ISBN 978-9934-619-63-2

Narrative Detection and Topic Modelling in the Baltics

PREPARED AND PUBLISHED BY THE
NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-63-2

Authors: Eduard Barbu, Somnath Banerjee, Marija Isupova, Yukai Zeng
Project Managers: Yukai Zeng, Marija Isupova
Content Editor: Monika Izandra Hanley
Design: Una Grants

Cover Image Content generated by AI (DALL-E)

Riga, April 2024

NATO STRATCOM COE
11b Kalnciema iela,
Riga, LV1048, Latvia
stratcomcoe.org
@stratcomcoe

Narrative Detection and Topic Modelling in the Baltics

Contents

Introduction

Current techniques for topic modelling and narrative detection are optimised for English language content, primarily catering to English-centric or other widely spoken languages, and do not perform as well on less spoken languages.[1] These less-spoken languages, including those of the Baltic region, often contain unique cultural and contextual nuances not captured well by models trained in major languages. This issue is compounded for less spoken languages, where the challenge of accurately capturing cultural nuances becomes even more pronounced, leading to inaccurate or incomplete analysis.[2] In strategic communications, these limitations pose a significant challenge in effectively analysing the digital information environment.[3] Furthermore, equalising strategic communications capabilities across allies is essential. As highlighted in our report, AI in Support of StratCom Capabilities[4], this research aims to bring parity in strategic communication tools and practices among allies.[5] Disinformation transcends geographic boundaries, and technological limitations in the digital space necessitate the development of robust capabilities for defence against such challenges. This research aims to illuminate gaps in current methodologies that require further exploration and resolution. This report evaluates the potential of the languages of the Baltic States—Estonian, Latvian, and Lithuanian—for topic modelling and narrative detection. Specifically, we will focus on annotated datasets and open-source tools suitable for executing these tasks.

The structure of this report is organised as follows. Sections 2 and 3 delve into our primary areas of interest: topic modelling and narrative detection, respectively. Given the term's ambiguous interpretation within existing literature, narrative detection receives an in-depth examination. With diverse approaches identified for narrative detection, we have aimed to distil a shared framework from the myriad tasks associated with this designation.

In Section 4, we evaluate the existing capabilities in executing the two tasks for each Baltic language, primarily focusing on research conducted in academic settings. The emergence of Large Language Models (LLMs) like Generative Pre-trained Transformer 4 (GPT-4) has greatly enhanced the capabilities of Natural Language Processing (NLP) technologies. GPT-4 is the latest iteration of OpenAI's advanced language model series.[6] It represents a significant advancement in artificial intelligence, building upon the success of its predecessor, GPT-3.[7] These models are based on transformer architecture, a type of deep neural network, and obtain state-of-the-art performance for various language-related tasks. However, developing and training such advanced LLMs demands significant computational power and expertise, restricting this ability to a few entities with access to high-end computational resources and specialised knowledge. The cost of training a model like GPT-4 can vary widely, ranging from several million to tens of millions of dollars, influenced by factors like scale, training efficiency, and model-specific requirements. Currently, there are no Baltic language LLMs with capabilities comparable to GPT-4. Therefore, our review concentrates on works preceding the introduction of these advanced architectures.

Nevertheless, section 5 evaluates GPT-4's effectiveness in narrative detection. In particular, for the Estonian language, we evaluate the capabilities of GPT-4 for performing narrative detection using a small set of Estonian news articles about unauthorised mass migration from Belarus to Latvia.

Topic Modelling

Topic modelling is a collection of statistical techniques designed to uncover the latent topics embedded within a corpus of documents. It provides a structured method to structure vast amounts of textual data. Such models are good at organising large datasets, reducing dimensionality, and unveiling hidden semantic patterns. Various topic modelling methodologies include:

- **Latent Dirichlet Allocation (LDA)**[8]: Often considered the cornerstone of topic modelling, LDA sees documents as composed of various topics, while a topic comprises different words. LDA identifies topics by examining word distribution across documents, subsequently determining both topic prevalence in documents and word prevalence within topics.

- **Hierarchical Topic Models**[9]: This extends basic topic models by identifying a hierarchy or tree of topics, where topics at each level of the tree are derived from the topics at the level above. This can be useful in capturing more fine-grained or specific topics nested within broader ones.

- **Unsupervised Topic Models:** These models, like LDA, do not use any labels or metadata and rely purely on the word distributions within documents to identify topics. They aim to capture the data's inherent structures without prior knowledge or external guidance.

