since its introduction, LDA has been widely applied across domains including scientific literature analysis, social media monitoring, and digital humanities [11].

Several extensions of LDA have been proposed to address limitations in scalability, flexibility, and model selection. For example, *Hierarchical Dirichlet Processes* (HDP) allow the number of topics to be inferred automatically rather than specified in advance [12]. This nonparametric extension is particularly valuable when the appropriate number of topics is unknown, which is common in practice.

Other work has explored matrix factorization approaches such as *Nonnegative Matrix Factorization* (NMF) [13, 14], which provides an alternative decomposition of document-term matrices while maintaining interpretable nonnegative components. The *Correlated Topic Model* (CTM) [15] replaces the Dirichlet prior with a logistic-normal distribution, enabling the model to capture correlations between topics.

More recently, neural approaches have emerged that integrate deep learning techniques with topic modeling frameworks.

Figure 2.1 illustrates the evolution of topic modeling approaches, moving from classical statistical models toward embedding-based methods capable of capturing richer semantic relationships. This transition reflects the increasing relevance of semantic embeddings in topic discovery tasks. Accordingly, the proposed system is designed to support different topic modeling techniques, provided that they operate over embedding-based document representations.

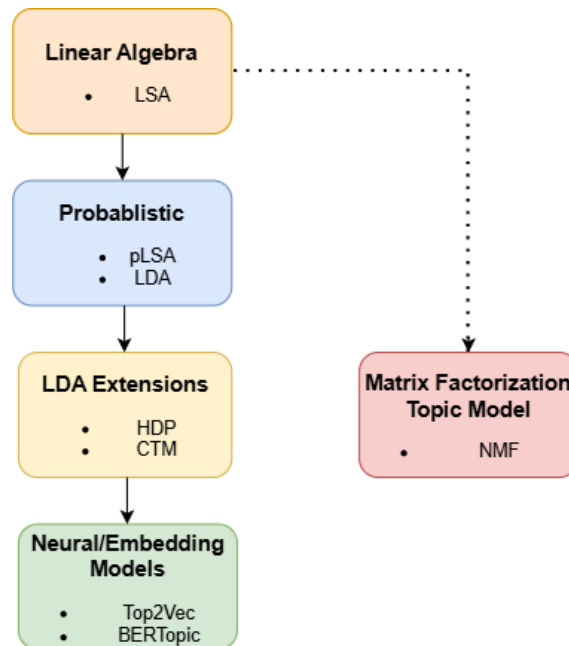


Figure 2.1: Evolution of topic modeling.

The method proposed in this work, BERTilda, follows the embedding-based paradigm and extends it to track the evolution of topics over time, as described in detail in Chapter 4.

2.1. Embedding-Based Topic Models

Recent advances in representation learning have led to a shift from bag-of-words representations toward semantic embeddings learned by neural language models. This shift also changes the overall topic modeling pipeline, as illustrated in Figure 2.2.

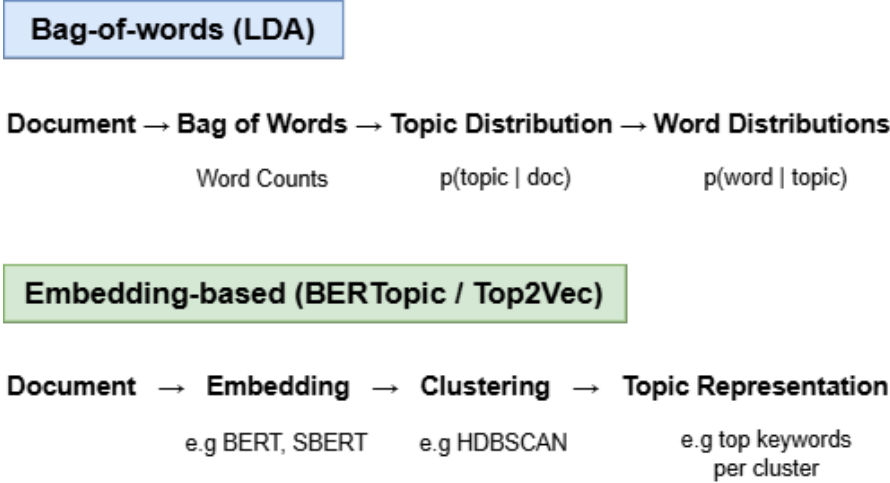


Figure 2.2: Topic Modeling Pipelines.

Pretrained models such as BERT [16] and Sentence-BERT (SBERT) [17] encode text into dense vector representations that capture semantic similarity between sentences and documents.

2.1.1 Top2Vec

Embedding-based topic models leverage these representations to identify clusters of semantically related documents. One example is Top2Vec [6], which is based on the assumption that topics correspond to dense regions in a semantic vector space. The method learns joint representations of documents and words, typically using Doc2Vec, such that semantic similarity is reflected by distance in the embedding space.

Topic discovery is performed by identifying dense clusters of document vectors, commonly after dimensionality reduction with UMAP and clustering with HDBSCAN. Each cluster is interpreted as a topic, represented by the centroid of its document vectors, and described by the words closest to that centroid in the embedding space.

Top2Vec does not require a predefined number of topics and relies on minimal preprocessing, as topic structure emerges directly from the geometry of the learned representations. The model also supports estimating topic sizes and hierarchical topic reduction by merging similar topics into more general themes.

2.1.2 BERTopic

A more recent and widely adopted framework is BERTopic [7], which builds on transformer-based embeddings while introducing a modular pipeline for topic construction and representation. The method typically relies on Sentence-BERT embeddings to encode documents into a semantic vector space, followed by dimensionality reduction techniques such as UMAP [18], which projects high-dimensional embeddings into a lower-dimensional space while preserving local neighborhood structure and semantic relationships between documents. Density-based clustering algorithms, most commonly HDBSCAN, are then applied to identify groups of semantically similar documents without requiring a predefined number of topics. The overall pipeline is illustrated in Figure 2.3.

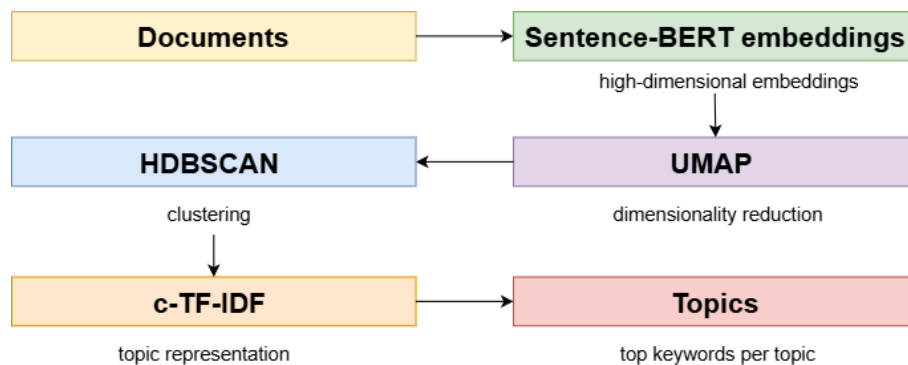


Figure 2.3: BERTopic Pipeline.

A key contribution of BERTopic lies in its separation between semantic grouping and topic representation. Instead of deriving topics directly from the embedding space, the model employs a class-based TF-IDF (c-TF-IDF) weighting scheme to extract words that are most distinctive for each cluster relative to the corpus. This approach improves interpretability by emphasizing terms that characterize each topic while down-weighting common or uninformative words.

The modular design of BERTopic allows for flexible customization of each stage of the pipeline, including the choice of embedding models, clustering algorithms, and topic representation methods. Despite its strong performance, the method may face limitations

when dealing with documents that cover multiple topics, as it assumes that each document can be assigned to a single dominant cluster representation. However, this limitation is less critical in short-text corpora such as tweets, where documents are typically focused on a single theme. In such settings, BERTopic has been shown to perform effectively [19], particularly compared to traditional bag-of-words approaches that struggle with sparse co-occurrence patterns. By leveraging contextual embeddings, the model captures semantic similarity even in the absence of shared vocabulary.

Other research has explored integrating embeddings directly into the probabilistic modeling framework. The Embedded Topic Model (ETM) [8] represents topics in continuous embedding spaces, enabling improved modeling of semantic relationships between words. Contextualized topic models have also been used [20].

2.2. Topic Modeling for Short and Noisy Texts

Many classical topic models were originally designed for relatively long documents such as scientific articles or news reports. However, modern text collections increasingly consist of short and informal messages, particularly on social media platforms such as X (formerly Twitter). Short texts pose significant challenges for topic modeling because they contain very few words, resulting in sparse document-term representations and limited co-occurrence information. Several strategies have been proposed to mitigate this problem.

2.2.1 Aggregation-Based Approaches

One family of methods addresses sparsity by aggregating short documents into longer pseudo-documents before applying topic models. Documents can be grouped by author [21] or by hashtag [22], creating richer representations that provide sufficient co-occurrence signal. While effective, this strategy introduces a design decision regarding aggregation criteria and may obscure within-group variation.

2.2.2 Alternative Approaches

Rather than relying on explicit aggregation criteria, several methods have been proposed to address short-text sparsity directly at the modeling level. Some approaches focus specifically on social media data. For example, Twitter-LDA [23] incorporates metadata such as hashtags and user information to improve topic discovery in Twitter streams. It

assumes that each tweet expresses a single topic, which is more appropriate than the standard LDA assumption of mixed-topic documents for such short texts.

Other methods address sparsity by modeling word co-occurrence patterns across the entire corpus. The *Biterm Topic Model* (BTM) [24], for instance, generates biterms (unordered pairs of words) rather than individual words. This allows the model to share statistical information across the entire corpus, effectively circumventing the per-document sparsity problem. BTM has been applied successfully to Twitter data and other short-text collections and demonstrates competitive performance compared with LDA [24].

A different perspective is adopted by sparse prior approaches such as the *Sparse Additive Generative Model* (SAGE) [25], which encourages most topic-word weights to be zero. This leads to more focused and interpretable topic representations even when only limited training data is available.

More recently, the availability of pretrained word and sentence embeddings has transformed approaches to topic modeling for short texts. By mapping documents into dense semantic spaces, embedding-based methods can leverage distributional similarity learned from large external corpora to compensate for sparse local co-occurrence information [26].

Some approaches extend probabilistic topic models by incorporating semantic information derived from embeddings. For example, the *GPU-DMM* model [27] enhances the Dirichlet Multinomial Mixture framework by promoting the assignment of semantically similar words to the same topic using pretrained word embeddings. This enables the model to exploit semantic relations between words for more effective topic discovery in short and noisy texts.

2.3. Dynamic Topic Models

Many real-world text collections evolve over time, reflecting changes in public discourse, scientific research trends, or social events. Dynamic topic models aim to capture these temporal dynamics by modeling how topics change across time.

One of the earliest approaches in this direction is the *Dynamic Topic Model* (DTM) introduced by Blei and Lafferty [2]. In this model, the parameters of each topic evolve over time according to a state-space model, allowing gradual changes in word distributions.

Another influential approach is *Topics over Time* (TOT) [3], which models the temporal

distribution of documents associated with each topic using continuous timestamp distributions. Continuous-time extensions further generalize these models to handle irregular temporal intervals [4].

For streaming scenarios, where data arrives continuously and revisiting the full corpus is computationally impractical, online variants update topic representations incrementally without reprocessing the entire history [28, 29]. These approaches enable scalable topic modeling in dynamic environments, such as social media streams, where topics evolve rapidly over time. Surveys further summarize methods for topic detection and tracking in such settings, [30].

More recent work has incorporated neural architectures into dynamic topic modeling. For example, the *Dynamic Embedded Topic Model* (DETM) [5] represents topics in embedding space while modeling their temporal evolution. Another approach introduces structured dependencies that can support branching/merging behavior [31, 32]. Contrastive and chain-free evolution-tracking formulations further relax one-to-one temporal identities [33].

Despite these advances, many dynamic topic models primarily capture smooth temporal changes in topic distributions. As a result, they may struggle to represent discrete structural changes such as topic splits, merges, or sudden topic emergence.

2.4. Tracking Structural Change in Clusters and Topics

The problem of tracking how clusters or communities evolve over time is not unique to text analysis. In network science, a substantial body of work has studied how communities in dynamic graphs change across snapshots, providing both conceptual frameworks and algorithmic tools that are directly applicable to topic tracking in temporal corpora.

2.4.1 Community Detection in Static Graphs

Before addressing temporal evolution, it is useful to review community detection in static graphs, as temporal methods typically extend static approaches. Classical algorithms such as the Louvain method [34] and spectral clustering [35] identify groups of nodes that are more densely connected internally than with the rest of the graph.

In the context of topic modeling, topics can be conceptualized as communities of semantically related documents. This analogy motivates the use of clustering algorithms from

network science for topic discovery, and conversely suggests that frameworks developed for community evolution can be applied to topic evolution.

2.4.2 Frameworks for Community Evolution

Palla et al. [36] introduced one of the earliest systematic studies of community evolution in dynamic networks, describing how communities may grow, contract, merge, or split over time. The work further discusses the emergence and disappearance of communities across temporal snapshots, laying the conceptual foundation for later lifecycle taxonomies used in dynamic community analysis.

These frameworks are directly applicable to topic evolution analysis, as topics in temporal corpora can be interpreted as dynamic communities of semantically related documents, where events such as merging and splitting reflect the convergence, diversification, or transformation of discussion themes over time.

2.4.3 Cluster Evolution and Topic Tracking

2.4.3.1 Evolutionary Clustering Objectives

Beyond probabilistic modeling, cluster evolution has been studied via objectives that balance snapshot fit and temporal smoothness. Evolutionary clustering [37] formalizes this as a dual objective, a clustering solution should accurately represent the current data snapshot while remaining consistent with the cluster structure observed at the previous time step. The trade-off between these two criteria is controlled by a smoothness parameter, which determines how much the solution is allowed to deviate from the historical structure in order to fit the current data. This formulation is particularly relevant in settings where abrupt changes in cluster assignments may reflect genuine structural transitions or simply stochastic fluctuation, and where the goal is to distinguish between the two.

2.4.3.2 Snapshot Matching for Split and Merge Detection

A complementary line of work addresses the problem of detecting discrete structural events by matching clusters across consecutive temporal snapshots. Rather than optimizing a global smoothness objective, these methods post-process independently computed snapshot clusterings by establishing correspondences between clusters at adjacent time steps using set-based or vector-based similarity measures.

Greene et al. [38] proposed a community matching framework in which clusters at consecutive snapshots are linked via Jaccard-based overlap of their member sets. Lifecycle events such as birth, death, merge, split, expansion and contraction are then identified from the resulting inter-snapshot community matchings. A notable property of this approach is its independence from any specific clustering algorithm, allowing it to operate over both disjoint and overlapping community structures.

Asur et al. [39] similarly formalized a taxonomy of community lifecycle events, defining continuation, appearance, disappearance, join, leave, split, merge, formation, and dissolution in terms of membership overlap between consecutive snapshots. These taxonomies were further extended and formalized by Bródka et al. [40], who introduced the GED framework for tracking group evolution in temporal networks. Their method defines a set of events including continuing, growing, shrinking, splitting, merging, dissolving, and forming, using an inclusion measure that accounts for both member overlap and node importance within groups.

Visual analytics tools have been used to highlight branching and convergence patterns in topic streams. TextFlow [41] visualizes topic evolution through a flow metaphor that highlights branching and convergence patterns between topics over time, facilitating the analysis of complex transitions in large document collections. Another visual analytics system is HierarchicalTopics [42], which organizes evolving topics into hierarchical structures to reveal thematic relationships and their temporal changes.

Many snapshot matching approaches rely exclusively on pairwise similarity between clusters, which may lead to ambiguous correspondences in dense semantic spaces. BERTilda builds on this snapshot matching paradigm but adds a document-level flow constraint. Rather than relying solely on embedding similarity between topic centroids, matches are further conditioned on the proportion of documents that migrate between candidate topic pairs across consecutive windows. This coverage criterion helps distinguish substantive structural transitions from weak semantic proximity between otherwise similar topics.

2.5. Evaluation of Topic Models

Evaluating the quality of topic models is a challenge. Unlike supervised learning, where prediction accuracy provides a clear optimization target, topic modeling produces outputs whose quality can only be assessed indirectly, through proxy metrics or human judgment.

2.5.1 Automatic Coherence Metrics

Automated coherence measures are the most widely used proxies for topic quality. These metrics assess how semantically related the top words of a topic are to one another, typically by measuring the co-occurrence or distributional similarity of word pairs. The most commonly used coherence formulations include UMass [43], UCI, NPML, and CV, each differing in the reference corpus used and the specific association measure computed [44].

Coherence measures correlate imperfectly with human judgments [45, 46], and different coherence formulations can behave inconsistently across models and domains [47, 48]. A fundamental difficulty is that these metrics capture lexical coherence but may miss semantic coherence expressed through synonymy or paraphrase, which is particularly relevant for short texts where low word overlap does not necessarily imply different meanings.

Recent work proposes contextualized coherence using masked language models [49] and purpose-oriented evaluation enabled by large language models [50].

2.5.2 Topic Diversity

While coherence captures the internal semantic consistency of individual topics, it does not account for redundancy across topics. Models may produce highly coherent topics that substantially overlap in vocabulary, reducing their interpretability and usefulness.

To address this limitation, topic diversity metrics are used as a complementary measure of topic quality. These metrics quantify the extent to which top words differ across topics, penalizing redundant or highly overlapping representations. The most widely adopted formulations include:

- Proportion of Unique Words [8], defined as the fraction of distinct terms among the top- k words across all topics
- Inverted Rank-Biased Overlap (IRBO) [51], which accounts for the rank ordering of words rather than treating topic lists as unordered sets, assigning lower scores to topics with overlapping high-ranked terms and higher scores to more dissimilar topics.

2.5.3 Temporal Stability Metrics

Evaluating dynamic topic models requires metrics that extend beyond within-window topic quality to assess the temporal behavior of learned representations. A model may produce coherent topics at each individual timestep while still failing to support meaningful longitudinal analysis if those representations fluctuate excessively across successive windows. Instability of this kind obstructs the identification of persistent themes, distorts the interpretation of topic transitions, and undermines the reliability of downstream tasks such as trend detection and lifecycle modeling [52, 53].