- **Supervised Topic Models**[10]: While conventional topic models are unsupervised and focus solely on discovering topics from the data, supervised topic models incorporate metadata or labels associated with documents to influence the topic discovery process. This can be particularly beneficial when certain topics are expected or desired based on external factors or known categories.

Topic models are powerful tools that help make sense of large amounts of information, including highlighting hidden patterns and themes. For instance, they help websites suggest articles you might like based on your reading. They can take a big pile of information and create a summary that captures the main points. Researchers use them to sift through massive datasets to identify key topics.

Neural topic modelling[11] is an approach that applies neural network architectures to the problem of topic modelling, which traditionally has been addressed by probabilistic methods like LDA. Neural topic models can potentially model complex patterns in data and are more flexible in incorporating additional contexts such as document metadata or pre-trained word embeddings. They can also be scaled to handle very large datasets efficiently.

On social media, they help spot what topics are trending or becoming popular. They are even used in fields like history and literature to study big collections of texts, helping scholars uncover new insights about the past and cultural trends. Below is an example of a hypothetical topic model applied to a political text about the recent situation in the Middle East. The source of the text is part of an article published by *The Atlantic* by Gal Beckerman on 7 October 2023, The Middle East Region Is Quieter Today than It Has Been in Two Decades.[12]

The provided text is segmented based on hypothetical topic modelling results into distinct blocks corresponding with paragraphs where the text subject changes. Above each block, a bolded topic label is added to describe the main theme of that paragraph. The text is divided into four sections, each with its own topic: National Security & Middle East Developments, Conflict & Hamas Attack,

Military Mobilisation & Israeli Response, and American Perception & Analysis.

National Security & Middle East Developments: Just eight days ago, National Security Adviser Jake Sullivan, speaking at The Atlantic Festival, rattled off a long list of positive developments in the Middle East, developments that were allowing the Biden administration to focus on other regions and other problems. A truce was holding in Yemen. Iranian attacks against U.S. forces had stopped. America's presence in Iraq was "stable." The good news crescendoed with this statement: "The Middle East region is quieter today than it has been in two decades."

Conflict & Hamas Attack: One week later, a shocking, multifront attack launched by Iranian-supported Hamas against Israel has turned the Middle East into a maelstrom. The assault, almost 50 years to the day after the surprise Arab attack on Israel that marked the opening of the Yom Kippur War, could represent a paradigm-shifting moment as big as 9/11. So far, more than 100 Israelis are confirmed dead and many hundreds more gravely injured in a coordinated attack by Hamas terrorists who infiltrated by land, sea, and air. A thousand tragedies will unfold—at the moment, an unknown number of Israeli civilians and soldiers might be held hostage in Gaza. As of this writing, nearly 200 are reported dead in Israeli reprisal raids.

Military Mobilisation & Israeli Response: The Israeli army has activated at least 100,000 reservists, and a full-scale ground invasion of Gaza is plausible, if not probable. Behind this moment are failures of intelligence but also of imagination. The Israeli government, led by Prime Minister Benjamin Netanyahu, who has styled himself as "Mr. Security" for decades, will have much to answer for in the coming weeks and months.

American Perception & Analysis: But Sullivan's comments, made onstage in Washington to *The Atlantic's* editor-in-chief, Jeffrey Goldberg, also suggest how little sense there was among Biden officials that something like this could happen. "Challenges remain," Sullivan said in his comments last week. "Iran's nuclear weapons program, the tensions between Israelis and Palestinians. But the amount of time I have to spend on crisis and conflict in the Middle East today, compared to any of my predecessors going back to 9/11, is significantly reduced."

Narrative Detection

Despite the diversity in data and operations labelled as narrative detection[13], three elements stand consistent in all automatic narrative detection efforts: named entity recognition (NER), identifying the semantic relationships between these entities (relationship extraction), and identifying plots.

Named Entity Recognition (NER)

NER and its subtask, fine-grained NER, are essential components of information extraction that aim to categorise named entities in unstructured text into set categories like persons, organisations, and locations, among others. Fine-grained NER takes this further by identifying more specific categories within these broader groups, such as differentiating a city from a country.

The two main methods for automated entity learning are rule-based systems and machine learning techniques. Rule-based systems use manually devised patterns to identify entities within text. However, this method's rigidity can hinder its ability to adapt to the nuanced nature of language, potentially affecting identification accuracy.