A family of temporal diagnostics concerns *topic coherence over time* [54]. Rather than computing coherence independently at each timestep, these measures assess whether semantically coherent structures are preserved across successive temporal intervals. Sustained coherence is particularly critical in long-running corpora, where gradual vocabulary shift can otherwise produce noisy or fragmented topic trajectories.

A complementary family of measures addresses *topic drift*, quantifying the magnitude and direction of change in topic representations between consecutive time windows. Drift analysis is closely related to the concept drift literature in streaming and evolving data environments [55, 56], where the central challenge is to distinguish genuine distributional change from stochastic fluctuation. In the context of topic modeling, drift metrics help characterize whether topic evolution reflects genuine thematic transitions in the corpus or instability introduced by the modeling process itself [57].

In this work, we combine within-window topic quality metrics with temporal diagnostics (drift and volatility) and human validation of discrete lifecycle labels. This multi-faceted evaluation approach addresses the limitations of any single metric and provides a more complete picture of model performance across both the static and temporal dimensions.

2.6. Political Discourse Analysis

The application domain of this thesis is political communication, a field with a rich tradition of computational and quantitative analysis. Understanding how political actors discuss issues, frame problems, and respond to events is central to political science, communication studies, and the emerging field of computational social science.

2.6.1 Agenda-Setting and Framing

Two foundational concepts in political communication research are *agenda-setting* and *framing*. Agenda-setting theory [58] holds that the media and political actors influence public perception not necessarily by telling people what to think, but by determining which issues receive attention. Framing theory [59] extends this idea by noting that the way an issue is presented, the language used, the causal attributions made, and the proposed remedies all shape how audiences interpret it.

Computational approaches to detecting frames in political text have proliferated with the availability of large-scale political corpora. Card et al. [60] developed the Media Frames Corpus, a large annotated dataset of news articles covering immigration policy, where each article is labeled according to the primary framing strategy employed. Supervised classifiers trained on this corpus have been applied to detect framing in social media and legislative speech.

In the context of the present thesis, the concept of framing is particularly relevant to the interpretation of topic evolution events. A topic split, for example, may correspond not merely to a thematic fragmentation but to the emergence of competing frames around the same underlying issue: one thread emphasizing economic consequences, another emphasizing security concerns. Tracking such divergences provides insight into how political actors strategically reframe issues over time.

2.6.2 Political Polarization and Ideological Analysis

A growing body of computational research has examined political polarization, the degree to which political actors and the public occupy increasingly distinct ideological positions. Gentzkow et al. [61] developed a measure of partisan language by identifying the vocabulary most strongly associated with Republican and Democratic legislators in the U.S. Congressional Record, demonstrating a marked increase in partisan polarization over recent decades.

Related work has used topic models to study the structure of political discourse [62], analyzing how members of Congress allocate their attention across policy domains.

Social media has intensified the study of polarization because it offers near-real-time access to the public statements of political figures and citizens. Bail et al. [63] found that

exposure to opposing viewpoints on Twitter can actually increase rather than reduce ideological polarization, a finding that underscores the importance of understanding the structure and dynamics of political discourse on these platforms.

2.6.3 Computational Analysis of Congressional Communication

Members of the U.S. Congress have increasingly used social media as a primary channel for communicating with constituents, amplifying political messages, and responding to current events [64]. Congressional Twitter behavior has attracted significant research attention, examining how legislators use the platform [65].

Mendonça and Figueira [19] applied BERTopic to tweets from members of the 117th U.S. Congress, demonstrating the viability of transformer-based topic models for extracting coherent thematic structures from the short and highly noisy text produced by political actors on social media. That work serves as a direct methodological predecessor to the present thesis, which extends the analysis to the 119th Congress and adds a temporal dimension focused on detecting how topics split, merge, and disappear over time.

The preceding review highlights several open challenges that motivate the present work. Classical dynamic topic models, such as Dynamic Topic Model (DTM) and Topics Over Time (TOT), are effective at capturing gradual changes in topic distributions over time. However, these probabilistic approaches generally assume smooth temporal evolution and are therefore limited in their ability to identify more abrupt structural changes, such as continuation, merge, split, and disappearance events. In contrast, research on community evolution provides a principled framework for modeling these transitions, but has been applied mainly to network-based data and has not been integrated with embedding-based topic models designed for short and noisy text corpora.

Another important gap concerns evaluation. Existing studies tend to focus either on static topic quality metrics, such as coherence and diversity, or on qualitative analyses of topic evolution over time. More comprehensive evaluation frameworks that jointly assess topic quality, temporal stability, and the interpretability of lifecycle events remain underexplored.

This thesis addresses these limitations through the proposal of BERTilda, a framework that combines embedding-based topic modeling with the detection of topic transitions in political text corpora, while evaluating the resulting representations across both static and temporal dimensions.

3. Problem Formulation

Understanding how themes evolve in longitudinal text collections is a central challenge in temporal text analysis. In many domains, such as political communication, news reporting, or social media discussions, the topics addressed by authors change continuously over time. These changes may involve gradual shifts in vocabulary, the emergence of new themes, or structural transformations where previously unified topics fragment into multiple subtopics or where distinct topics converge into a broader narrative.

This thesis focuses on the problem of detecting and characterizing the evolution of topics over time in temporal text corpora. In particular, the goal is to identify not only the persistence or disappearance of topics but also structural lifecycle events such as splits and merges, which correspond to significant reorganizations of thematic structure.

To formalize this problem, this chapter introduces the key concepts and notation used throughout the thesis. We first define the notion of a temporal text corpus and describe how such corpora are segmented into temporal windows. We then introduce the concept of snapshot topics obtained independently within each window. Based on these definitions, we formalize the problem of aligning topics across consecutive time windows and define the lifecycle events that characterize topic evolution. Finally, we discuss the main challenges and ambiguities inherent in this problem.

3.1. Temporal Text Corpora

Many real-world text datasets are not static snapshots of knowledge but rather the product of ongoing discourse, accumulating documents continuously over months, years, or even decades. Political speeches, parliamentary records, news archives, and social media platforms all produce text in this way. Each document carries a creation date, and together they form a chronological record of how people discussed, debated, and responded to the world around them.

The temporal dimension is not simply a detail contained in the metadata. It plays a fundamental role in how the content itself should be understood. A speech delivered in 2001 and another delivered in 2023 may employ similar vocabulary, yet they can refer to very different events, priorities and political contexts. For this reason, the meaning of a topic cannot be separated from the moment in which it emerges. Analyzing these documents

without considering their chronological context would overlook important changes in how issues are discussed and understood over time.

In this work, we consider a temporal text corpus, defined as a collection of timestamped documents. Formally, let

$$\mathcal{D} = \{d_1, d_2, \dots, d_N\} \quad (3.1)$$

be a corpus of N documents. Each document d_i is associated with a timestamp $\text{time}(d_i)$, indicating when the document was produced.

The temporal dimension of the corpus allows us to analyze how thematic content evolves over time. Rather than treating the corpus as a static collection, we aim to study how topics appear, persist, transform, or disappear as new documents are generated.

Temporal text corpora may vary widely in document length and production frequency. For instance, social media datasets often contain very short texts produced at high frequency, while political speech corpora may consist of longer documents produced less frequently. Despite these differences, the fundamental problem remains the same: identifying the evolution of themes across time.

To make temporal analysis tractable, the corpus is typically organized along a timeline and segmented into a sequence of time intervals [2] [3], as illustrated in Figure 3.1.

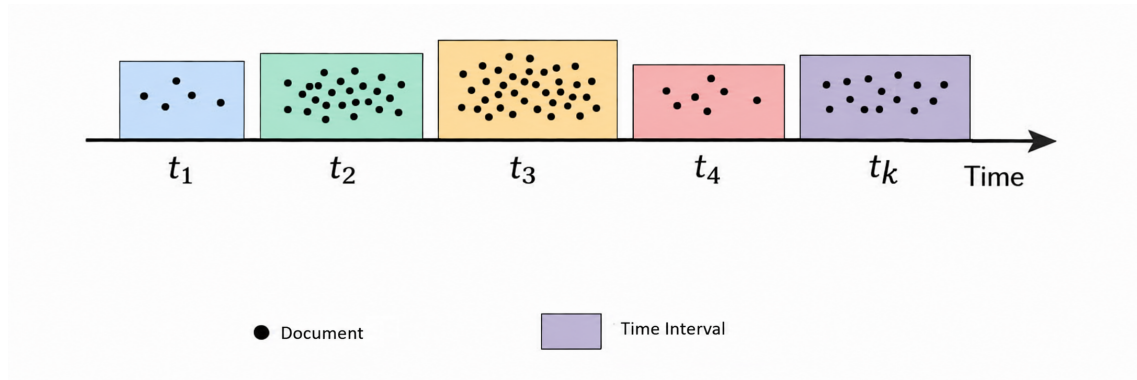


Figure 3.1: Documents distributed over time.

3.2. Temporal Segmentation

In order to analyze topic evolution, the temporal corpus must be divided into a sequence of time intervals, often referred to as temporal windows. Each window contains the subset of documents produced during a specific time period.

Let $T \in \mathbb{N}$ denote the total number of temporal windows obtained after segmentation. For each window t , we define its start and end times as s_t and e_t , respectively. The set of

documents belonging to window t is defined as

$$D_t = \{d \in \mathcal{D} : s_t \leq \text{time}(d) < e_t\} \quad (3.2)$$

This segmentation produces a sequence of document collections

$$D_1, D_2, \dots, D_T \quad (3.3)$$

that represent the corpus at different points in time.

Note: The temporal segmentation procedure does not guarantee coverage of the entire corpus up to the final timestamp. Specifically, the last time interval is included only if a full window of size w fits within the range. As a result, the final remaining documents may be excluded if they do not form a complete temporal window.

Temporal windows can be defined using different strategies. A common approach is to use fixed-length windows, such as weeks, months, or years, depending on the nature of the data. In a sliding window formulation, each window is characterized by a window length Δ_w and a step size Δ_s . The window length Δ_w determines the temporal span of each window, while the step size Δ_s defines the temporal distance between the starting points of consecutive windows.

$\Delta_s = \Delta_w$ corresponds to non-overlapping (disjoint) windows. When $\Delta_s < \Delta_w$, consecutive windows overlap, allowing documents to contribute to multiple temporal contexts. Overlapping windows can help smooth abrupt changes and provide a more continuous view of topic evolution, as shown in Figure 3.2.

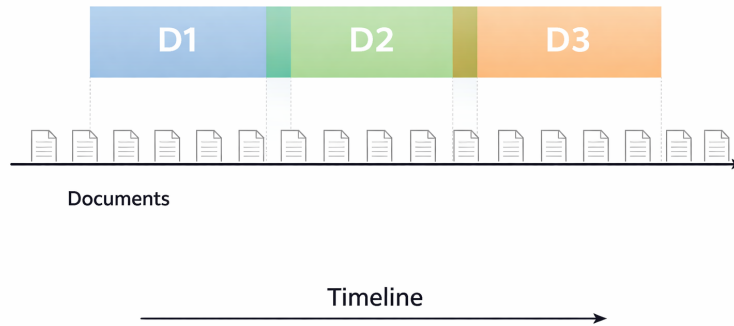


Figure 3.2: Temporal Window Segmentation.

The choice of temporal granularity has a significant impact on the analysis. If the windows are too large, short-term topic dynamics may be obscured. Conversely, very small

windows may produce unstable or fragmented topics due to insufficient data within each window. Choosing an appropriate granularity is, therefore, a critical design decision.

3.3. Snapshot Topics

Once the corpus has been segmented into temporal windows, topics can be extracted independently from each window. This approach treats each window as a snapshot of the corpus at a particular point in time.

Recall that D_t denotes the set of documents in window t . Applying a topic modeling method to this subset yields K_t topics, where K_t denotes the number of topics identified in window t . The topics are indexed by the set

$$\mathcal{K}_t = \{1, \dots, K_t\}. \quad (3.4)$$

Each topic represents a coherent semantic theme shared by a subset of documents in that window. Documents are assigned to topics according to the topic modeling method used. For each $d \in D_t$, let

$$z_t(d) \in \mathcal{K}_t \quad (3.5)$$

denote the topic assigned to document d . In probabilistic models, $z_t(d)$ may be obtained as $\arg \max_k P(k \mid d)$, whereas in embedding-based models it may correspond to the nearest-cluster assignment.

The set of documents associated with topic k in window t is therefore

$$D_{t,k} = \{d \in D_t : z_t(d) = k\} \quad (3.6)$$

Topics can be represented in various ways, depending on the modeling approach. In traditional probabilistic topic models [2] [10] [14] [1] [3], topics are often represented as probability distributions over words. In embedding-based topic models [7] [6] [8] [5], topics can instead be represented by vector representations derived from document embeddings.

Regardless of the specific representation, each window yields a set of topics that capture the dominant themes present in the documents of that time period. Figure 3.3 illustrates this concept by representing topics as circles associated with each temporal window.

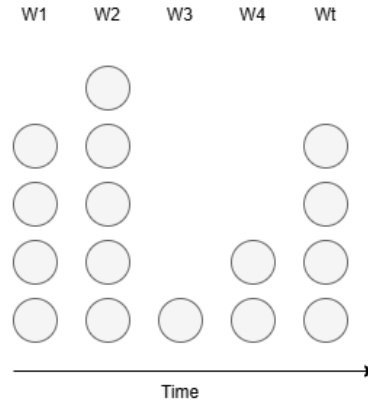


Figure 3.3: Example of snapshot Topics per Window W_t . Each circle represents a topic.

Because topics are discovered independently within each window, with no correspondence enforced between windows, their identities are not preserved across time.

3.4. Inter-window Topic Alignment

When topics are discovered independently within each temporal window, there is no inherent correspondence between the topics obtained at different time steps. In other words, the topic labeled k in window t is not guaranteed to represent the same theme as topic k in window $t + 1$. As a result, analyzing topic evolution requires solving a topic alignment problem across consecutive windows.

Given two sets of topics, $\mathcal{K}_t$ and $\mathcal{K}_{t+1}$, the goal is to determine which topics in window t correspond to which topics in window $t + 1$. Establishing these correspondences allows us to track how themes evolve over time.

However, this alignment is not straightforward. Several situations may arise when comparing topics across windows:

- A topic may persist with similar content across time.
- A topic may gradually shift in vocabulary while retaining its core theme.
- A topic may split into several more specialized subtopics.
- Multiple topics may merge into a broader discussion.
- A topic may disappear entirely if the associated discussion ceases.

Figure 3.4 presents an example of how topics across multiple temporal windows may relate to one another.

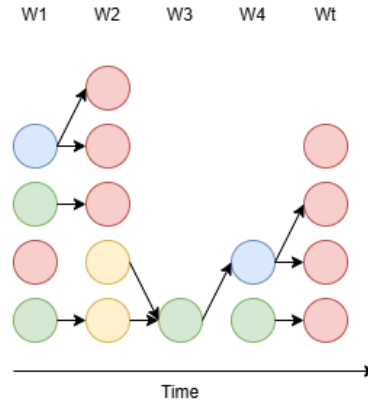


Figure 3.4: Topic Alignment over time.

Because snapshot models discover topics independently in each window, these structural changes are not explicitly modeled but must be inferred by comparing those obtained in consecutive windows.

The alignment problem therefore consists of determining which topics are related across time and identifying the nature of their relationships.

3.5. Topic Lifecycle Events

Once correspondences between topics across windows are established, it becomes possible to characterize how topics evolve over time through a set of lifecycle events. These events describe the structural transformations that may occur between topics in consecutive windows.

Continuation A continuation occurs when a topic persists across time with relatively stable semantics. In this case, a topic in window t corresponds to a single topic in window $t + 1$ that represents the same underlying theme.

Split A split occurs when a topic in window t gives rise to multiple topics in window $t + 1$. This situation typically arises when a broad discussion becomes fragmented into more specialized subtopics.

Merge A merge occurs when multiple topics in window t converge into a single topic in window $t + 1$. This may happen when previously separate discussions become tightly connected or when a broader narrative emerges.

Disappearance A disappearance occurs when a topic present in window t has no clear successor in window $t + 1$. This typically indicates that the corresponding discussion has faded from the corpus.

Unclear Transition In some cases, the relationship between topics across windows cannot be confidently determined. Such cases arise when none of the structural conditions for continuation, split, merge, or disappearance are met, and are labeled as unclear transitions.

Figure 3.5 presents a schematic representation of these possible transitions between consecutive windows.

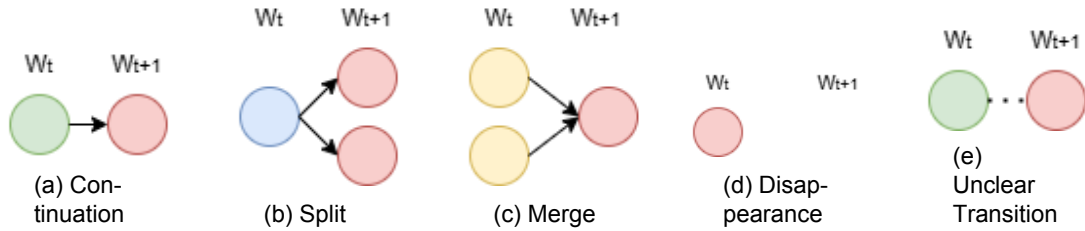


Figure 3.5: Illustration of the five topic lifecycle events.

These lifecycle events provide a structured framework for describing how thematic structures change over time within a temporal corpus.

3.6. Problem Statement

Given a temporal corpus $\mathcal{D}$ segmented into windows $\{D_1, \dots, D_T\}$ and the corresponding snapshot topic sets $\{\mathcal{K}_1, \dots, \mathcal{K}_T\}$, the problem addressed in this thesis is to find, for every pair of consecutive windows $(t, t + 1)$, a relation

$$R_t \subseteq \mathcal{K}_t \times \mathcal{K}_{t+1}$$

together with a labelling function

$$\ell_t : R_t \rightarrow \{\text{cont, split, merge, disap, unclear}\}$$

that jointly describe how topics evolve from window t to window $t + 1$.

3.7. Challenges and Ambiguity

Detecting topic evolution in temporal corpora presents several challenges. These challenges arise from both the nature of language and the properties of topic modeling methods.