On the other hand, machine learning techniques provide more flexibility and accuracy. They operate by training models on datasets with text tokens mapped to specific entity tags, allowing the models to identify and categorise entities in new texts effectively. This adaptability enables more robust and accurate entity recognition.

The latest advances in NER use pretrained models based on transformer architecture.[14] These architectures[15] undergo a two-step process of pretraining and fine-tuning. In pretraining, models are exposed to large datasets, learning to predict the next word and capturing contextual information through attention mechanisms. In the fine-tuning stage, the pre- trained models are further trained on smaller, task-specific datasets. Despite these advances, challenges remain, especially in the domain adaptability of NER models and identifying specific entity types, such as events, which vary significantly based on narrative context.

Entities like actors, events, and timelines play a significant role in extracting semantic insights from narratives.

In the provided text sample, specific words or phrases are marked to denote various entities, including actors, events, and timelines. These entities are encapsulated in brackets with their respective category noted in subscript.

Consider the following examples:

- **Summer of 1969** is identified as a timeline
- **Neil Armstrong** and **Buzz Aldrin** are labelled as actors.
- **Apollo 11 mission** is marked as an event.

This approach to tagging allows for the effortless identification and classification of

diverse entities within the text, facilitating the extraction of valuable information from narratives, as demonstrated in example 3.1.

> In the [summer of 1969]$_{timeline}$, [Neil Armstrong]$_{actor}$ and [Buzz Aldrin]$_{actor}$ made history by becoming the first humans to land on the moon during the [Apollo 11 mission]$_{event}$. This incredible [event]$_{event}$ was watched on television by an estimated 600 million people worldwide. [Armstrong]$_{actor}$'s famous words as he stepped onto the lunar surface, "That's one small step for man, one giant leap for mankind," are still remembered and quoted today.
>
> Fast forward to [July 20, 2019]$_{timeline}$, which marked the [50th anniversary]$_{timeline}$ of the [Apollo 11 moon landing]$_{event}$. Various [events]$_{event}$ and celebrations were held around the world to commemorate this significant milestone in human history. [NASA]$_{actor}$ celebrated the anniversary with a live television broadcast and events at the [Kennedy Space Center]$_{actor}$, highlighting the achievements of [Armstrong]$_{actor}$, [Aldrin]$_{actor}$, and the many others who contributed to the successful [Apollo 11 mission]$_{event}$.

Relationship extraction (RE)

The next layer in implementing narrative detection building on the NER is relationship extraction (RE). As NER, RE is a sub-task of information extraction in NLP that focuses on identifying and classifying semantic relationships between entities mentioned in a text. For narrative detection, the goal is to convert unstructured text data into structured information automatically. This involves extracting the relationships between actors, events, and timelines.[16] Older approaches to RE were based on expert-devised lexico-syntactic patterns.[17] The patterns are designed to capture how relationships between entities are typically expressed in natural language text. Examples of lexico-syntactic patterns for encoding the hypernym relation for the English language are "X is a Y," "X such as Y," "X or other Y," and "X and other Y." Applying the second pattern ("X such as Y") to the text below would yield the following extracted relationship: state-controlled media outlets –> sources of disinformation, online proxy media outlets –> sources of disinformation, social media operations –> sources of disinformation.

> "State-controlled media outlets such as RT and Sputnik, along with online proxy media outlets and malign social media operations, are often implicated in spreading disinformation to advance national interests."

Here, "sources of disinformation" is identified as the hypernym (general term), and "state-controlled media outlets," "online proxy media outlets," and "social media operations" are identified as hyponyms (specific terms), indicating that these are types of sources that can be involved in the dissemination of disinformation.[18] Cutting-edge techniques for relation extraction in NLP have markedly advanced with the emergence of deep learning and neural networks. State-of-the-art methods include transformer models such as BERT[19], GPT-x[20], and Graph Neural Network[21] architectures. Additionally, the distant supervision approach[22] auto-labels training data using large knowledge base entries or leveraging Wikipedia. In narrative detection, the typical relationships focus on the interaction and connections among actors, events, and timelines within the narrative.

Common types of relations include:

1. Causal Relationships:

Description: Identifies cause-and-effect relationships between events or actions within the narrative.

Example: An event (E1) leads to another event (E2).

2. Temporal Relationships:

Description: Establishes the chronological order of events or actions.

Example: An event (E1) occurs before another event (E2).

3. Spatial Relationships:

Description: Outlines the geographic or spatial connections between actors or events.

Example: An actor (A1) is located in a place (P1).

4. Social Relationships:

Description: Describes the interactions and relationships between actors.

Example: An actor (A1) is a friend of another actor (A2).

5. Participation Relationships

Description: Describes how actors are involved in events.

Example: An actor (C1) participates in an event (E1).

Considering the text from example 3.1, a Relationship Extraction Module for narrative detection should identify the following types of relations.

- **Temporal Relationship**

– [Summer of 1969] is temporally related to the event [Apollo 11 mission]. This relationship establishes that the Apollo 11 mission occurred during the summer of 1969.

- **Participation Relationship:**

– the actors [Neil Armstrong] and [Buzz Aldrin] participate in [Apollo 11 mission]. This relationship identifies Neil Armstrong and Buzz Aldrin as the key actors in the Apollo 11 mission.

- **Spatial Relationship**

– [Neil Armstrong] is spatially related to [moon]. This relationship identifies that Neil Armstrong was on the moon.

- **Social Relationships:**

– [NASA] is socially (organisationally) related to [Armstrong] and [Aldrin]. In the context of the relationship between NASA, Armstrong, and Aldrin, it may be more accurate to classify it as an organisational or affiliation relationship. This would represent the professional connection between Armstrong and Aldrin as astronauts and NASA as the space agency they worked for. Organisational relationships could be taught as a subtype of social relationships. This classification allows for a more granular and specific understanding of the relationships while still falling under the broader category of social relationships.

Plot discovery and story evolution

In our prior explorations of NER and RE, we discussed the identification of entities and the connections between them. From an NER perspective, this means distinguishing key 'events' and 'characters' as separate entities. Extending this with Relationship Extraction allows us to understand the evolving dynamics among characters and their actions over time. Building on this foundation, identifying plots becomes the subsequent phase in narrative analysis.[23] Essentially, plots chart the course of events and character developments within a narrative, intertwining various sub-stories. The principal components of plots, also known as Freytag's Pyramid, consist of the following steps:

- **Exposition**: Introduces characters, setting, and conflict. The **setting** is the backdrop, specifying time, place, and environment influencing characters and events.
- **Rising Action**: Events post-setup, escalating tension and complexity, with characters facing challenges leading up to the climax.
- **Climax**: The story's critical moment where the main conflict intensifies, marking a turning point and setting the stage for resolution.
- **Falling Action**: Events after the climax, signifying conflict resolution beginnings, with decreasing tension and complications.
- **Resolution**: The story's final phase, addressing unresolved matters, depicting characters' outcomes, and ending the narrative.

For example, consider the following narrative about the unfolding of the COVID-19 pandemic.

> In December 2019, the city of Wuhan in China reported mysterious cases of pneumonia, marking the beginning of an unprecedented global health crisis. Scientists and doctors raced against time to identify this new threat, revealing a novel coronavirus. As the world watched with bated breath, the virus spread across continents, bringing bustling metropolises to a standstill.
>
> Governments scrambled to respond, imposing lockdowns and closing borders. Frontline healthcare workers became the heroes of this tale, battling the surge in patients despite dwindling resources. Researchers and pharmaceutical companies worldwide collaborated in the search for a vaccine at a pace never seen before in medical history.
>
> Despite the fear and despair, communities came together in solidarity. Families sang from balconies, and strangers delivered groceries to the elderly. Yet, the socioeconomic disparities became evident, and debates about public health versus economic stability raged on.
>
> By the end of 2020, vaccines were rolled out, providing a glimmer of hope. Slowly, the world adapted to a 'new normal,' redefining work, travel, and social interactions. The story is ongoing, with lessons on resilience, cooperation, and the indomitable human spirit.

- **Exposition**: Discovering mysterious pneumonia cases in Wuhan, China, and identifying a novel coronavirus.
- **Rising Action**: The spread of the virus worldwide, governments' responses with lockdowns and border closures, and the global race for a vaccine.
- **Climax**: The intense strain on healthcare systems, the socio-economic debates, and

the world grappling with an unforeseen challenge.

- **Falling Action**: The development and distribution of vaccines, the emergence of global solidarity, and the adaptation to new ways of living and working.

- **Resolution**: The world is beginning to embrace a 'new normal' and drawing lessons from the pandemic, even though the story of COVID-19 is yet to see a definitive end.