One major difficulty is semantic ambiguity. Topic boundaries are often fuzzy, and documents may relate to multiple themes simultaneously. As a result, the relationship between topics across windows may not always be clear-cut.

Another challenge concerns topic granularity. The level of abstraction at which topics are discovered may vary across time. A topic identified in one window may later appear as several more specific subtopics, or multiple topics may be combined into a broader theme. Distinguishing between genuine structural changes and modeling artifacts is therefore difficult [66].

Split and merge events are particularly challenging to detect [67]. These events involve many-to-many relationships between topics across windows, which cannot be captured by simple one-to-one matching strategies.

Finally, many existing approaches to temporal topic modeling assume that topics evolve gradually over time [2] [5] [3]. While this assumption can capture smooth semantic drift, it does not naturally account for abrupt structural changes such as splits or merges.

We propose BERTilda, a topic-evolution detection method that addresses these challenges through bidirectional document attribution and rule-based lifecycle detection, as described in Chapter 4.

4. Methodology

This chapter presents the proposed methodology used to detect and characterize topic evolution in temporal text corpora. The approach, named **BERTilda** (BERT-based Temporal Identification, Lifecycle-Detection, and Analysis), models topic evolution as an alignment and event-detection problem across consecutive temporal windows.

The key idea behind BERTilda is to treat topics discovered independently in each window as nodes in a temporal structure and to establish relationships between them using both semantic similarity and document flow signals. These relationships are then used to construct a temporal topic graph and to identify lifecycle events such as continuation, split, merge, disappearance, or unclear transitions.

The chapter first provides an overview of the framework and then describes each component of the methodology in detail, including topic discovery, topic representation, cross-window alignment, and lifecycle event detection.

4.1. Overview of the Proposed Framework

The BERTilda framework is designed as a multi-stage pipeline that transforms a temporal text corpus into a structured representation of topic evolution.

Given a timestamped document corpus, the process consists of the following main steps:

1. Segment the corpus into temporal windows.
2. Discover topics independently within each window using a snapshot topic model.
3. Represent topics in a shared semantic embedding space.
4. Compute semantic similarity between topics in consecutive windows.
5. Estimate bidirectional document coverage between topics across windows.
6. Apply rule-based event detection to classify lifecycle events.
7. Construct a temporal topic graph based on the lifecycle events detected.

This process produces a directed temporal graph in which nodes correspond to topics at specific time windows and edges represent validated relationships between topics across time.

An overview of the BERTilda pipeline is presented in Figure 4.1.

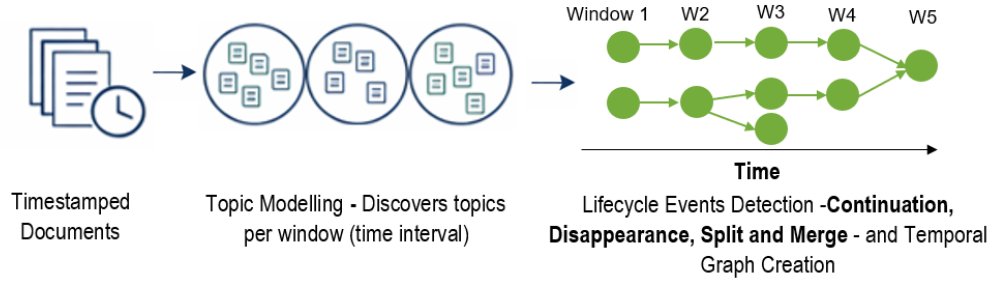


Figure 4.1: BERTilda's Pipeline

The framework is designed to be modular in order to allow flexibility in the topic discovery stage. In particular, the snapshot topic modeling component is not tied to a specific algorithm. Any topic discovery method can be integrated provided that it produces the core information required by the subsequent stages of the framework.

Specifically, the framework requires document embeddings that represent each document in a shared semantic vector space, topic embeddings that characterize each discovered topic (for example, computed as the centroid of the embeddings of the documents assigned to the topic), and document-topic assignments indicating which topic each document belongs to within a given temporal window.

These elements allow the framework to compute semantic similarity between topics across windows and to estimate document flows through the bidirectional coverage mechanism described in the following sections.

Topic descriptors such as top words can also be extracted to facilitate the interpretation of the discovered topics. However, these descriptors are not required for the temporal alignment procedure itself, which relies primarily on the embedding representations and the document-topic assignments.

4.2. Topic Discovery per Temporal Window

As described in Chapter 3, the temporal corpus is first divided into a sequence of windows

$$D_1, D_2, \dots, D_T \quad (4.1)$$

where each window contains the documents produced during a specific time interval.

Within each window t , topics are discovered independently using a snapshot topic model.

Let

$$\mathcal{K}_t = \{1, 2, \dots, K_t\} \quad (4.2)$$

denote the set of topics identified in window t .

Each document $d \in D_t$ is assigned to a topic according to

$$z_t(d) \in \mathcal{K}_t \quad (4.3)$$

and the set of documents belonging to topic k is defined as

$$D_{t,k} = \{d \in D_t : z_t(d) = k\} \quad (4.4)$$

In this work, topic discovery is implemented using **BERTopic** [7], an embedding-based topic modeling approach that combines transformer-based document embeddings [17], density-based clustering (Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [68]), and Class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) representations for topic descriptors.

Embedding-based topic models are particularly suitable for short and noisy texts, such as social media posts, because they capture semantic similarity beyond simple word co-occurrence patterns [8, 19, 69].

By discovering topics independently within each temporal window, BERTilda avoids imposing artificial temporal continuity. However, this design introduces an important challenge, as correspondences between topics across time must be explicitly determined.

4.3. Topic Representation and Similarity

Each document d is mapped to an embedding vector

$$\phi(d) \in \mathbb{R}^m \quad (4.5)$$

obtained using a pretrained sentence transformer encoder.

For each time window t , documents are grouped into topics. Let $D_{t,k}$ denote the set of documents assigned to topic k in window t .

The representation of a topic (t, k) is defined as the centroid of the embeddings of the documents assigned to that topic:

$$\psi(t, k) = \frac{1}{|D_{t,k}|} \sum_{d \in D_{t,k}} \phi(d) \quad (4.6)$$

Because all document embeddings are generated using the same encoder, topic representations lie in a shared semantic space, allowing topics from different windows to be directly compared.

The semantic similarity between topics from consecutive windows is computed using cosine similarity:

$$\text{sim}((t, k), (t + 1, j)) = \cos(\psi(t, k), \psi(t + 1, j)) \quad (4.7)$$

High similarity values suggest that the two topics may correspond to related themes across time.

Although semantic similarity provides an important signal for topic alignment, it is often insufficient to disambiguate complex structural relations such as splits or merges. For this reason, BERTilda introduces an additional signal based on document-level coverage across windows.

4.4. Bidirectional Coverage Estimation

Semantic similarity captures proximity between topic representations but does not directly measure how documents flow between topics over time.

To address this limitation, BERTilda introduces a *bidirectional coverage* mechanism that estimates how documents from one topic relate to topics in the next window.

This mechanism consists of two complementary components: forward attribution and backward attribution.

4.4.1 Forward Attribution (Outflow)

For each document $d \in D_{t,k}$, the most similar topic in the next window is identified:

$$j^*(d) = \arg \max_{j \in \mathcal{K}_{t+1}} \cos(\phi(d), \psi(t + 1, j)) \quad (4.8)$$

If the similarity exceeds a predefined threshold τ_{doc} , the document is considered to flow from topic (t, k) to topic $(t + 1, j^*)$.

Let

$$F_t^{fwd}(k \rightarrow j) \quad (4.9)$$

denote the number of documents from topic k in window t attributed to topic j in window $t + 1$.

The outflow coverage is defined as

$$cov_{out}(k \rightarrow j) = \frac{F_t^{fwd}(k \rightarrow j)}{|D_{t,k}|} \quad (4.10)$$

This quantity measures the fraction of documents from topic (t, k) that appear to transition toward topic $(t + 1, j)$.

4.4.2 Backward Attribution (Inflow)

Symmetrically, backward attribution estimates where the documents of a topic in window $t + 1$ originated.

For each document $d \in D_{t+1,j}$, the most similar topic in the previous window is identified:

$$k^*(d) = \arg \max_{k \in \mathcal{K}_t} \cos(\phi(d), \psi(t, k)) \quad (4.11)$$

Let

$$F_t^{bwd}(k \rightarrow j) \quad (4.12)$$

denote the number of documents in topic $(t + 1, j)$ attributed back to topic (t, k) .

The inflow coverage is defined as

$$cov_{in}(k \rightarrow j) = \frac{F_t^{bwd}(k \rightarrow j)}{|D_{t+1,j}|} \quad (4.13)$$

Together, outflow and inflow provide complementary perspectives on topic transitions. Outflow indicates where a topic appears to evolve toward, while inflow indicates the sources from which a topic emerges.

4.5. Event Labeling Rules

After computing semantic similarity and bidirectional coverage between topics in consecutive windows, BERTilda determines the relationships between topics using a set of rule-based criteria.

These rules identify how topics evolve between window t and window $t + 1$, allowing the detection of lifecycle events such as continuation, split, merge, disappearance, or unclear transitions.

A candidate relation between topic (t, k) and topic $(t + 1, j)$ is considered valid if it satisfies the following minimum conditions:

$$\text{sim}((t, k), (t + 1, j)) \geq \tau_{\text{topic}}, \quad \text{cov}_{\text{out}}(k \rightarrow j) \geq \tau_{\text{cov}}^{\text{out}}, \quad \text{cov}_{\text{in}}(k \rightarrow j) \geq \tau_{\text{cov}}^{\text{in}} \quad (4.14)$$

These thresholds ensure that both semantic similarity and document flow support the relationship.

Let

$$S_t(k) = \{j : (k \rightarrow j) \in \mathcal{E}_t\} \quad (4.15)$$

denote the set of successor topics in window $t + 1$ associated with topic (t, k) , and

$$P_{t+1}(j) = \{k : (k \rightarrow j) \in \mathcal{E}_t\} \quad (4.16)$$

the set of predecessor topics in window t associated with topic $(t + 1, j)$.

Based on these sets, lifecycle events are defined as follows.

4.5.1 Continuation

Topic (t, k) is labeled as *continued* if it has exactly one validated successor and the relation satisfies stricter similarity and coverage constraints:

$$\text{sim}(k, j) \geq \alpha_{\text{cont}}, \quad \text{cov}_{\text{out}}(k \rightarrow j) \geq \alpha_{\text{cont}}^{\text{out}}, \quad \text{cov}_{\text{in}}(k \rightarrow j) \geq \alpha_{\text{cont}}^{\text{in}} \quad (4.17)$$

where $S_t(k) = \{j\}$.

4.5.2 Split

A topic is labeled as a *split* if it has multiple validated successors and a substantial fraction of its document mass flows toward them:

$$|S_t(k)| \geq 2, \quad \sum_{j \in S_t(k)} \text{cov}_{\text{out}}(k \rightarrow j) \geq \alpha_{\text{split}}^{\text{out}}, \quad \text{cov}_{\text{in}}(k \rightarrow j) \geq \alpha_{\text{split}}^{\text{in}} \quad \forall j \in S_t(k). \quad (4.18)$$

This condition captures situations where a previously unified topic fragments into several more specialized subtopics.

4.5.3 Disappearance

A topic is considered to *disappear* if it has no validated successor, or if the total document flow toward the next window is negligible in either direction, indicating that the topic's content does not meaningfully persist:

$$S_t(k) = \emptyset \quad \text{or} \quad \sum_{j \in S_t(k)} \text{cov}_{\text{out}}(k \rightarrow j) < \alpha_{\text{disp}}^{\text{out}} \quad \text{or} \quad \sum_{j \in S_t(k)} \text{cov}_{\text{in}}(k \rightarrow j) < \alpha_{\text{disp}}^{\text{in}}. \quad (4.19)$$

This situation typically indicates that the discussion associated with the topic fades from the corpus.

4.5.4 Merge

Separately, topic $(t+1, j)$ is labeled as a *merge target* if it receives substantial inflow from multiple predecessor topics:

$$|P_{t+1}(j)| \geq 2, \quad \sum_{k \in P_{t+1}(j)} \text{cov}_{\text{in}}(k \rightarrow j) \geq \alpha_{\text{merge}}^{\text{in}}, \quad \text{and} \quad (4.20)$$

$$\text{cov}_{\text{out}}(k \rightarrow j) \geq \alpha_{\text{merge}}^{\text{out}} \quad \forall k \in P_{t+1}(j).$$

This condition captures cases where multiple previously distinct topics converge into a broader discussion.

4.5.5 Unclear Transitions

If none of the previous conditions are satisfied, the transition is labeled as *unclear*. Such cases typically arise when similarity or coverage signals are weak or ambiguous, making it difficult to confidently assign a specific lifecycle event.

4.6. Temporal Topic Graph Construction

Once topic relationships and lifecycle events have been identified, BERTilda constructs a temporal topic graph that summarizes the evolution of topics across time.

In this graph, each node represents a topic associated with a specific temporal window (t, k) . Directed edges represent validated relationships between topics in consecutive windows.

An edge $(t, k) \rightarrow (t + 1, j)$ is created whenever a valid relation between the two topics has been detected according to the similarity and coverage constraints described previously.

The resulting graph therefore captures the structural evolution of topics through time. Different patterns of connectivity correspond to different lifecycle events:

- A single edge between two topics represents a continuation.
- Multiple outgoing edges from a topic correspond to a split event.
- Multiple incoming edges to a topic correspond to a merge event.
- Topics without outgoing edges correspond to disappearance events.
- Dashed edges represent unclear transitions.

This temporal graph provides a compact and interpretable representation of topic evolution, allowing researchers to trace how themes persist, branch, merge, or vanish across time. Figure 4.2 illustrates an example of such a graph across consecutive windows.

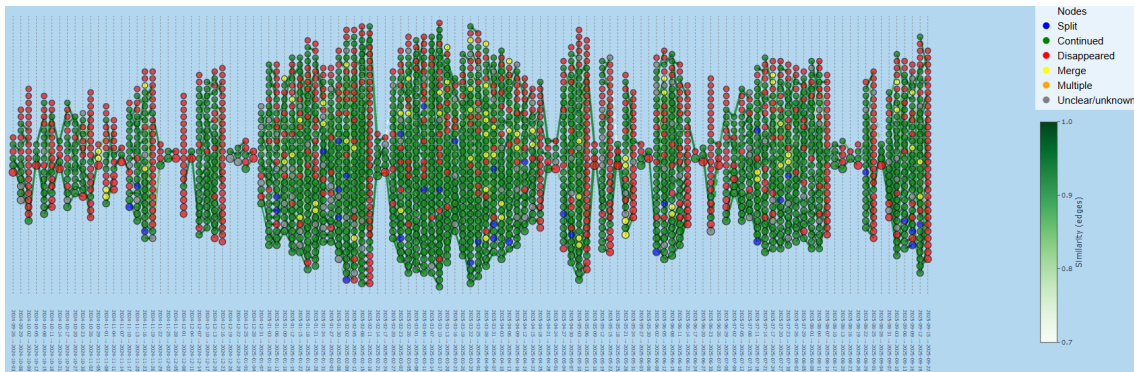


Figure 4.2: Example of a temporal topic graph illustrating topic evolution across consecutive time windows.

4.7. BERTilda Algorithm

The complete procedure implemented by BERTilda is summarized below.

Algorithm 1 BERTilda: temporal topic graph and lifecycle labeling

Require: Timestamped corpus $\mathcal{D}$; window scheme $(\Delta_{\text{window}}, \Delta_{\text{step}})$; thresholds

$\tau_{\text{doc}}, \tau_{\text{topic}}, \tau_{\text{cov}}^{\text{out}}, \tau_{\text{cov}}^{\text{in}}$.

- 1: **for** $t = 1$ **to** T **do**
 - 2: Fit a snapshot topic model on $\mathcal{D}_t$; obtain topics $\mathcal{K}_t$ and representations $\psi(t, \cdot)$.
 - 3: **end for**
 - 4: **for** $t = 1$ **to** $T - 1$ **do**
 - 5: Compute topic similarities $\text{sim}((t, k), (t+1, j))$.
 - 6: Compute bidirectional coverage $\text{cov}_{\text{out}}(k \rightarrow j)$ and $\text{cov}_{\text{in}}(k \rightarrow j)$ by cross-window attribution.
 - 7: Retain edges by Eq. 4.14; define $S_t(k)$ and $P_{t+1}(j)$ by Eq. 4.15 and Eq. 4.16.
 - 8: Label continuation/split/disappearance/merge/unclear.
 - 9: **end for**
-

4.8. Implementation Considerations

For short texts, document embeddings are computed using a fixed sentence-transformer encoder (*all-MiniLM-L6-v2*) [70] within each window, ensuring that topic representations lie in a consistent semantic space across time.

Topic discovery uses BERTopic (transformer embeddings, HDBSCAN clustering, and c-TF-IDF descriptors), with a minimum topic size to filter noisy micro-clusters.

For each adjacent window pair, the dominant cost is cross-window attribution, which naively requires comparing each document embedding to all topic embeddings in the neighboring window. In practice, K_t is typically far smaller than $|\mathcal{D}_t|$, and efficient matrix multiplication makes attribution practical. Candidate edges are additionally pruned by

$\tau_{\text{topic}}, \tau_{\text{cov}}^{\text{out}}$, and $\tau_{\text{cov}}^{\text{in}}$.

Further implementation details and parameter configurations used in the experiments are described in Chapter 5. The complete implementation, including the source code, datasets and interactive dashboard, is publicly available at: <https://github.com/ossclaudia/BERTilda>.

5. Experimental Design

This chapter describes the experimental design used to evaluate the proposed methodology, covering the research questions guiding the evaluation, the datasets and preprocessing pipeline, the experimental configuration, the baselines used for comparison, and the evaluation metrics employed to assess both topic quality and event detection. Qualitative case studies are also outlined. Results are presented and discussed in the following chapter.

5.1. Research Questions

The evaluation of the proposed framework is guided by several research questions that aim to assess different aspects of the methodology, including topic quality, temporal robustness, and the validity of detected topic evolution events.

RQ1: Topic Quality. Does the snapshot topic discovery component yield coherent and diverse topics on short political texts, relative to common baselines?

RQ2: Temporal Stability. How stable are topic-quality diagnostics over time compared to dynamic topic models with explicit temporal priors?