Story Evolution delves deeper into the understanding of how narratives progress over time. However, while the plot gives us the overarching structure, the evolution of the story brings out the finer details and the nuances. Usually, the story evolution has two components.[24]

1. Shift Detection: This involves recognising significant changes or deviations in the narrative. These shifts could be in the form of major turning points, character dynamics alterations, theme changes, or tone shifts.

2. Causal Relation Extraction: A deeper understanding of the narrative necessitates comprehending the cause-and-effect relationships embedded within. This component aims to unveil the reasons behind every significant event or decision in the story. By understanding causality, one can predict future narrative progressions or backtrack to the origins of certain events or conflicts.

The analysis of the evolution of the previous story about COVID-19 helps to understand the pivotal moments (shifts) in the narrative and the causal relationships that drive the story's progression.

- **Shift Detection**: The story witnessed several major shifts. From the initial discovery of the virus to its global spread, to vaccine development, and finally to global adaptation.

- **Causal Relation Extraction**: The cause-and-effect relationships are evident throughout. The discovery of the virus (cause) led to global alertness (effect). The rapid spread of infections (cause) resulted in lockdowns and a global crisis (effect). Vaccine development (cause) led to a decrease in infections and a ray of hope (effect).

The interplay between Plots and Story Evolution in the COVID-19 Narrative works as follows.

1. Exposition and Shift Detection: The exposition of the novel virus sets the initial shift, moving the world from a state of normalcy to heightened alertness.

2. Rising Action and Causal Relation Extraction: The rising action, characterised by the global spread, had a direct causal relationship with worldwide responses like lockdowns and travel bans.

3. Climax and Major Turning Point: The major turning point was the climax, where infections peaked. It dictated global responses and catalysed rapid vaccine research and development.

4. Falling Action and Adaptation: The distribution of vaccines and global adaptation to the new normal showcases a causative relationship where the availability of a preventive measure (vaccine) led to a decrease in infections and eventual societal adaptation.

5. Resolution and Ongoing Evolution: The resolution is ongoing and an evolving response to the pandemic. This continuous evolution is influenced by the learnings from the previous plot points and their causative outcomes.

Narrative Detection and Topic Modelling in the Baltics

Estonian

EstNLTK[25] is an open-source tool for Estonian NLP. EstNLTK provides common NLP functionality such as paragraph, sentence, and word tokenisation, morphological analysis, and NER. Tools like EstNLTK and open-source implementations of topic modelling packages, like LDA, are sufficient for performing topic modelling in Estonian. The following is a typical workflow for performing traditional topic modelling using EstNLTK:

1. Data Collection: Gather the textual data in Estonian on which you intend to perform topic modelling.

2. Text Pre-processing with EstNLTK:

A **Tokenisation:** Split the text into sentences and words using EstNLTK's tokenisation functions.

B **Morphological Analysis:** Analyse the morphological structure of words to extract root forms and assign part-of-speech-tags.

C **NER (Optional):** Identify and possibly exclude named entities if they are not relevant to your topic modelling goals.

3. Topic Modelling: After pre-processing the text, transition it into a bag-of-words or Term Frequency - Inverse Document Frequency (TF-IDF) representation. To extract topics from this representation, leverage an open-source topic modelling package such as LDA. It is essential to highlight that EstNLTK does not come equipped with LDA, so turning to other libraries like Gensim or Scikit-learn can be beneficial. Once the modelling begins, determining the optimal number of topics is crucial. Achieving this might require several iterations, adjusting the topic count, and assessing the resulting topics for coherence and distinctiveness.

4. Evaluation & Analysis: Review the relevance of words associated with each topic. To quantitatively evaluate the topics, we can employ topic coherence metrics. Additionally, visualisation tools, such as pyLDAvis, offer insightful perspectives into the topics and their significance.

5. Post-processing (optional): Depending on the application, you might wish to cluster topics further, associate them with metadata, or integrate them into other applications.

6. Iterative Refinement: Refinement is often needed in topic modelling. Based on the results, adjustments may be required in pre-processing, modelling, or the number of topics.

A major advantage of neural topic modelling is its capacity to largely operate in a language-agnostic manner. This quality arises because neural networks, especially those employed for topic modelling, are intrinsically built to discern patterns from data, irrespective of the language involved. Consequently, neural topic modelling can be effectively applied to the Estonian language in a way that is independent of language-specific constraints.