RQ3: Event Detection Validity. Are the detected topic evolution events (continuations, splits, merges, and disappearances) semantically valid when evaluated by human annotators?

RQ4: Comparative Performance. How does the proposed approach compare with alternative event detection strategies in terms of event detection accuracy?

RQ5: Robustness. How sensitive is the event detection process to the thresholds used for topic alignment and event classification?

RQ6: Narrative Interpretability. Do qualitative case studies illustrate that detected events correspond to meaningful narrative reorganizations?

These research questions directly inform the choice of datasets, baselines, and evaluation metrics described in the remainder of this chapter.

5.2. Datasets

The evaluation was conducted using three corpora representing different forms of political discourse and temporal resolutions. These datasets allow the proposed approach to be tested across both short-term and long-term temporal scenarios. Table 5.1 summarizes the main characteristics of the datasets used in this study.

Table 5.1: Summary of the datasets used.

Dataset	#Docs	Time Span	Type	Use
Congressional Tweets	357,896	2024–2025	Social media (short)	Main evaluation
UN General Debates	7,507	1970–2015	Political speeches	Temporal models
State of the Union	228	1790–2018	Political speeches	Temporal models

5.2.1 Congressional Tweets

The first dataset consists of tweets published by members of the 119th United States Congress during the period 2024–2025. This dataset captures short and highly dynamic political discourse on social media, making it particularly suitable for evaluating fine-grained topic evolution and event detection.

The data extraction was performed using TwExportly [71], a Google Chrome extension designed to export tweets from any account on X [72]. Initially, a dataset was manually created containing the personal information of members of the 119th U.S. Congress [73]. Subsequently, an extensive search was conducted to identify the X accounts associated with these members. Whenever available, professional accounts were prioritized; otherwise, personal accounts were used. Out of the 545 members of Congress, 544 accounts were successfully identified, leaving only one member without a known X account.

TwExportly allows the extraction of tweets within a specified timeframe. Although the 119th Congress officially began on January 3, 2025 [74], the collection period was extended in order to capture discussions both before and during the congressional term. Tweets were therefore extracted from September 25, 2024, to September 24, 2025, the day before the extraction process was conducted.

The extraction process resulted in 544 individual datasets, each corresponding to a specific member of Congress and containing their tweets along with associated metadata such as the number of favorites and the posting date. In addition, the dataset containing

the personal information and X account of each member was used to merge all individual datasets into a single combined dataset that includes both the members' personal information and their tweet data.

The temporal segmentation used for this and all datasets is described in Section 5.4.

5.2.2 UN General Debates

The second dataset corresponds to the United Nations General Debate corpus [75], which contains speeches delivered by world leaders during the annual UN General Assembly sessions between 1970 and 2015.

5.2.3 State of the Union Addresses

The third dataset consists of State of the Union addresses [76] delivered by U.S. presidents between 1790 and 2018. These speeches represent one of the longest historical records of political discourse.

5.3. Data Preprocessing

Before applying topic modeling, all datasets were subjected to preprocessing procedures designed to remove noise, normalize textual content, and reduce lexical variability. The objective of this process was to obtain a consistent token-based representation suitable for topic discovery across political corpora.

Given the structural differences between the datasets in terms of document length and composition, the preprocessing pipeline was adapted to each corpus. However, the overall goal was to maintain comparable token-level representations suitable for modeling.

5.3.1 State of the Union Addresses

The State of the Union corpus consists of long political speeches, which motivated a preprocessing strategy that emphasizes segmentation into finer-grained textual units and vocabulary filtering.

First, a corpus-level vocabulary filtering step was applied using document frequency statistics. A `CountVectorizer` model was used to identify terms appearing in more than 70% of the documents, which were subsequently removed to reduce the influence of extremely common terms.

Each speech was then segmented into paragraphs, enabling the decomposition of long speeches into smaller thematic units and improving topic granularity.

Each paragraph was processed by converting text to lowercase and removing punctuation and numeric characters. The resulting text was then analyzed using the spaCy English language model (`en_core_web_sm`) to perform lemmatization and remove stopwords as well as non-informative tokens.

Paragraphs containing fewer than four tokens after preprocessing were discarded to eliminate extremely short or uninformative segments.

Finally, each paragraph was represented as a list of tokens and associated with the corresponding year of the speech.

5.3.2 UN General Debates

The United Nations General Debate corpus required a two-pass preprocessing strategy in order to compute corpus-level vocabulary statistics prior to paragraph segmentation.

In the first pass, full speeches were preprocessed by converting text to lowercase and removing punctuation, digits, and control characters. The cleaned text was then processed using spaCy to perform tokenization and lemmatization, while removing stopwords and tokens shorter than two characters.

Document frequency statistics were then computed over the entire corpus using a document-term matrix generated by a `CountVectorizer`. Terms appearing in fewer than ten documents or in more than 70% of the documents were removed from the vocabulary in order to eliminate both rare and overly common terms.

In the second pass, each speech was segmented into paragraphs. Each paragraph was then processed using the same cleaning and lemmatization pipeline, after which tokens not belonging to the filtered vocabulary were removed.

The final representation consists of tokenized paragraphs associated with the year of the corresponding speech.

5.3.3 Congressional Tweets

The congressional tweets dataset consists of short-form social media messages and therefore requires additional cleaning steps to remove platform-specific artifacts.

First, tweets were filtered to retain only English-language messages. The tweet text was then cleaned by removing HTML artifacts, URLs appearing directly in the text, mentions, emojis, and other non-alphabetic characters.

In the original dataset, URLs and hashtags are also stored in separate columns. Information from these fields was therefore incorporated by converting hashtag strings and relevant URL components into textual tokens when possible.

Subsequently, the resulting text was normalized to lowercase, stopwords were removed, and the remaining tokens were lemmatized using the spaCy English language model.

The final processed dataset consists of tweets represented as token lists associated with their corresponding timestamps (publication date).

Figure 5.1 provides an overview of the preprocessing workflow used for the congressional tweets dataset.

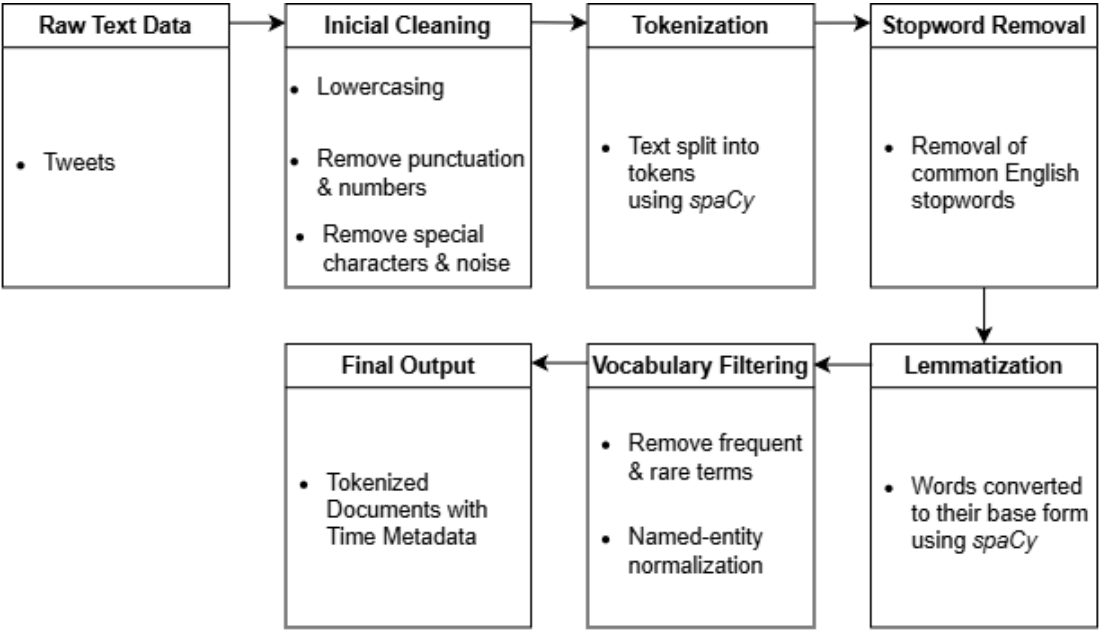


Figure 5.1: Overview of the preprocessing workflow used to prepare textual tweet data for topic modeling.

Table 5.2 reports the number of documents before and after preprocessing. Due to paragraph segmentation applied to long-form speeches, the number of documents increases for the UN General Debates and State of the Union datasets, while the congressional tweets dataset experiences a slight reduction due to filtering.

Table 5.2: Dataset sizes before and after preprocessing.

Dataset	Raw Documents	After Preprocessing
Congressional Tweets	357,896	345,865
UN General Debates	7,507	26,362
State of the Union	228	19,283

5.3.4 Overall Preprocessing Design

Overall, the preprocessing framework was designed to balance consistency with adaptation to each dataset. While the exact steps differ across corpora, all datasets are transformed into a shared format based on tokenized text units with temporal labels.

This approach ensures that all datasets can be used in a consistent way during topic modeling, while still respecting the differences in structure and language between them.

It should be noted that the Sentence-BERT encoder received token sequences reconstructed from lemmatized and preprocessed tokens, rather than minimally cleaned natural text. While the standard recommendation for BERTopic is to preserve syntactic structure for the embedding step [7], this approach was retained. The potential limitation is acknowledged and future work could assess the impact of applying the encoder to minimally preprocessed text.

5.4. Experimental Configuration

5.4.1 Temporal Windowing

To capture topic evolution over time, each dataset was segmented into overlapping temporal windows. The window size and step were selected according to the temporal granularity of each corpus:

Dataset	Time Span	Window Size	Step
Congress Tweets	2024–2025	7 days	3 days
UN General Debates	1970–2015	3 years	2 years
State of the Union	1790–2018	7 years	3 years

Table 5.3: Temporal segmentation used in the experiments.

Within each window, topics were independently discovered using BERTopic [7] with a transformer-based embedding model (MiniLM) and HDBSCAN clustering.

Topics containing fewer than a minimum number of documents were discarded in order to avoid noise. The minimum size was set to 50 documents for congressional tweets, 20 documents for the UN debates, and 1 document for State of the Union speeches.

Topic descriptors were defined using the top- N words obtained from the class-based TF-IDF representation.

5.4.2 Threshold Parameters

The event detection process relies on several thresholds controlling topic alignment and event classification.

Table 5.4: Default decision thresholds used in the alignment and event-labeling stages.

Symbol	Meaning	Value
<i>Similarity and edge validation</i>		
τ_{doc}	min doc-to-topic similarity for attribution	0.40
τ_{topic}	min topic-to-topic similarity for candidate edge	0.70
$\tau_{\text{cov}}^{\text{out}} = \tau_{\text{cov}}^{\text{in}}$	min coverage for a validated edge	0.30
<i>Event-labeling thresholds</i>		
α_{cont}	continuation similarity threshold	0.90
$\alpha_{\text{cont}}^{\text{out}} = \alpha_{\text{cont}}^{\text{in}}$	min inflow/outflow coverage for continuation	0.50
$\alpha_{\text{split}}^{\text{out}}$	min total outflow coverage for split	0.70
$\alpha_{\text{split}}^{\text{in}}$	min inflow coverage per successor	0.50
$\alpha_{\text{disp}}^{\text{out}} = \alpha_{\text{disp}}^{\text{in}}$	min inflow/outflow coverage for disappearance	0.40
$\alpha_{\text{merge}}^{\text{out}}$	min outflow coverage per predecessor	0.50
$\alpha_{\text{merge}}^{\text{in}}$	min total inflow coverage for merge target	0.70

These parameters were selected based on preliminary experiments and sensitivity analysis.

5.5. Baselines and Comparison Methods

To evaluate the effectiveness of the proposed framework, several baseline methods were considered.

5.5.1 Topic Discovery Baselines

Topic quality within each temporal window was compared with the following methods:

- **Top2Vec [6]** An embedding-based topic modeling approach that jointly learns document and word embeddings.
- **Latent Dirichlet Allocation (LDA) [1]** A classical probabilistic topic modeling technique based on bag-of-words representations.
- **Non-negative Matrix Factorization (NMF) [14]** A matrix factorization approach commonly used for topic extraction.

These baselines allow the comparison of transformer-based topic modeling with both traditional probabilistic and embedding-based methods.

5.5.2 Temporal Modeling Baselines

To contextualize the temporal behavior of topics, the experiments also include comparisons with dynamic topic models:

- **DTM [2]**
- **Dynamic Embedded Topic Model (DETM) [5]**

Because topic identities differ across modeling families, these models are used primarily to analyze stability and temporal behavior rather than tracking accuracy.

5.5.3 Event Detection Baselines

To isolate the contribution of the proposed event detection mechanism, three minimal baselines were implemented:

- **Similarity-only** Topic correspondences are computed using only normalized embedding similarity.
- **Lexical-only** Topic similarity is computed using lexical similarity of c-TF-IDF topic descriptors.
- **Forward-only** Topic transitions are computed using forward document attribution only, without backward flow estimation.

To ensure a fair comparison, all baselines share the same preprocessing pipeline, temporal windowing configuration, and snapshot topics.

5.6. Gold Set and Human Annotation Protocol

To evaluate the semantic validity of detected topic evolution events, a manually annotated gold set was constructed from the Congressional Tweets dataset.

A subset of detected events produced by the proposed system was randomly sampled and balanced across event types (continuations, splits, merges, and disappearances). Each sampled instance corresponds to a candidate transition between topics in adjacent temporal windows.

Three independent annotators evaluated these items following a structured annotation protocol. Instead of directly labeling event types, annotators were presented with candidate topic correspondences across adjacent windows, including topic keywords and representative tweets, and were asked to judge whether the proposed relationships were semantically meaningful.

Event validity was determined using two agreement criteria:

- **Majority agreement**, where at least two annotators confirm the event;
- **Unanimous agreement**, where all three annotators confirm the event.

A detailed description of each baseline is provided in Chapter 7.

5.7. Evaluation Metrics

The evaluation focuses on both topic quality and the validity of detected topic evolution events.

5.7.1 Topic Quality Metrics

Within-window topic quality was evaluated using several commonly used metrics [77–80]:

- **Coherence Value (CV)** measures the semantic similarity between words within a topic using contextual word co-occurrence and distributional information. The metric ranges from 0 to 1, where higher values indicate more semantically coherent topics.
- **Normalized Pointwise Mutual Information (NPMI) Coherence** measures how strongly words within a topic co-occur compared to how often they would be expected to co-occur by chance. It ranges from -1 to 1, where higher values indicate stronger association between topic words.

- **UMass Coherence** is a corpus-based metric that evaluates how frequently words in a topic co-occur within the same documents. It typically yields negative values, with a range from approximately -14 to 0, where values closer to zero indicate better topic interpretability.
- **Topic Diversity** measures the proportion of unique words across all topics, indicating how distinct the topics are from each other. It ranges from 0 to 1, where higher values indicate less redundancy and greater separation between topics.

To summarize overall topic quality, coherence and diversity were combined into a single score, rewarding only models that achieve both coherence and diversity simultaneously:

$$TopicQuality(TQ) = CV \times Diversity \quad (5.1)$$

Although automatic coherence metrics provide useful diagnostics, they are interpreted cautiously due to known limitations in evaluating semantic quality.

5.7.2 Temporal Stability Metrics

To assess how topic quality evolves over time, temporal dynamics were analyzed for three metrics: CV coherence, topic diversity, and the combined topic quality score TQ .

- **Drift**, estimated using linear regression over time:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \quad (5.2)$$

where y_t represents the value of a topic quality metric at time t , β_1 measures the temporal drift, and statistical significance is assessed using a two-sided t -test.

- **Volatility**, which captures short-term fluctuations in topic metrics beyond the global linear drift. For each metric time series y_t (CV coherence, topic diversity, and the combined topic quality score TQ), we compute a rolling mean and rolling Standard Deviation (SD) over a moving window of size $w = 3$, defined as:

$$\bar{y}_t^{(w)} = \frac{1}{w} \sum_{i=t-w+1}^t y_i, \quad (5.3)$$

$$\sigma_t^{(w)} = \sqrt{\frac{1}{w} \sum_{i=t-w+1}^t \left(y_i - \bar{y}_t^{(w)} \right)^2}. \quad (5.4)$$

The rolling mean $\bar{y}_t^{(w)}$ provides a smoothed local estimate of each metric, while the rolling standard deviation $\sigma_t^{(w)}$ indicates the degree of local variability or temporal irregularity. These statistics are summarized using their overall averages across time windows and the maximum observed short-term deviation.

Note that for Tomotopy DTM and DETM a window of 3 corresponds to three consecutive years, whereas for BERTilda it corresponds to three overlapping seven-year windows. Consequently, the rolling statistics for BERTilda span a longer temporal interval and should be interpreted with this difference in mind.

5.7.3 Event Validation Metrics

The validity of detected topic evolution events was evaluated using human annotations. For each predicted event, annotators assessed whether the detected transition corresponded to a meaningful semantic change in the topic structure.

Evaluation is based on **validation rate**, which measures the proportion of predicted events that are confirmed as valid by human annotators:

$$ValidationRate = \frac{TP}{TP + FP} \quad (5.5)$$

where TP denotes the number of true positive events (predicted events validated by annotators) and FP denotes the number of false positive events (predicted events judged invalid).

Because multiple event types may be detected (e.g., topic splits, merges, continuations, and disappearances), overall performance is summarized using **macro-average validation rate**. This metric computes the validation rate independently for each event type and then averages the results:

$$ValidationRate_{macro} = \frac{1}{K} \sum_{k=1}^K ValidationRate_k \quad (5.6)$$

where K represents the number of event types and $ValidationRate_k$ denotes the validation rate computed for event type k .

Macro-averaging ensures that each event type contributes equally to the final evaluation and provides a balanced comparison across different model configurations and parameter settings.

5.8. Case Studies

In addition to quantitative evaluation, qualitative case studies were conducted to illustrate the interpretability of the detected topic evolution events.

These case studies focus on selected political topics of the Congressional Tweets dataset and examine how they evolve across consecutive temporal windows. The objective is to demonstrate how the framework captures meaningful semantic transitions such as topic fragmentation, convergence, or disappearance.

The detailed interpretation of these case studies and the insights obtained from them are presented in the next chapter.

6. Results

We first report within-window topic quality and fragmentation on the congressional tweets corpus, then provide temporal stability diagnostics on historical corpora and illustrate the resulting event narratives via a qualitative case study.

6.1. Static Topic Quality and Fragmentation (Congress)

To ensure a fair comparison, all models were configured with equivalent parameters where applicable. BERTilda and Top2Vec were both configured using the same sentence embedding model (MiniLM) and set to return 10 top words per topic. For Top2Vec, a minimum cluster size of 50 documents per topic was enforced to reduce noise. For LDA and NMF, a global dictionary and vocabulary were constructed from all documents across the full dataset, ensuring that topic representations from different temporal windows are defined within the same feature space and are thus directly comparable over time. Given that both models require a pre-defined number of topics, $k = 10$ was selected after empirical exploration of alternative values.

Across all temporal windows, BERTilda produced 2,569 topics in total, while Top2Vec produced 5,717 topics, frequently exceeding 70 topics per week. This pronounced fragmentation suggests that Top2Vec tends to over-segment the corpus into highly specific micro-clusters, which may hinder interpretability at the window level.

Table 6.1 reports the static topic quality metrics averaged across all windows. BERTilda achieves strong coherence scores ($CV = 0.7381$, $NPMI = 0.1981$) and high topic diversity (0.9751), yielding an overall Topic Quality (TQ) score of 0.7200 . NMF attains the highest scores across all metrics ($TQ = 0.7539$), benefiting from its fixed vocabulary and global optimisation objective. LDA produces moderate coherence ($CV = 0.5716$, $TQ = 0.5541$), while Top2Vec lags considerably behind, with negative NPMI (-0.3005) and UMass (-11.5995) scores and the lowest TQ (0.3184), consistent with its tendency to produce fragmented and semantically diffuse topics.

Table 6.1: Static topic quality metrics on the congressional tweets dataset.

Model	CV	NPMI	UMass	Diversity	TQ
BERTilda (BERTopic + temporal graph)	0.7381	0.1981	-2.8489	0.9751	0.7200
Top2Vec	0.3565	-0.3005	-11.5995	0.8937	0.3184
NMF	0.7697	0.2155	-2.1038	0.9793	0.7539
LDA	0.5716	0.0448	-4.2256	0.9694	0.5541

Overall, NMF achieves the best performance across all evaluation metrics, while BERTilda also performs well, clearly outperforming both LDA and Top2Vec. A key advantage is that, unlike NMF and LDA, BERTilda does not require the number of topics to be specified in advance. It also works directly with contextual embeddings, which makes these results especially promising given the additional complexity introduced by temporal modeling.

6.2. Temporal Stability Diagnostics (Historical Corpora)

To assess whether within-window topic quality remains stable over time, we evaluate BERTilda against two dedicated dynamic topic models (Tomotopy DTM and DETM) on two historical corpora: the UN General Debates (1970–2015) and the State of the Union addresses (1790–2018).

Three complementary diagnostics are reported for each corpus: (i) mean topic quality metrics averaged across all temporal windows (Tables 6.2 and 6.5); (ii) linear temporal drift, estimated via ordinary least squares regression of each metric on time, with associated p -values (Tables 6.3 and 6.6); and (iii) short-term volatility, measured as the average and maximum rolling standard deviation over a window of three consecutive periods (Tables 6.4 and 6.7).

Because topic identity is not directly comparable across model families, these results should be interpreted as qualitative evidence of temporal consistency rather than as a direct accuracy comparison.

6.2.1 UN General Debates (1970–2015)

Table 6.2 reports mean topic quality metrics across all windows. BERTilda achieves the highest overall TQ (0.551 ± 0.095), slightly outperforming DETM (0.541 ± 0.012), while Tomotopy DTM performs considerably worse (0.211 ± 0.019). DETM shows near-perfect diversity (0.998), though NPMI and UMass scores are unavailable. BERTilda provides

a balanced profile, combining strong coherence ($CV = 0.593$, $NPMI = 0.084$) with high diversity (0.921).

Table 6.2: Overall topic quality metrics on UN General Debates (mean $\pm$ std across windows).

Model	CV	NPMI	UMass	Diversity	TQ
BERTilda	0.593 ± 0.066	0.084 ± 0.044	-1.105 ± 0.776	0.921 ± 0.074	0.551 ± 0.095
Tomotopy					
DTM	0.361 ± 0.037	-0.030 ± 0.013	-0.697 ± 0.614	0.589 ± 0.067	0.211 ± 0.019
DETM	0.542 ± 0.012	–	–	0.998 ± 0.000	0.541 ± 0.012

Temporal drift results (Table 6.3) show no significant trends for BERTilda in CV or TQ ($p = 0.564$, $p = 0.106$), indicating stable coherence over time. A small but statistically significant increase in diversity ($\beta_1 = 0.005$, $p = 0.006$) suggests a slight rise in lexical variety, though the effect size is minimal. In contrast, Tomotopy DTM shows significant drift across all metrics, particularly TQ ($p = 2.05 \times 10^{-9}$), while DETM shows no significant drift.

Table 6.3: Temporal drift on UN General Debates (drift = β_1).

Model	Metric	Drift	p-value
BERTilda	CV	0.0007	0.5635
	Diversity	0.0051	0.0062
	TQ	0.0031	0.1062
Tomotopy DTM	CV	0.0009	0.0204
	Diversity	0.0016	0.0310
	TQ	0.0010	2.05e-09
DETM	CV	2.13e-04	0.1215
	Diversity	-1.53e-07	0.9717
	TQ	2.13e-04	0.1219

Short-term volatility analysis (Table 6.4) further supports BERTilda’s stability, with moderate variation in TQ (mean rolling SD = 0.065, max = 0.135). Tomotopy DTM shows lower volatility but much lower overall quality, whereas DETM is the most stable model (max rolling SD = 0.023), reflecting its neural smoothing.

Model	Metric	Window	Rolling mean (avg)	Rolling std (avg)	Rolling std (max)
BERTilda	CV	3	0.598226	0.053813	0.100443
	Diversity	3	0.925166	0.037962	0.093322
	TQ	3	0.558000	0.065464	0.135208
Tomotopy DTM	CV	3	0.359558	0.019797	0.080005
	Diversity	3	0.590263	0.037128	0.104597
	TQ	3	0.210829	0.009855	0.023908
DETM	CV	3	0.542054	0.010595	0.023335
	Diversity	3	0.998442	0.000349	0.000603
	TQ	3	0.541210	0.010607	0.023162

Table 6.4: Short-term volatility metrics (rolling window = 3) for UN General Debates dataset

6.2.2 State of the Union (1790–2018)

A similar pattern of model performance is observed on the State of the Union corpus (Table 6.5). BERTilda achieves the highest TQ (0.580 ± 0.080), followed by DETM (0.548 ± 0.132) and Tomotopy DTM (0.431 ± 0.094).

Table 6.5: State of the Union: overall topic quality metrics (mean $\pm$ std across windows).

Model	CV	NPMI	UMass	Diversity	TQ
BERTilda	0.586 ± 0.081	0.006 ± 0.070	-5.916 ± 2.260	0.989 ± 0.007	0.580 ± 0.080
Tomotopy DTM	0.473 ± 0.079	-0.518 ± 0.086	-16.344 ± 3.926	0.904 ± 0.057	0.431 ± 0.094
DETM	0.550 ± 0.133	–	–	0.998 ± 0.001	0.548 ± 0.132

Temporal drift results (Table 6.6) reveal a small but statistically significant negative drift in BERTilda’s CV and TQ ($\beta_1 = -0.001$, $p < 0.001$ for both), which may reflect the gradual evolution of political language over two centuries rather than model instability. Diversity remains stable ($p = 0.100$). Tomotopy DTM shows a significant decline in diversity ($p < 0.001$) but stable CV and TQ, while DETM exhibits no significant drift across any metric.

Table 6.6: State of the Union: temporal drift on topic-quality metrics (drift = β_1).

Model	Metric	Drift	p-value
BERTilda	CV	-0.001	0.000
	Diversity	0.000	0.100
	TQ	-0.001	0.000
Tomotopy DTM	CV	0.000	0.655
	Diversity	0.000	0.000
	TQ	0.000	0.122
DETM	CV	0.000	0.173
	Diversity	0.000	0.265
	TQ	0.000	0.169

Short-term volatility (Table 6.7) is broadly consistent with the UN results for BERTilda, with an average rolling standard deviation in TQ of 0.039 and a maximum of 0.089. Tomotopy DTM displays considerably higher peak volatility in both diversity (0.187) and TQ (0.151), suggesting less consistent behavior across consecutive windows. DETM’s maximum rolling standard deviation in TQ reaches 0.233 on this corpus, a notable increase relative to the UN dataset, which may be attributable to the sparser document distribution across some historical periods.

Model	Metric	Window	Rolling mean (avg)	Rolling std (avg)	Rolling std (max)
BERTilda	CV	3	0.586355	0.039820	0.091979
	Diversity	3	0.989338	0.003438	0.015535
	TQ	3	0.579797	0.039254	0.089234
Tomotopy DTM	CV	3	0.472304	0.035421	0.115503
	Diversity	3	0.903866	0.026472	0.186796
	TQ	3	0.430403	0.042339	0.150781
DETM	CV	3	0.549288	0.070496	0.233219
	Diversity	3	0.998106	0.001069	0.004468
	TQ	3	0.548190	0.070286	0.232769

Table 6.7: Short-term volatility metrics (rolling window = 3) for the State of the Union dataset

Taken together, these diagnostics indicate that BERTilda maintains competitive and temporally stable topic quality across both corpora.

6.3. Qualitative Case Studies (Congress)

Having established that BERTilda maintains competitive and temporally stable topic quality across both historical corpora, we now turn to the congressional tweets dataset to examine whether the detected lifecycle events correspond to meaningful narrative reorganizations in practice.

This section presents a qualitative analysis of selected topics identified by BERTilda in the Twitter data from members of the 119th United States Congress. The goal is to illustrate, through concrete examples, how the system tracks the evolution of politically relevant discourse over time, capturing phenomena such as topic continuations, splits, merges, and disappearances across consecutive time windows.

The analysis is supported by the dashboard described in Section 6.3.1, which provides interactive visualizations of topic evolution, including the network graph, the topics' information table, and the documents over time chart. For each case study, the network graph is used as the primary visualization, complemented by qualitative inspection of representative tweets and top words to contextualize and interpret the detected transitions.

6.3.1 Dashboard

To facilitate the exploration and interpretation of BERTilda's outputs, a dedicated interactive dashboard was developed. The dashboard supports two primary filtering modes: filtering by date, which displays all topics identified within a user-selected time interval, and filtering by topic, which allows the full temporal evolution of a specific topic to be traced across consecutive windows, including its past, present, and future occurrences. The dashboard provides three complementary visualization components, described below.

6.3.1.1 Network Graph

The main visualization is a network graph in which each node represents a topic identified in a given time window. The position of each node along the x-axis corresponds to the time window in which it was detected, while the y-axis is automatically arranged to minimize overlap, with topics ordered by tweet count: larger topics are positioned lower and smaller topics higher. Node size is proportional to the number of tweets associated with the topic in that window, and node color encodes the detected temporal behavior, which may correspond to a continuation, split, merge, disappearance, or unclear transition (Figure 6.1).

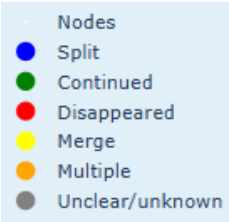


Figure 6.1: Node color legend.

Hovering over a node displays detailed information, including the window number, start and end dates, topic ID, tweet count, and the top 10 representative words (Figure 6.2).

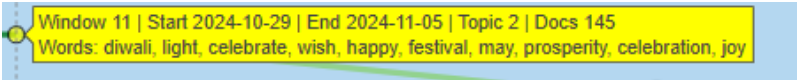


Figure 6.2: Example of node hover information.

Edges between nodes indicate detected relationships across consecutive windows. Edge thickness represents forward coverage, while edge color encodes cosine similarity between topic representations. Dashed edges denote unclear or ambiguous relations. Hovering over an edge reveals the specific values for coverage, similarity, and relationship type (Figure 6.3).

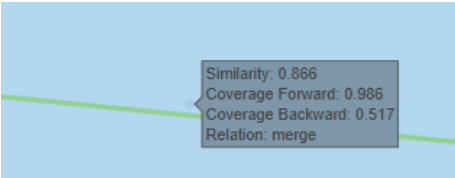


Figure 6.3: Example of edge hover information.

6.3.1.2 Topics’ Information Table

The topics’ information table provides a tabular view of all topics included in the current network graph. For each topic, the table displays the window number, topic ID, top 10 representative words, tweet count, and optionally five representative tweets. While much of this information is also accessible via node hover, the tabular format enables direct comparison across topics and provides qualitative insight into topic content through the display of sample documents (Figure 6.4).

Window	Topic	Top 10 Words	Document Count	Representative Tweets
1	12	passport, lebanon, travel, citizen, state, statedept, renew, flight, embassy, online	64	Tweet 1: "For additional information, see @usembassybeirut's full security alert at: https://t.co/a8qgDTQXpV ", Tweet 2: "The Department of State should reconsider Nashville as a location for its new passport agency. Nashville is a prime location to house a new passport facility that will benefit Tennesseans and neighboring states alike. https://t.co/vVypGEkqmj ", Tweet 3: "The Department of State's Online Passport Renewal system allows Americans to renew their passports through a secure and paperless process. Check to make sure your passports are up-to-date and visit the site below for more info: https://t.co/aly0SBVQph ", Tweet 4: "My #CA34 Passport Fair is only a few weeks away! If you need help renewing your passport or applying for the first time, this is your chance to speak to an agent on-site and receive direct guidance. 🗳️ RSVP today at https://t.co/nbz196dVZM https://t.co/zvTmhzk4w9 ", Tweet 5: "You can also follow @TravelGov for ongoing updates from the State Department. https://t.co/VSX4VZmBZY "

Figure 6.4: Example of topic information table: window 1, topic 12. Because the topic did not persist beyond its initial window, no predecessor or successor topics are shown.

6.3.1.3 Documents Over Time

The Documents Over Time component displays the total number of tweets per time window for the topics included in the network graph. When filtering by date, the chart reflects overall activity levels across the selected interval. When filtering by topic, it shows the temporal evolution of a specific topic, highlighting peaks, declines, and general dynamics of its presence across time windows (Figure 6.5).

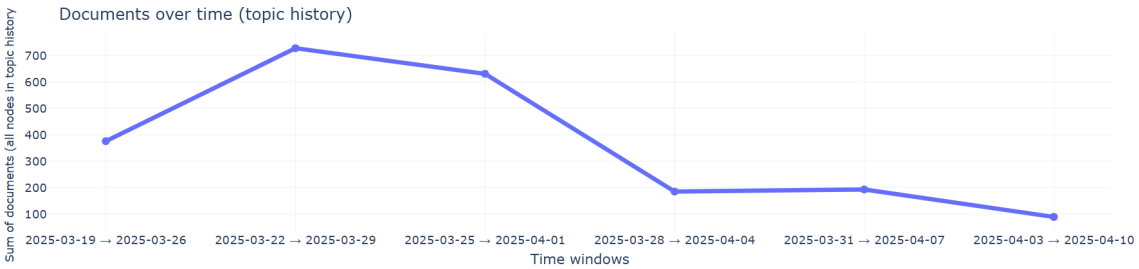


Figure 6.5: Documents over time: Example of topic 18 from window 63

6.3.2 Holocaust/Israel–Antisemitism.

From mid-April to mid-May, BERTilda tracks a topic that shifts from Holocaust remembrance to contemporary Israel/Hamas and antisemitism discourse. The system marks this transition as an unclear continuation, followed by a split that separates geopolitical conflict from campus-focused antisemitism debates. The resulting branches then continue independently until disappearance, as illustrated in Figure 6.6.

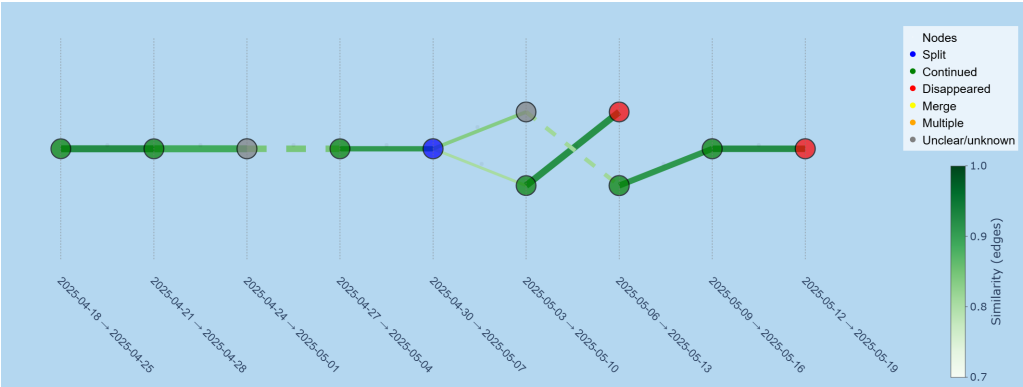


Figure 6.6: Example temporal evolution of a topic (Holocaust/Israel–Antisemitism), illustrating split, continuations, and disappearance across consecutive windows.

Window	Topic	Top 10 Words	Document Count
68	11	holocaust, hashoah, yom, million, antisemitism, never, jews, six, jewish, murder	142
69	2	holocaust, yom, hashoah, million, antisemitism, never, remember, murder, six, jewish	153
70	1	holocaust, antisemitism, jewish, yom, hashoah, million, never, remember, must, murder	207
71	8	israel, jewish, antisemitism, hostage, hamas, must, iran, independence, terrorist, hold	173
72	4	jewish, israel, antisemitism, iran, hamas, hostage, terrorist, ally, must, gaza	217
73	7	jewish, antisemitism, campus, college, student, university, columbia, edworkforcecmte, antisemitic, month	144
73	15	iran, israel, gaza, hamas, nuclear, hostage, benjamin, terrorist, aid, sanction	116
74	4	edan, hostage, alexander, hamas, home, release, israel, captivity, iran, finally	273
74	14	antisemitism, jewish, campus, college, university, student, columbia, contribution, month, antisemitic	139
75	2	edan, hostage, alexander, hamas, home, release, captivity, finally, hold, gaza	246
76	5	edan, hostage, alexander, hamas, home, release, captivity, finally, gaza, hold	231

Figure 6.7: Top words evolution of a topic (Holocaust/Israel–Antisemitism).