Previous research on Estonian topic modelling, the results of which can be reused in the refinement step, is presented in Barbu et al., 2018.[26] They empirically identified the best number of topics for a 185 million newspaper

corpus. These topics were translated into English and annotated by two individuals. Subsequently, Estonian and Princeton WordNet were employed to find word pairs in topics with notable taxonomic similarity. Tools developed in the European project "Microservices at your Service – Bridging the Gap between NLP Research and Industry"[27] can be adapted for narrative topic modelling and narrative detection tasks. This project focuses on dockerising specialised language technology services and equipping them with a web API, facilitating easy reuse by stakeholders such as researchers and the industry.[28] The catalogue of microservices for the Estonian language can be accessed at the following link in the footnote.[29]

Narrative detection has not been previously attempted for the Estonian language, leading to a lack of domain-specific annotated datasets. Yet, based on the task definition provided in the prior section, existing datasets and code can be repurposed to facilitate narrative detection in Estonian. Data and code support the initial two steps essential for narrative detection in Estonian: NER and RE.

The primary references for NER in the Estonian language are outlined in Sirts, 2023.[30] The first dataset comes annotated with a three-tier taxonomy of entities, encompassing common types such as Persons, Locations, and Organisations and distinctive types like Product and Title. Detailed statistics for this dataset are available at: github.com/TartuNLP/EstNER. The second dataset is a compilation of news and media texts annotated with a hierarchical entity structure. In-depth statistics for this latter dataset can be found at: github.com/TartuNLP/EstNER_new.

There is an NLP task that is usually glossed over when we talk about narrative detection coreference resolution. Coreference resolution in NLP determines which words (or phrases) in a text refer to the same entity. Specifically, it involves identifying all expressions, called mentions, in a text that refer to the same underlying entity or concept. These mentions can include pronouns (e.g., "she," "it"), definite noun phrases (e.g., "the car"), and names (e.g., "John"). For pronominal coreference resolution in Estonian, the EstAnaphora corpus is available and can be accessed at github.com/EstSyntax/EstAnaphora. This corpus comprises texts from Estonian newspapers, magazines, and a scientific journal covering the period from 1998 to 2007. With approximately 253,000 words, EstAnaphora includes 7,250 nominal coreference pairs. These pairs consist of a pronoun and its referent, which could be a common noun, a proper noun, or another pronoun. Event coreference involves identifying instances within a text or collection of texts where different expressions mention the same event. Essentially, it ascertains whether multiple mentions of an event within a document correspond to the same real-world occurrence, regardless of differences in phrasing or specifics. An initial investigation into event coreference can be found in Orasmaa, 2015.[31] While the study uses an annotated dataset, it remains unpublished but can be accessed upon request. Temporal expressions in the Estonian language are annotated within the Estonian TimeML Corpus (github.com/soras/EstTimeMLCorpus), a subset of the Estonian Dependency Treebank. [32] This corpus comprises 80 newspaper articles in Estonian, totalling around 22,000 word tokens. Each article comes with manually refined morphological and dependency syntactic annotations supplemented with temporal semantic annotations.

EstBERT[33], a model tailored for the Estonian language, is based on the BERT architecture developed by Google. BERT is a deep learning model in NLP that utilises bidirectional transformers to comprehend word contexts in sentences. Specifically fine-tuned on Estonian texts, EstBERT captures the language's nuances and grammar, making it appropriate for various Estonian NLP tasks, including NER. By leveraging this pretrained model, researchers achieve state-of-the-art performance in NER for Estonian.

To our knowledge, one of this study's co-authors conducted the only study on semantic relationship extraction for the Estonian language. Led by the Estonian Military Academy,

the study utilised a cutting-edge relation extraction system rooted in the BERT architecture to identify entities for military intelligence and their semantic relations automatically. While the code for relationship extraction is not publicly accessible, it can be obtained upon request from the author. A pronominal coreference resolver was presented in Barbu et al., 2020.[34] The pronominal coreference resolution module trained on the EstAnaphora corpus contains different ML algorithms and is freely available at Estonian Coreference GitHub.[35]

The third phase of narrative detection, which involves uncovering and tracing the evolution of plots, has yet to be explored for the Estonian language.