6.3.3 Town Halls and Constituent Services Assistance.

Two recurring themes run through this topic’s timeline, briefly converging before returning to their distinct trajectories. The first centers on community engagement and event outreach, characterized by terms such as *town hall*, *join*, *live*, *RSVP*, and *telephone*, reflecting constituents being invited to participate in public discussions. The second focuses on federal assistance and casework, with keywords like *passport*, *agency*, *federal*, *help*, and *casework*. Early in the timeline, these themes temporarily merge, likely reflecting communications that simultaneously promoted live sessions and highlighted available services. Once this overlap subsides, the two topics resume their separate paths, until each eventually disappears.

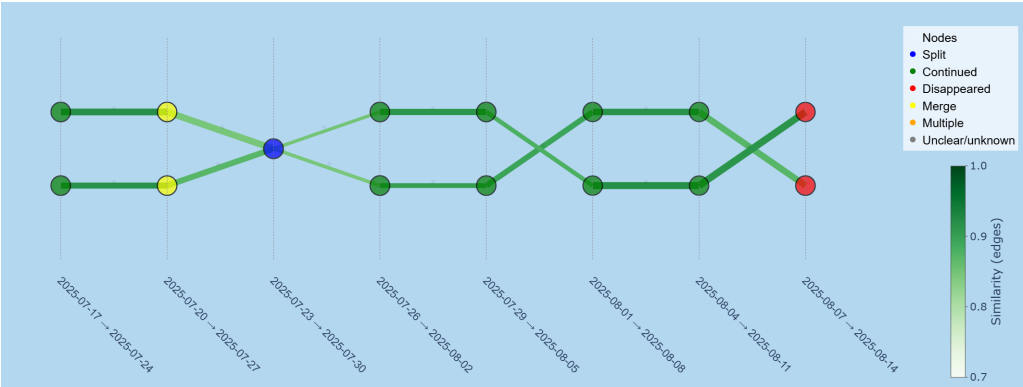


Figure 6.8: Example temporal evolution of a topic (Town Halls and Constituent Services Assistance), illustrating merge, split, continuations, and disappearance across consecutive windows.

Window	Topic	Top 10 Words	Document Count
98	17	hall, town, pm, telephone, july, join, question, live, rsvp, hear	92
98	18	office, passport, help, agency, hour, federal, team, mobile, july, social	86
99	12	hall, town, telephone, pm, join, newsletter, july, hear, question, tonight	104
99	18	office, hour, passport, help, agency, federal, team, mobile, pm, social	74
100	10	hall, town, office, pm, help, team, passport, agency, july, hour	133
101	12	town, hall, pm, join, newsletter, update, question, host, subscribe, hear	137
101	15	help, office, team, va, federal, agency, passport, need, veteran, casework	83
102	19	hall, town, join, pm, update, newsletter, host, rsvp, subscribe, telephone	81
102	21	office, va, help, veteran, federal, agency, need, team, service, mobile	73
103	23	office, help, agency, federal, team, passport, need, hour, mobile, assistance	69
103	25	hall, town, rsvp, august, pm, join, host, facebook, telephone, form	64
104	24	office, help, passport, agency, federal, team, need, hour, constituent, assistance	60
104	25	hall, town, rsvp, join, host, august, facebook, come, docs, google	60
105	14	hall, town, join, pm, update, newsletter, host, rsvp, subscribe, telephone	110

Figure 6.9: Part of top words evolution of a topic (Town Halls and Constituent Services Assistance).

6.3.4 Disaster Management and Agricultural Impact.

Window 79 surfaces two distinct topics: one centered on emergency and disaster management, with terms such as *emergency*, *disaster*, *federal*, *hurricane*, and *storm*, and another focused on agriculture, including *farmer*, *farm*, *agriculture*, *industry*, and *food*. In the following window, these converge into a broader topic encompassing *business*, *community*, and *federal response*, a transition that reflects the real-world entanglement of climate-related disasters and agricultural disruption. When storms and hurricanes strike, farms, rural economies, and local communities bear much of the impact, naturally drawing disaster response and agricultural concerns into a shared discourse. The merged topic subsequently disappears.

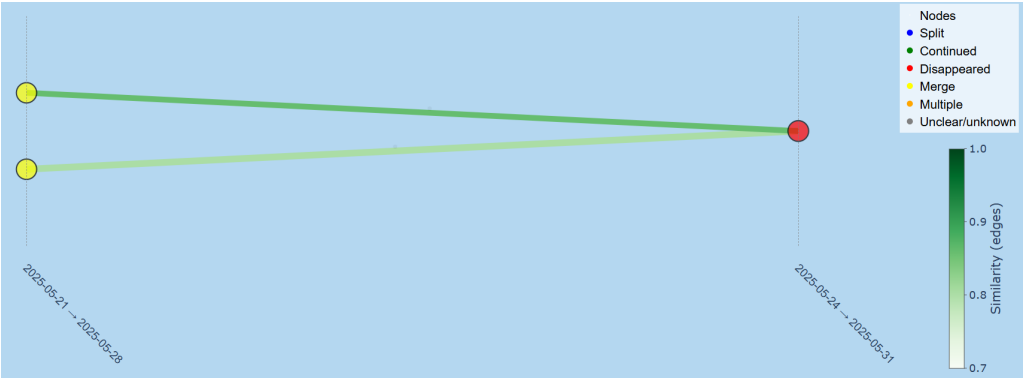


Figure 6.10: Example temporal evolution of a topic (Disaster Management and Agricultural Impact), illustrating merge and disappearance across consecutive windows.

Window	Topic	Top 10 Words	Document Count
79	10	emergency, disaster, management, federal, agency, storm, declaration, hurricane, request, weather	126
79	11	farmer, farm, agriculture, industry, great, discuss, support, food, economy, ag	114
80	4	business, federal, community, emergency, project, disaster, work, county, agency, small	263

Figure 6.11: Top words evolution of a topic (Disaster Management and Agricultural Impact).

6.3.5 Shooting Event to Political Polarization.

A violent school shooting initially dominates the discourse. The discussion is anchored in factual and location-based language, including *school*, *gun*, *shooting*, *student*, and *Evergreen*, which situate the conversation around the concrete details of the event. In the subsequent window the topic begins to diverge into two distinct strands. One strand continues to focus on the event itself but gradually loses prominence. The other slowly shifts in tone as the discussion moves away from the immediate details of the incident and toward broader ideological framing. Across the following windows the discourse becomes increasingly shaped by ideological vocabulary, with terms such as *political*, *radical*, *democracy*, *fascist*, and *antifa* appearing more frequently. Over time, a discussion that initially focuses on a specific act of violence gradually evolves into a broader ideological debate, while the original event becomes less central to the conversation.

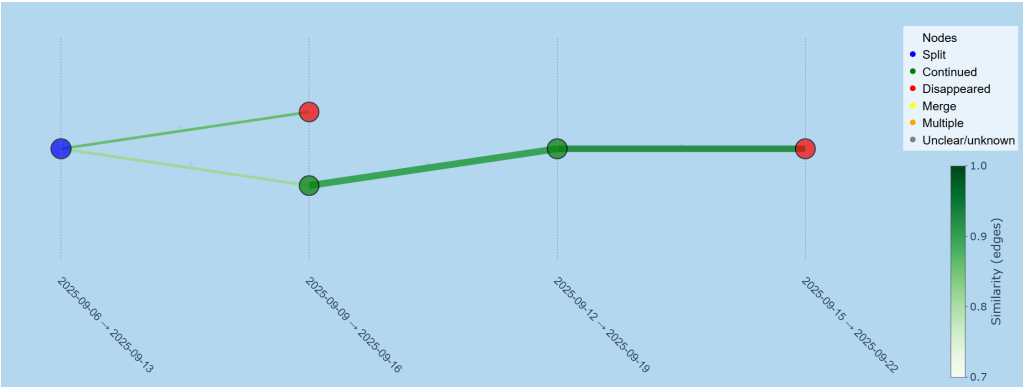


Figure 6.12: Example temporal evolution of a topic (Shooting Event to Political Polarization), illustrating split, continuations, and disappearance across consecutive windows.

Window	Topic	Top 10 Words	Document Count
115	4	violence, gun, school, political, student, evergreen, threat, shooting, firearm, high	265
116	12	violence, political, condemn, acceptable, democracy, never, country, must, place, freedom	111
116	18	school, gun, evergreen, shooting, high, violence, colorado, shoot, denver, community	76
117	10	violence, political, radical, must, condemn, leave, rhetoric, fascist, left, democracy	157
118	11	violence, political, radical, rhetoric, leave, antifa, people, must, left, fascist	163

Figure 6.13: Top words evolution of a topic (Shooting Event to Political Polarization).

7. Human Validation

Automatic evaluation metrics alone are insufficient to assess whether detected topic events are meaningful in practice. This chapter reports a human validation study designed to estimate the validation rate of BERTilda’s event detection against three baselines, using independent annotators to confirm whether predicted events correspond to genuine topic transitions.

7.1. Evaluation Setup

7.1.1 Annotation Protocol

Three independent annotators evaluated a curated subset of detected events from the congressional corpus, balanced across event types. Annotators were three master’s students from domains unrelated to either political or data science (mechanical engineering, sports training, and public administration), ensuring that detected events are interpretable without specialized domain or methodological knowledge.

Rather than asking annotators to label events directly (split/merge/continue/disappear), we present them with candidate topic correspondences across adjacent windows, reducing the task to a binary judgment of whether the topics are related, which does not require familiarity with the underlying methodology.

For continuation, split, and disappearance items, the focal topic is in window t , and annotators see its keywords and representative tweets, together with the three most similar candidate successors in $t+1$. For merge items, the focal topic is in $t+1$ and annotators see the three most similar candidate predecessors in t . Annotators select which candidates are genuinely related, from which we infer event validity.

7.1.2 Annotated Evaluation Sample

The evaluation sample was constructed by randomly selecting events from BERTilda’s detected transitions, stratified by event type (30 instances per type). This design was chosen to ensure that annotators evaluated a diverse and balanced set of continuation, split, merge, and disappearance cases, including relatively rare event categories.

The stratified sampling was performed over BERTilda’s predicted event labels, not over manually verified ground-truth events. Consequently, the balanced composition of the

annotated sample does not imply that all sampled instances represent valid event transitions.

The annotations are not meant to measure how common each event type is in the entire corpus, nor the overall accuracy of each baseline method across every possible transition. Instead, the annotated sample provides a shared set of topic candidates that allows all methods to be compared fairly under the same human judgments.

Annotators evaluated these items, and we report the annotator-confirmed validation rate on detected events, under both majority and unanimous agreement across three annotators.

7.2. Event Detection Baselines

7.2.1 Baseline Definitions

To isolate the contribution of bidirectional coverage and document-level flow, we implement three minimal baselines that operate on the same snapshot topics and windows as BERTilda.

- (i) *Similarity-only*: build the temporal graph using only topic similarity and derive pseudo-flow by normalizing similarities per source and target.
- (ii) *Lexical-only*: replace embedding similarity by top-word overlap and normalize analogously.
- (iii) *Forward-only*: compute topic outflow using only forward attribution, approximating inflow without backward attribution.

7.2.2 Protocol and Comparability Safeguards

All event-detection baselines reuse the same preprocessing, windowing, snapshot topics, and (where applicable) topic representations $\psi(t, k)$. This avoids conflating event detection quality with differences in topic discovery. For each adjacent window pair, we compute the topic sets $\mathcal{K}_t, \mathcal{K}_{t+1}$ once and store $n_{t,k}$, c-TF-IDF descriptors, and $\psi(t, k)$.

Similarity-only computes a similarity matrix $S_{k,j} = \text{sim}((t, k), (t+1, j))$ and retains entries above τ_{topic} . To make rules comparable to coverage-based thresholds, we derive normalized weights

$$\widetilde{\text{cov}}_{\text{out}}(k \rightarrow j) = \frac{S_{k,j}}{\sum_{j'} S_{k,j'}}, \quad \widetilde{\text{cov}}_{\text{in}}(k \rightarrow j) = \frac{S_{k,j}}{\sum_{k'} S_{k',j}},$$

and apply the same event rules using $(\widetilde{\text{cov}}_{\text{out}}, \widetilde{\text{cov}}_{\text{in}})$. Only the entries of $S_{k,j}$ above the threshold τ_{topic} are considered when computing the normalized weights.

Lexical-only repeats the procedure using a lexical similarity matrix (Jaccard overlap of top- N words).

Forward-only runs only the forward attribution step, computes $F_t^{\text{fwd}}(k \rightarrow j)$, and sets

$$\text{cov}_{\text{out}}^{\text{fwd}}(k \rightarrow j) = \frac{F_t^{\text{fwd}}(k \rightarrow j)}{n_{t,k}}, \quad \text{cov}_{\text{in}}^{\text{fwd}}(k \rightarrow j) = \frac{F_t^{\text{fwd}}(k \rightarrow j)}{n_{t+1,j}},$$

thereby approximating inflow without backward attribution.

Finally, all methods share the same default decision thresholds for alignment and event labeling (Table 5.4) and apply the same post-processing (minimum topic size and candidate-edge pruning).

7.3. Validation Results

We report results using two agreement criteria (majority and unanimous vote) to reflect both the general level of agreement and a stricter measure of how clearly the events can be interpreted.

7.3.1 Distribution of Predicted Events

Table 7.1 shows the distribution of predicted event labels across the 120 annotated gold-set items. The table reports only four event categories (continue, disappear, split, and merge). Any remaining items correspond to cases where, for the selected topic, a method classified the relationship with its candidate successor(s) or predecessor(s) as unclear. BERTilda is the only method that produces a perfectly balanced distribution, a direct consequence of the stratified sampling procedure. The baselines deviate systematically: similarity-only and forward-only both over-predict disappearances, the latter due to its inability to account for inflow through backward attribution. Lexical-only collapses almost entirely into disappearances, with no splits or merges detected, confirming that top-word overlap alone is insufficient to capture topical continuity across windows.

Table 7.1: Distribution of predicted event labels on the 120 annotated gold-set items.

Method	Continue	Disappear	Split	Merge
BERTilda (similarity + bidirectional coverage)	30	30	30	30
Similarity-only (normalized topic similarity)	20	48	21	20
Lexical-only (normalized c-TF-IDF similarity)	4	106	0	0
Forward-only (no backward attribution)	35	43	23	14

These distributional imbalances help contextualize the validation rates reported below.

7.3.2 Human Validation Rates

Majority Vote Criterion Across all four event types, BERTilda yields the highest macro-average validation rate among the compared methods (0.775), ahead of similarity-only (0.571) and forward-only (0.725). For Lexical-only, the reported macro-average (0.488) is computed only over continuation and disappearance, since the method never predicts split or merge events.

Although forward-only attains higher validation rates for split and merge events, BERTilda performs substantially better on continuation and, especially, disappearance events. This advantage is consistent with Table 7.1, where forward-only predicts considerably more disappearance events, reflecting the limitations of relying solely on forward attribution.

Table 7.2: Annotator-confirmed validation rate on the common annotated event sample, using majority vote across three annotators. Macro-average is computed over the event types predicted by each method.

Method	Continue	Disappear	Split	Merge	Macro-avg
BERTilda (similarity + bidirectional coverage)	0.867	0.800	0.767	0.667	0.775
Similarity-only (normalized topic similarity)	0.750	0.417	0.667	0.450	0.571
Lexical-only (normalized c-TF-IDF similarity)	0.750	0.226	NA	NA	0.488
Forward-only (no backward attribution)	0.800	0.558	0.826	0.714	0.725

Unanimous Vote Criterion Under the stricter unanimous-vote criterion, inter-annotator agreement remained substantial (Fleiss' $\kappa = 0.611$, Krippendorff's $\alpha = 0.612$), but absolute validation rates decreased, especially for the more ambiguous split and merge cases. This pattern is expected: unanimous vote rewards only the least ambiguous instances and therefore provides a stricter lower bound on event interpretability.

Even under unanimous vote, forward-only slightly outperforms BERTilda on split and merge events, whereas BERTilda remains strongest on disappearance validation. This pattern mirrors the majority-vote results and suggests that bidirectional coverage primarily improves the identification of topic continuations and disappearances, while forward attribution alone can still capture some complex split and merge relationships.

Table 7.3: Annotator-confirmed validation rate on the common annotated event sample, using unanimous vote across three annotators. Macro-average is computed over the event types predicted by each method.

Method	Continue	Disappear	Split	Merge	Macro-avg
BERTilda (similarity + bidirectional coverage)	0.533	0.567	0.367	0.400	0.467
Similarity-only (normalized topic similarity)	0.550	0.312	0.381	0.300	0.386
Lexical-only (normalized c-TF-IDF similarity)	0.750	0.160	NA	NA	0.455
Forward-only (no backward attribution)	0.457	0.395	0.391	0.429	0.418

Taken together, the majority and unanimous vote results indicate that BERTilda provides the most balanced event detection across all event types. Although forward-only achieves slightly higher validation rates for split and merge events, BERTilda consistently performs better on continuation and disappearance events while also attaining the highest overall macro-average validation rate. Therefore, these results suggest that combining semantic similarity with bidirectional coverage provides a more reliable basis for event detection than the simpler alignment variants considered.

8. Robustness to threshold perturbations

A practical concern for any threshold-based system is whether its performance depends critically on a specific, hand-tuned configuration. That is, whether small deviations from the default thresholds cause meaningful degradation. Robustness, in this context, means that the method’s validated rate and predicted event distribution remain stable across a range of plausible threshold values, rather than being an artifact of one particular setting.

To assess this, we performed a local one-at-a-time sensitivity analysis around the default configuration on the congressional gold set. Preprocessing, windowing, snapshot topics, topic representations, and the 120 annotated items with majority-vote evaluation protocol were kept fixed and only one threshold was varied at a time. We focused on the thresholds most directly tied to alignment and labeling, namely τ_{topic} , τ_{doc} , $\alpha_{\text{split}}^{\text{out}}$, $\alpha_{\text{merge}}^{\text{in}}$, and $\alpha_{\text{disp}}^{\text{out}}$, and compared BERTilda against forward-only, the strongest ablation.

Table 8.1: Macro-average validated validation rate under local threshold perturbations. Each entry reports min/max/mean/std across the tested values.

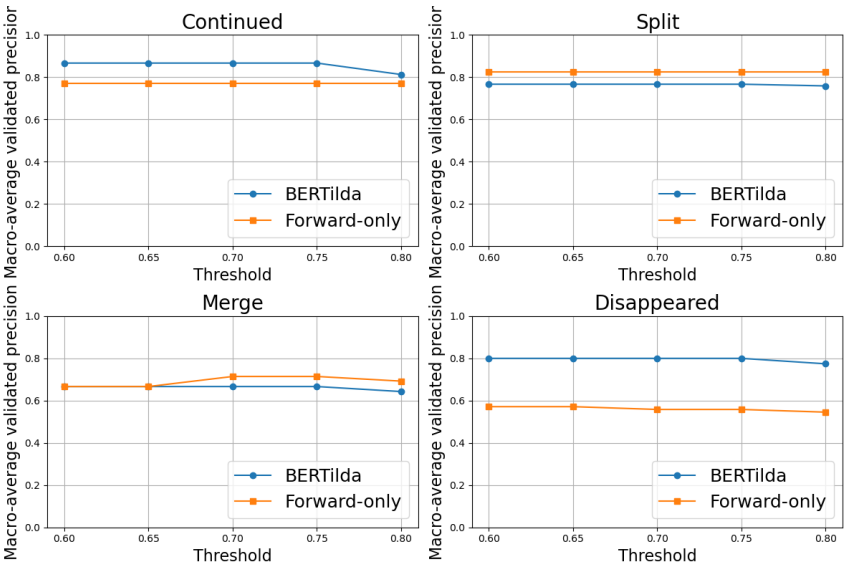
Threshold	BERTilda	Forward-only
τ_{topic}	0.747 / 0.775 / 0.769 / 0.013	0.709 / 0.717 / 0.712 / 0.005
τ_{doc}	0.722 / 0.782 / 0.760 / 0.028	0.686 / 0.721 / 0.707 / 0.002
$\alpha_{\text{disp}}^{\text{out}}$	0.775 / 0.782 / 0.778 / 0.004	0.714 / 0.724 / 0.719 / 0.005
$\alpha_{\text{split}}^{\text{out}}$	0.771 / 0.776 / 0.774 / 0.002	0.713 / 0.721 / 0.717 / 0.003
$\alpha_{\text{merge}}^{\text{in}}$	0.750 / 0.775 / 0.768 / 0.011	0.689 / 0.717 / 0.609 / 0.013

Across the five perturbation analyses, BERTilda’s macro-average validated rate remained between 0.722 and 0.782, whereas forward-only ranged from 0.686 to 0.724. The advantage was most stable for disappearance detection (Fig 8.1). Split and merge were more sensitive, but the method did not collapse into degenerate label distributions under small threshold changes. This supports the claim that BERTilda’s gains are not confined to one hand-tuned threshold configuration.

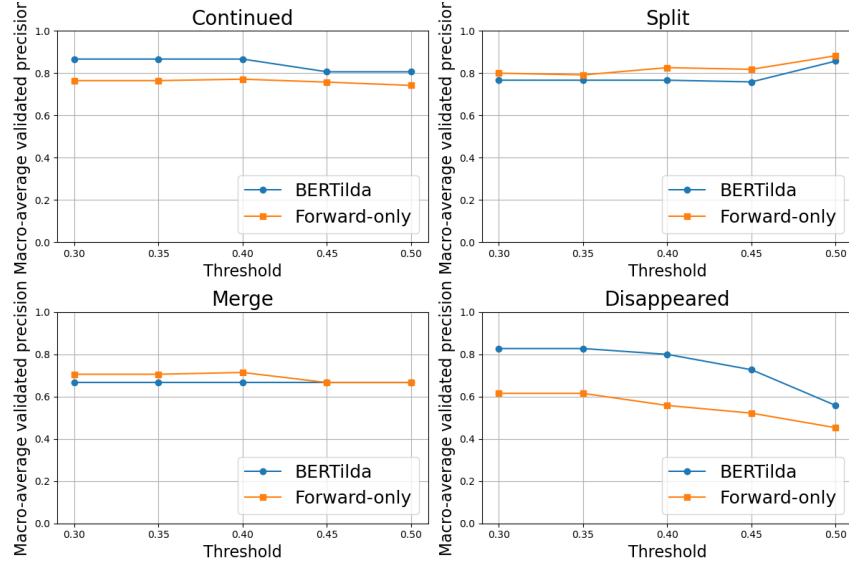
Table 8.2: Range of predicted event counts across all tested threshold perturbations.

Method	Continue	Disappear	Split	Merge
BERTilda	30–32	29–43	21–30	18–30
Forward-only	31–35	39–53	17–25	9–17

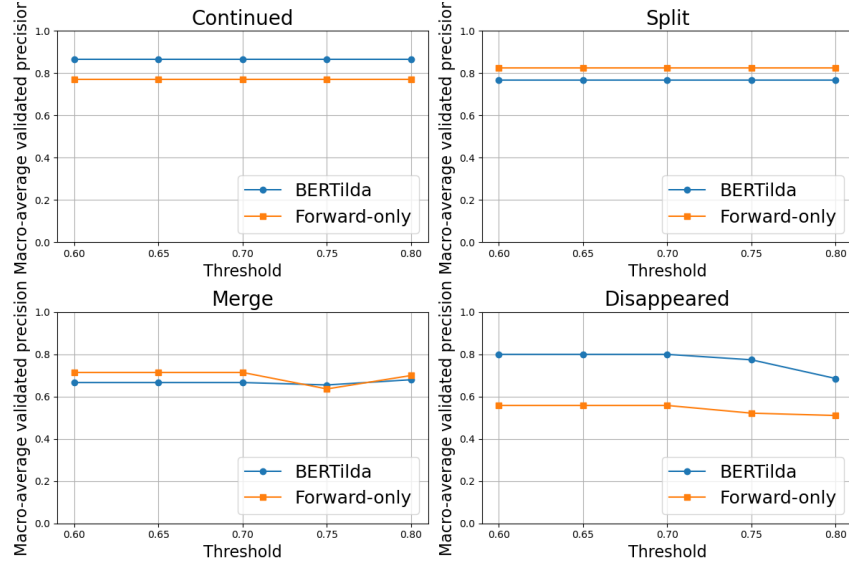
The event-count ranges help explain why macro validation rate remains informative in this setting. Forward-only tends to produce more disappearances and fewer merges, whereas BERTilda preserves a broader event distribution over the tested thresholds. These results further confirm that bidirectional attribution helps stabilize inflow estimates and reduces collapse into a narrower set of labels.



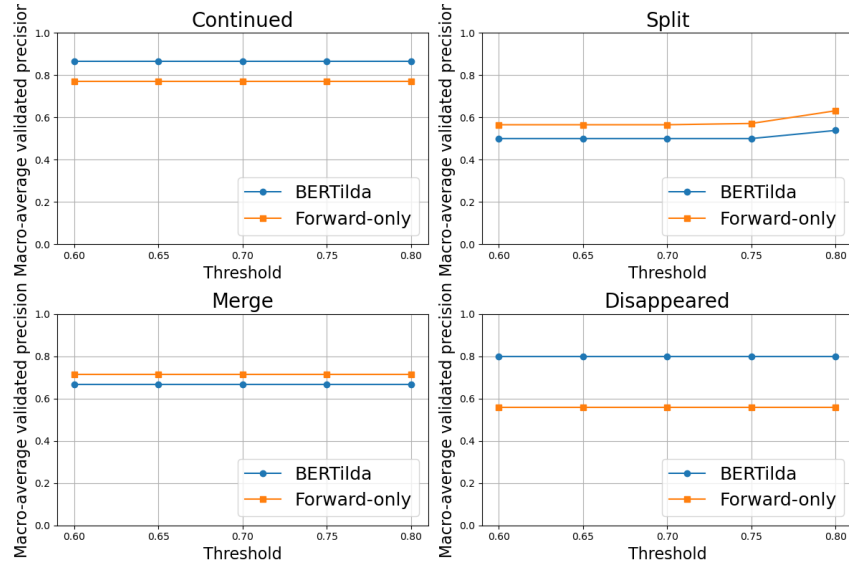
(a) τ_{topic}



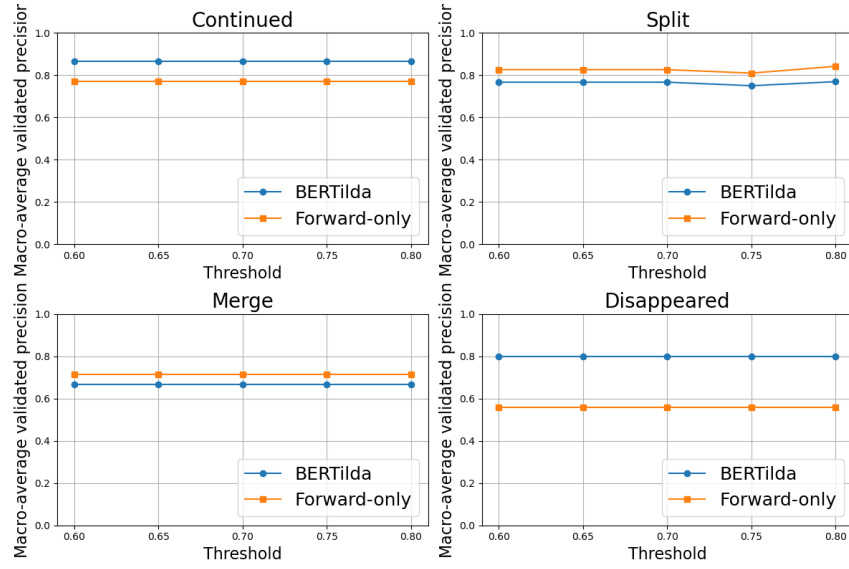
(b) τ_{doc}



(c) $\alpha_{\text{merge}}^{\text{in}}$



(d) a_{split}^{out}



(e) a_{disp}^{out}

Figure 8.1: Local threshold-sensitivity analyses on the congressional gold set.

9. Discussion

The results obtained in this work provide consistent evidence that BERTilda constitutes a viable and interpretable framework for modeling topic evolution in dynamic political corpora. By explicitly separating topic discovery from temporal alignment and by introducing a graph-based representation of topic transitions, the approach enables the identification of discrete lifecycle events that are not captured by traditional dynamic topic models.

A first important finding concerns the quality of the snapshot topics and their behavior over time. The results show that it is possible to maintain strong topic quality and temporal stability without imposing explicit temporal priors. Although NMF achieves higher coherence and diversity scores on individual snapshots, its objective is static topic quality rather than temporal analysis. BERTilda instead trades a small reduction in snapshot quality for the ability to model topic evolution and detect lifecycle events, while maintaining competitive coherence and diversity and avoiding excessive temporal volatility. This supports the argument that separating topic discovery from temporal modeling preserves flexibility while enabling reliable longitudinal analysis.

Beyond topic quality, the central contribution of this work lies in the detection and interpretation of topic evolution events. The proposed framework produces lifecycle representations that are not only structurally coherent, but also largely interpretable under human validation. In particular, continuation and disappearance events achieve the highest validation rates, suggesting that the combination of semantic similarity and bidirectional document coverage provides a robust signal for identifying stable or vanishing topics. These results indicate that the method successfully captures clear and semantically grounded transitions in political discourse.

In contrast, split and merge events remain inherently more ambiguous. Splits often reflect a gradual specialization of a topic into more specific subthemes, rather than a clear and abrupt fragmentation into distinct and unrelated topics, while merges may capture partial thematic overlap rather than the formation of a single unified narrative. These findings suggest that certain lifecycle events should be understood as points along a continuum rather than as sharply defined categories.

A further insight emerges from the comparison with alternative event detection strategies. Methods based solely on similarity signals, whether embedding-based or lexical, tend

to produce less reliable and less interpretable alignments. In particular, lexical similarity collapses most transitions into disappearances, while embedding-based similarity alone is insufficient to capture the structure of topic transitions.

The comparison with the forward-only baseline reveals a more nuanced trade-off. Although forward-only achieves higher validation rates for split and merge events, the absence of backward attribution leads to an overestimation of disappearance events. As a result, BERTilda achieves higher validation rates for continuation and disappearance events and the highest overall macro-average validation rate. These findings suggest that combining semantic similarity with bidirectional coverage provides a more reliable event detection framework overall, whereas forward-only offers stronger performance for specific transition types, particularly splits and merges.

The robustness analysis further indicates that the framework behaves consistently under moderate variations in threshold parameters. The overall distribution of lifecycle events and the main qualitative conclusions remain stable across a reasonable range of values, suggesting that the method does not depend on highly specific parameter configurations. At the same time, certain event types, particularly splits and merges, show greater sensitivity to threshold variation. Small changes in similarity or coverage thresholds can affect the classification of borderline cases, reinforcing the idea that these transitions are inherently gradual and context-dependent. This sensitivity does not undermine the validity of the approach, but points to the importance of careful parameter selection and motivates future work on adaptive or data-driven thresholding strategies.

Despite these strengths, several limitations must be acknowledged. Threshold calibration remains a critical component of the framework, as the choice of similarity and coverage thresholds directly influences the density of the temporal graph and the distribution of lifecycle events. This issue is especially relevant for split and merge detection, where small variations can lead to qualitatively different interpretations. Developing principled methods for threshold selection remains an important direction for future research.

Additionally, the use of overlapping temporal windows introduces a structural bias toward continuity. While this design improves stability and reduces fragmentation, it may also inflate continuation rates and obscure sharper transitions. A systematic evaluation of this effect, for example through ablation studies with non-overlapping windows or alternative temporal formulations, would provide a clearer understanding of its impact.

Comparisons with dynamic topic models such as DTM and DETM must also be interpreted with caution, as topic identity is defined differently across modeling paradigms. Metrics such as drift and volatility should therefore be understood as behavioral diagnostics rather than direct measures of tracking accuracy. Recognizing this distinction is important to avoid misleading conclusions when comparing fundamentally different approaches.

Another methodological alternative would be to analyze the corpus as a whole rather than segmenting it into sliding temporal windows. A global analysis would eliminate the need for inter-window topic alignment and threshold-based event detection, while potentially producing more stable topic representations. However, this simplification would come at the cost of losing temporal resolution, making it impossible to identify when topics emerge, disappear, split, or merge over time. Sliding windows therefore introduce additional methodological complexity, but they are essential for capturing the dynamic evolution of political discourse and the lifecycle events that motivate this work.

The evaluation of lifecycle events focuses primarily on validation rate with respect to annotator-confirmed cases. While this provides a reliable assessment of correctness, it does not capture recall. Estimating recall would require the construction of a more exhaustive gold standard, including systematically annotated negative cases, which was beyond the scope of this work.

Finally, with only 120 annotated events spread across four categories, the reported validation rates are subject to sampling uncertainty that was not formally quantified. An alternative approach would have been to pool the 120 events in order to estimate the sampling variability of the validation rates and more rigorously test whether the differences observed between BERTilda and the baselines are statistically significant, rather than purely descriptive. This analysis is left for future work, ideally using a larger and more comprehensive gold standard.

Overall, the findings demonstrate that BERTilda provides a flexible and interpretable framework for modeling topic evolution, while also highlighting key challenges that remain open for future research.

10. Conclusions

10.1. Summary of the Work

This work addressed the problem of modeling and interpreting topic evolution in temporal text corpora, with a particular focus on detecting structural lifecycle events such as splits and merges. As discussed throughout Chapters 2 and 3, existing approaches to dynamic topic modeling primarily capture smooth temporal drift, often failing to explicitly represent discrete structural events in topics over time.

To address this limitation, we proposed BERTilda, a graph-based framework for topic lifecycle tracking that operates on independently discovered snapshot topics. The central idea is to decouple topic discovery from temporal alignment and to model topic evolution as an explicit matching and event detection problem across consecutive time windows.

The core methodological contribution of this work lies in an explainable alignment mechanism that combines semantic similarity with bidirectional document flow. This design enables the construction of a temporal topic graph, where nodes represent topics and edges encode validated relationships across time. On top of this structure, rule-based criteria allow the labeling of lifecycle events, including continuation, split, merge, disappearance, and unclear transitions. This formulation provides both a formal framework for event detection and an interpretable representation that supports qualitative analysis.

10.2. Research Questions

The empirical evaluation demonstrates that the proposed approach is effective across multiple dimensions. With respect to RQ1, the use of embedding-based snapshot topic models yields coherent and diverse topics, even in short and noisy political texts. Regarding RQ2, the results show that stable and meaningful temporal behavior can be achieved without relying on temporal priors, supporting the viability of the snapshot-based paradigm.

For RQ3, the human evaluation confirms that many of the detected lifecycle events, particularly continuations and disappearances, are semantically valid and interpretable. In relation to RQ4, comparisons with alternative alignment strategies indicate that combining semantic similarity with document flow leads to more balanced event detection across the full lifecycle, at the cost of performance on specific event types. With respect to RQ5, the

robustness analysis shows that the framework remains stable under moderate variations in threshold parameters, although more complex events such as splits and merges exhibit higher sensitivity. Finally, RQ6 is supported by qualitative case studies, which illustrate that the detected events correspond to meaningful reorganizations of political narratives over time.

Beyond quantitative performance, an important contribution of this work lies in its emphasis on interpretability. The temporal topic graph provides an analyst-facing representation that makes the reasoning behind event detection explicit, enabling inspection, validation, and error analysis. This is particularly valuable in domains such as political discourse analysis, where understanding how narratives evolve is as important as detecting that they change.

10.3. Limitations and Future Work

At the same time, several limitations remain. These are not incidental but follow from the same design choice that gives the framework its interpretability: the explicit, rule-based nature of the alignment is what makes event detection inspectable, and also what makes it dependent on threshold calibration and sensitive to the temporal segmentation.

Building on these findings, several directions for future work emerge. First, more systematic robustness analyses are needed, including principled threshold calibration and deeper exploration of temporal segmentation strategies. Second, extending the document-to-topic attribution mechanism to better support multi-membership could improve the handling of ambiguous topics. Finally, future research should expand the evaluation to additional domains and explore alternative alignment formulations, such as transport-based or distributional approaches, to further enhance the detection of complex structural events.