Lithuanian

In 2012, a report[36] published on the language technology support for European languages evaluated Lithuanian language as '*extremely under resourced and with very weak support*'. Utka et al., 2016[37] argued that the crucial reasons for the lack of progress of NLP tools for Lithuanian through re- search were: i) Global research communities were less interested as only three million speakers make Lithuanian commercially unattractive, ii) Lithuanian R&D mainly depended on national and European structural funds. However, a significant boost was noticed in NLP technologies for Lithuanian after 2012, due to the successful implementation of the national programme *The Lithuanian Language for Information Society (2012-2015)* and the involvement of international language technology communities and infrastructures such as METANET[38] and CLARIN ERIC.[39] Following this, the Government of Lithuania took a number of initiative programmes to boost NLP research for Lithuanian.

The purpose of this section is to evaluate the NLP capabilities of Lithuanian with respect to topic modelling and narrative detection. In the context of the Lithuanian language, the following subsections discuss the existing research and research gaps for handling topic modelling and narrative detection tasks.

Topic Modelling

In spite of the recent success stories of various NLP tasks employing transformer-based language models and neural networks, the proposed topic modelling approaches for Lithuanian are yet to leverage that. While surveying the studies carried out on topic modelling for Lithuanian language texts, we found only two topic modelling papers that were published in the recent past. However, neither of them employed any neural approaches. Rather, those studies used traditional ML approaches: LDA[40] and Correlated Topic Model.[41]

Mandravickaite et al., 2020[42] employed LDA as the topic modelling approach to identify trending topics in different media sources, which includes Lithuanian news portal (*delfi.lt*) and two alternative/unconventional media channels – namely, *sarmatas.lt* and *netiesa.lt*. In the pre-processing step, lemmatisation and lower-casing were achieved using the core Lithuanian models of SpaCy.[43] An open-source Python library[44], specifically developed for Lithuanian text processing, was used to remove stopwords, numbers, symbols, and punctuation marks.

Later, in 2021, Rabitz et al., 2021[45] employed the topic modelling approach for Lithuanian mass media discourse on climate change. The corpus consists of 583 Lithuanian news articles published on three popular news websites between 2017 and 2018. The authors

argued that the Correlated Topic Model (CTM) is more realistic than LDA because it assumes that the occurrences of different topics within a collection of documents may be correlated to each other. Moreover, in support of their statement, they cited Blei and Lafferty, 2007,[46] that confirmed the superiority of CTM over LDA. Hence, in this work, the CTM was chosen over LDA as the topic modelling approach.

Taking into account the aforementioned approaches in the previous paragraphs, the workflow for performing topic modelling could be as follows:

1. Data Collection: Acquire Lithuanian textual data intended for performing topic modelling.

2. Pre-processing: A standard pre-processing pipeline (which includes lowercasing, correction of character encoding and formatting issues, removal of numbers and common Lithuanian stopwords, and removal of punctuation) can be achieved following the work of Mandravickaite et al., 2020,[47] which employed Core Lithuanian models of SpaCy[48] and a dedicated open-source Python library[49] for Lithuanian NLP. Also, it can make use of the existing coreference resolution approaches.[50]

In topic modelling, often only noun and adjective terms are considered, which improves the outcome. Identifying those noun and adjective terms requires part-of-speech (POS) tagging tools. We observed that POS tagging tools are available for Lithuanian.[51] Hence, the models to be developed could leverage the existing POS tagging tools.

3. Topic modelling: Following the studies of Mandravickaite et al., 2020[52]; Rabitz et al., 2021[53], the LDA and CTM can be employed as topic modelling approaches for Lithuanian texts.

4. Evaluation & Analysis: The effectiveness of the employed topic modelling approaches can be evaluated by taking into account two main aspects: coherence and relevance. Coherence measures how well the words in a topic are related to each other, based on their semantic similarity or frequency. Relevance measures how well the topics capture the main themes or aspects of the documents, based on their importance or specificity. There are various metrics and tools to calculate coherence and relevance, such as C_V, U_Mass, topic coherence pipeline, etc.

The workflow presented above is based on traditional ML approaches. The textual contents are represented as a bag of words. The state-of-the-art BERTopic[54] can work as a multilingual model if embeddings for the specific language are provided to it as input. Considering this, topic modelling in Lithuanian can leverage this state-of-the-art topic modelling model. Embeddings can be generated using monolingual and multilingual language models (LMs) that support Lithuanian. We found the presence of both monolingual (namely, LitBERTa-uncased) as well as multilingual LMs (namely, mBERT, XLM-R, and LitLat BERT), which can be used for embedding generation for the Lithuanian texts.