In conclusion, this thesis demonstrates that modeling topic evolution as an alignment and graph construction problem provides a flexible, interpretable, and effective alternative to traditional dynamic topic models. Rather than optimizing solely for snapshot topic quality or individual event types, the proposed framework trades modest reductions in static topic coherence and in the detection of specific split and merge cases for a transparent representation of topic evolution that supports the analysis of the full topic lifecycle. These results show that separating topic discovery from temporal alignment enables the detection and analysis of topic lifecycle events while maintaining reliable topic quality.

More fundamentally, the discrete event taxonomy imposes categorical boundaries on what the analysis suggests is a continuous phenomenon: splits and merges appear to lie on a spectrum of thematic reorganization rather than forming sharply separable classes. Future formulations might therefore represent lifecycle events as graded or probabilistic relations rather than hard labels.

This perspective is not limited to political discourse. Because the framework treats temporal evolution as an alignment problem over successive topic snapshots, it can be applied to any temporally segmented text collection in which understanding how themes evolve is as important as identifying the themes themselves.

References

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, Mar. 2003. [Cited on pages 2, 3, 7, 23, and 44.]
- [2] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 113–120. [Cited on pages 3, 12, 21, 23, 27, and 44.]
- [3] X. Wang and A. McCallum, “Topics over time: a non-markov continuous-time model of topical trends,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 424–433. [Cited on pages 3, 12, 21, 23, and 27.]
- [4] C. Wang, D. Blei, and D. Heckerman, “Continuous time dynamic topic models,” 2015. [Cited on pages 3 and 13.]
- [5] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “The dynamic embedded topic model,” 2019. [Cited on pages 3, 13, 23, 27, and 44.]
- [6] D. Angelov, “Top2vec: Distributed representations of topics,” *CoRR*, vol. abs/2008.09470, 2020. [Cited on pages 3, 9, 23, and 44.]
- [7] M. Grootendorst, 2022. [Cited on pages 3, 4, 10, 23, 30, and 42.]
- [8] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “Topic modeling in embedding spaces,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Cited on pages 3, 11, 16, 23, and 30.]
- [9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. [Cited on page 7.]
- [10] T. Hofmann, “Probabilistic latent semantic analysis,” in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 289–296. [Cited on pages 7 and 23.]

- [11] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, no. 4, p. 77–84, Apr. 2012. [Online]. Available: <https://doi.org/10.1145/2133806.2133826> [Cited on page 7.]
- [12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006. [Online]. Available: <https://doi.org/10.1198/016214506000000302> [Cited on page 8.]
- [13] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999. [Cited on page 8.]
- [14] —, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001. [Cited on pages 8, 23, and 44.]
- [15] D. M. Blei and J. D. Lafferty, “A correlated topic model of science,” *The annals of applied statistics*, pp. 17–35, 2007. [Cited on page 8.]
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/> [Cited on page 9.]
- [17] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084> [Cited on pages 9 and 30.]
- [18] L. McInnes and J. Healy, “Umap: Uniform manifold approximation and projection for dimension reduction,” 02 2018. [Cited on page 10.]
- [19] M. Mendonça and □Figueira, “Topic extraction: Bertopic’s insight into the 117th congress’s twitterverse,” *Informatics*, vol. 11, no. 1, 2024. [Online]. Available: <https://www.mdpi.com/2227-9709/11/1/8> [Cited on pages 11, 19, and 30.]
- [20] F. Bianchi, S. Terragni, and D. Hovy, “Pre-training is a hot topic: Contextualized document embeddings improve topic coherence,” in *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 759–766. [Cited on page 11.]

- [21] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 80–88. [Online]. Available: <https://doi.org/10.1145/1964858.1964870> [Cited on page 11.]
- [22] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, “Improving lda topic models for microblogs via tweet pooling and automatic labeling,” in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’13. New York, NY, USA: Association for Computing Machinery, 2013, p. 889–892. [Online]. Available: <https://doi.org/10.1145/2484028.2484166> [Cited on page 11.]
- [23] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, “Comparing twitter and traditional media using topic models,” in *European conference on information retrieval*. Springer, 2011, pp. 338–349. [Cited on page 11.]
- [24] X. Yan, J. Guo, Y. Lan, and X. Cheng, “A biterm topic model for short texts,” in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW ’13. New York, NY, USA: Association for Computing Machinery, 2013, p. 1445–1456. [Online]. Available: <https://doi.org/10.1145/2488388.2488514> [Cited on page 12.]
- [25] J. Eisenstein, A. Ahmed, and E. P. Xing, “Sparse additive generative models of text,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 1041–1048. [Cited on page 12.]
- [26] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, “Representation learning for very short texts using weighted word embedding aggregation,” *Pattern Recognition Letters*, vol. 80, p. 150–156, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2016.06.012> [Cited on page 12.]
- [27] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Topic modeling for short texts with auxiliary word embeddings,” in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’16.

- New York, NY, USA: Association for Computing Machinery, 2016, p. 165–174. [Online]. Available: <https://doi.org/10.1145/2911451.2911499> [Cited on page 12.]
- [28] M. D. Hoffman, D. M. Blei, and F. Bach, “Online learning for latent dirichlet allocation,” in *Advances in Neural Information Processing Systems*, 2010. [Cited on page 13.]
- [29] L. Al-Sumait, D. Barbará, and C. Domeniconi, “On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking,” in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 3–12. [Cited on page 13.]
- [30] R. Ibrahim, A. Elbagoury, M. Kamel, and F. Karray, “Tools and approaches for topic detection from twitter streams: survey,” *Knowledge and Information Systems*, vol. 54, pp. 511–539, 2018. [Cited on page 13.]
- [31] N. Miyamoto, M. Isonuma, S. Takase, J. Mori, and I. Sakata, “Dynamic structured neural topic model with self-attention mechanism,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5916–5930. [Online]. Available: <https://aclanthology.org/2023.findings-acl.366/> [Cited on page 13.]
- [32] K. Cvejoski, R. J. Sánchez, and C. Ojeda, “Neural dynamic focused topic model,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. [Cited on page 13.]
- [33] X. Wu, X. Dong, L. Pan, T. Nguyen, and A. T. Luu, “Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion,” in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. [Cited on page 13.]
- [34] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, oct 2008. [Online]. Available: <https://doi.org/10.1088/1742-5468/2008/10/P10008> [Cited on page 13.]
- [35] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007. [Cited on page 13.]

- [36] G. Palla, A.-L. Barabási, and T. Vicsek, “Quantifying social group evolution,” *Nature*, vol. 446, no. 7136, pp. 664–667, 2007. [Cited on page 14.]
- [37] D. Chakrabarti, R. Kumar, and A. Tomkins, “Evolutionary clustering,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 554–560. [Cited on page 14.]
- [38] D. Greene, D. Doyle, and P. Cunningham, “Tracking the evolution of communities in dynamic social networks,” in *2010 international conference on advances in social networks analysis and mining*. IEEE, 2010, pp. 176–183. [Cited on page 15.]
- [39] S. Asur, S. Parthasarathy, and D. Ucar, “An event-based framework for characterizing the evolutionary behavior of interaction graphs,” *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 4, Dec. 2009. [Online]. Available: <https://doi.org/10.1145/1631162.1631164> [Cited on page 15.]
- [40] P. Bródka, S. Saganowski, and P. Kazienko, “Ged: the method for group evolution discovery in social networks,” *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 1–14, 2013. [Cited on page 15.]
- [41] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, “Textflow: Towards better understanding of evolving topics in text,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011. [Cited on page 15.]
- [42] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky, “Hierarchical topics: Visually exploring large text collections using topic hierarchies,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2002–2011, 2013. [Cited on page 15.]
- [43] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. Mccallum, “Optimizing semantic coherence in topic models,” 01 2011, pp. 262–272. [Cited on page 16.]
- [44] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: <https://doi.org/10.1145/2684822.2685324> [Cited on page 16.]

- [45] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, “Reading tea leaves: How humans interpret topic models,” in *Advances in Neural Information Processing Systems*, 2009. [Cited on page 16.]
- [46] A. Hoyle, P. Goel, D. Peskov, J. Boyd-Graber, and P. Resnik, “Is automated topic model evaluation broken?: The incoherence of coherence,” in *Advances in Neural Information Processing Systems*, 2021. [Cited on page 16.]
- [47] M. Roder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408. [Cited on page 16.]
- [48] M. Rüdiger, D. Antons, A. Joshi, and T. O. Salge, “Topic modeling revisited: New evidence on algorithm performance and quality metrics,” *PLoS ONE*, vol. 17, no. 4, 2022. [Cited on page 16.]
- [49] H. Rahimi, D. Mimno, J. L. Hoover, H. Naacke, C. Constantin, and B. Amann, “Contextualized topic coherence metrics,” in *Findings of the European Chapter of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 1760–1773. [Cited on page 16.]
- [50] Z. Tan and J. D’Souza, “Toward purpose-oriented topic model evaluation enabled by large language models,” *International Journal on Digital Libraries*, vol. 26, 2025. [Cited on page 16.]
- [51] S. Adhya, A. Lahiri, D. Sanyal, and P. Das, “Improving contextualized topic models with negative sampling,” 03 2023. [Cited on page 16.]
- [52] R. Churchill and L. Singh, “The evolution of topic modeling,” *ACM Computing Surveys*, 2022. [Cited on page 17.]
- [53] A. Hosseiny Marani and E. P. S. Baumer, “A review of stability in topic modeling: Metrics for assessing and techniques for improving stability,” *ACM Computing Surveys*, 2024. [Cited on page 17.]
- [54] C. Karakkaparambil James, M. Nagda, N. Haji Ghassemi, M. Kloft, and S. Fellenz, “Evaluating dynamic topic models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association

- for Computational Linguistics, Aug. 2024, pp. 160–176. [Online]. Available: <https://aclanthology.org/2024.acl-long.11/> [Cited on page 17.]
- [55] S. Mehta *et al.*, “Concept drift in streaming data classification: algorithms, platforms and issues,” *Procedia computer science*, vol. 122, pp. 804–811, 2017. [Cited on page 17.]
- [56] A. Pesaranghader, H. L. Viktor, and E. Paquet, “Mcdiarmid drift detection methods for evolving data streams,” in *2018 International joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–9. [Cited on page 17.]
- [57] D. Knights, M. Mozer, and N. Nicolov, “Detecting topic drift with compound topic models,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, no. 1, 2009, pp. 242–245. [Cited on page 17.]
- [58] R. Coleman, M. McCombs, D. Shaw, and D. Weaver, “Agenda setting,” in *The handbook of journalism studies*. Routledge, 2009, pp. 167–180. [Cited on page 18.]
- [59] R. M. Entman *et al.*, “Framing: Towards clarification of a fractured paradigm,” *McQuail’s reader in mass communication theory*, vol. 390, p. 397, 1993. [Cited on page 18.]
- [60] D. Card, A. Boydston, J. H. Gross, P. Resnik, and N. A. Smith, “The media frames corpus: Annotations of frames across issues,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 438–444. [Cited on page 18.]
- [61] M. Gentzkow, J. M. Shapiro, and M. Taddy, “Measuring group differences in high-dimensional choices: method and application to congressional speech,” *Econometrica*, vol. 87, no. 4, pp. 1307–1340, 2019. [Cited on page 18.]
- [62] J. Grimmer, “A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases,” *Political Analysis*, vol. 18, no. 1, p. 1–35, 2010. [Cited on page 18.]
- [63] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, “Exposure to opposing views on social media can increase political polarization,” *Proceedings of the National*

- Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018. [Online]. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.1804840115> [Cited on page 18.]
- [64] J. Golbeck, J. M. Grimes, and A. Rogers, “Twitter use by the u.s. congress,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 8, pp. 1612–1621, 2010. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21344> [Cited on page 19.]
- [65] P. BARBERÁ, A. CASAS, J. NAGLER, P. J. EGAN, R. BONNEAU, J. T. JOST, and J. A. TUCKER, “Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data,” *American Political Science Review*, vol. 113, no. 4, p. 883–901, 2019. [Cited on page 19.]
- [66] D. Greene, D. O’Callaghan, and P. Cunningham, “How many topics? stability analysis for topic models,” 2014. [Online]. Available: <https://arxiv.org/abs/1404.4606> [Cited on page 27.]
- [67] V. Andrei and O. Arandjelovic, “Complex temporal topic evolution modelling using the kullback-leibler divergence and the bhattacharyya distance,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2016, 12 2016. [Cited on page 27.]
- [68] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *The Journal of Open Source Software*, vol. 2, p. 205, 03 2017. [Cited on page 30.]
- [69] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Topic modeling for short texts with auxiliary word embeddings,” in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 165–174. [Online]. Available: <https://doi.org/10.1145/2911451.2911499> [Cited on page 30.]
- [70] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.10957> [Cited on page 36.]
- [71] 100xTools, “Twexportly: Export tweets from any account,” 2025. [Online]. Available: <https://chromewebstore.google.com/detail/twexportly-export-tweets/hbibehafoapglhcgfhlpifagloecmhfh> [Cited on page 38.]
- [72] X. Corporation, “X app (formerly twitter),” 2025. [Online]. Available: <https://x.com> [Cited on page 38.]

- [73] C. R. Service, “Membership of the 119th congress: A profile,” 2025. [Online]. Available: <https://crsreports.congress.gov/product/pdf/R/R48535> [Cited on page 38.]
- [74] U.S. Senate, “Dates of sessions of the congress,” 2025. [Online]. Available: <https://www.senate.gov/legislative/DatesofSessionsofCongress.htm> [Cited on page 38.]
- [75] United Nations / Kaggle, “Un general debates dataset,” <https://www.kaggle.com/datasets/unitednations/un-general-debates>, accessed in: 01 Dez 2025. [Cited on page 39.]
- [76] Rachael Tatman / Kaggle, “State of the union corpus (1790–2018),” <https://www.kaggle.com/datasets/rtatman/state-of-the-union-corpus-1989-2017>, accessed in: 01 Dez 2025. [Cited on page 39.]
- [77] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 399–408, 02 2015. [Cited on page 45.]
- [78] S. Sawant, J. Yu, K. Pandya, C.-K. Ngan, and R. Bardeli, “An enhanced bertopic framework and algorithm for improving topic coherence and diversity,” in *2022 IEEE 24th Int Conf on High Performance Computing Communications; 8th Int Conf on Data Science Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys)*, 2022, pp. 2251–2257.
- [79] P. Pitichotchokphokhin, P. Chuangkrud, K. Kalakan, B. Suntisrivaraporn, T. Leelanupab, and N. Kanungsukkasem, “Discover underlying topics in thai news articles: A comparative study of probabilistic and matrix factorization approaches,” in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2020, pp. 759–762.
- [80] K. Shaikh, H. Ahmed, M. F. U. Haque, P. C. Santos, H. R. Khan, K. Arshad, and K. Assaleh, “Analyzing risc-v compiler toolchain by adopting topic modeling,” *IEEE Access*, vol. 14, pp. 33 178–33 198, 2026. [Cited on page 45.]

Appendix

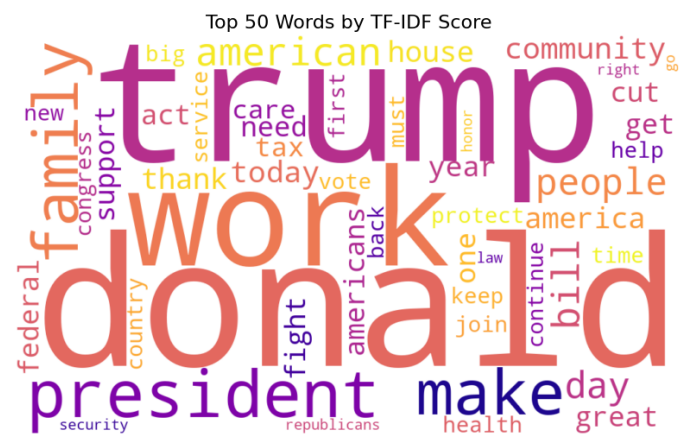


Figure 1: Word Cloud of the Top 50 words ranked by their TF-IDF scores.

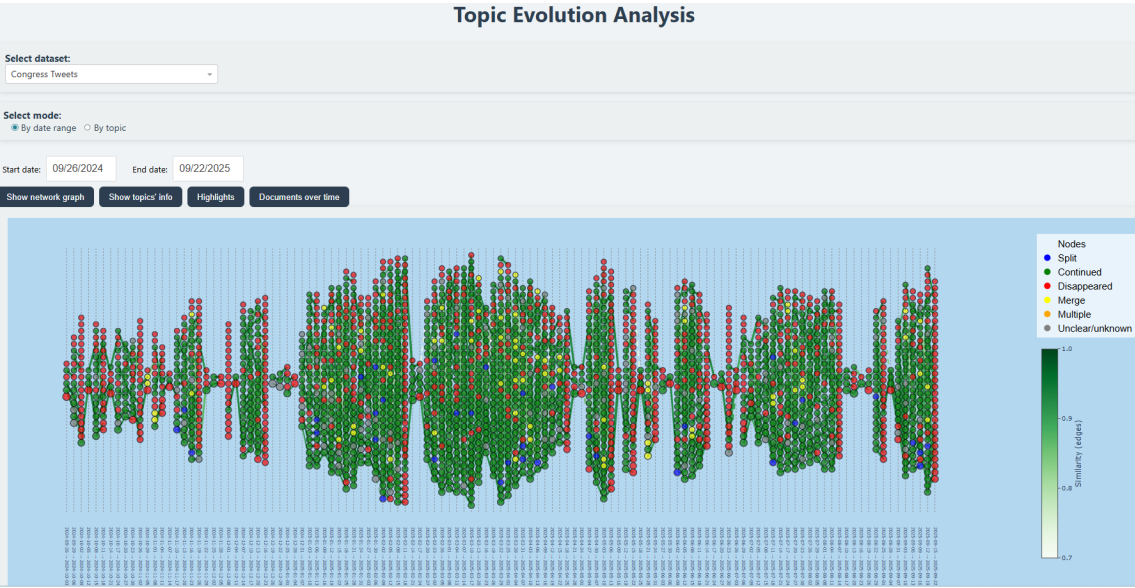


Figure 2: Dashboard Layout