Narrative Detection

Despite the gradual progress in language processing tasks in Lithuanian since 2012, we could not find any research work on narrative detection in Lithuanian. However, the guidelines described in Section 3 could give an ideal kick-start for narrative detection in Lithuanian.

1. Named Entity Recognition: Since 2012, an open-source freely available named entity recognition toolkit called TildeNER[55] has been available for Lithuanian NER tasks. TildeNER was released under the Apache 2.07 license and can be freely acquired through the toolkit for multi-level alignment and information extraction from comparable corpora, a

public deliverable of the ACCURAT[56] project. It supports basic along with some additional NE classes, namely organisation, person, location, date, time, and money. The F-measure effectiveness of the model is 0.65. In the recent past, a BERT-based multilingual model named LitLat BERT[57] was introduced. The transformer-based LLM was trained on Lithuanian, Latvian, and English corpora. Lithuanian corpora were composed of Lithuanian Wikipedia from 2018[58], the Lithuanian part of the Directorate-General for Translation (DGT) corpus,[59] and the LtTenTen14 corpus.[60] On the TildeNER dataset, the LitLat BERT achieved an 0.85 F-measure, which is a drastic improvement over the Lithuanian NER task. Last year, in the context of Baltic languages, Viksna and SKADINA, 2007[61] evaluated the NER in multilingual settings. For Lithuanian, they observed that on the EUR-LEX dataset,[62] the mBERT model performed better than other multilingual transformer models such as XLM-R and LitLat BERT.

2. Relationship extraction: This NLP task has not been explored with to its fullest potential. To the best of our knowledge, Vileiniškis et al., 201[63] presented an approach to semantic search over domain-specific Lithuanian web documents. In this information retrieval task, they leverage information extraction, more specifically, relation extraction to achieve their goal. They employed a rule-based approach that looks for specific lexico-semantic patterns, combining information from prior lexical, morphological and named entity annotations. The proposed ruleset works on political and economic events. Hence, the present status of NLP capabilities in Lithuanian is significantly behind the moderate approaches to relation extraction. Therefore, extensive research is needed to handle relation extraction in Lithuanian texts.

3. Plot discovery and story evolution: Like other Baltic languages, NLP researchers in the Lithuanian language have not explored these areas. Nevertheless, future research in these areas could leverage the existing monolingual (namely, LitBERTa-uncased) as well as multilingual LLMs (namely, mBERT, XLM-R, and LitLat BERT).

Latvian

The situation with NLP capabilities for Latvian is similar to Lithuanian. Latvia was put in the "little to no language technology support" category in the 2012 report about the availability of language resources for European languages.[64] Since then, more than a decade later, significant advancements have been achieved in developing language resources and tools, such as text and speech corpora and pre-trained language models.[65] One of the notable examples is NLP-PIPE—a publicly available pipeline for Latvian NLP.[66] However, there is still a need to develop more advanced capabilities to be able to perform tasks such as narrative detection in Latvian.

The necessity to develop such resources is also acknowledged by the Latvian government, with ongoing projects such as the Research on Modern Latvian Language and Development of Language Technology[67], which aims to advance and further develop language resources and tools for modern Latvian. The other notable initiative is the crowd-sourcing activity Balsu Talka, asking Latvian-speaking people to contribute their voices and assist in the creating a Latvian voice corpus.[68]

Topic Modelling

There are multiple case studies that leveraged LDA as a generative probabilistic model for extracting topics from Latvian corpora. Notable examples include topic analyses in Latvian legal documents.[69] The authors also used Hierarchical Dirichlet Process (HDP)[70] to automatically detect the number of topics. The other example is a recent paper that explores the methodology of LDA topic modelling for analysing a dataset of historical newspapers in Latvian.[71] This paper is a part of a series of case studies to explore text analysis techniques for the analysis of the collection of digitised historical newspapers of the National Library of Latvia. The potential workflow for performing topic modelling in Latvian is the same as described in the corresponding sections about Lithuanian and Estonian. The useful tool would be the previously mentioned pipeline for Latvian natural language processing, NLP-PIPE. It also includes separate NLP components—for tokenisation, morphological analysis, dependency parsing, and NER.[72] NLP-PIPE source code is available on GitHub[73] and as a web-based demo.[74] It is worth mentioning that its capability for NER is being integrated into the Latvian literature platform[75] allowing the automatic identification of person names, place names, and events mentioned in texts (Branco et al., 2023).[76]

The steps to perform topic modelling might look like this:

1. Data Collection: Gather the textual data in Latvian on which you intend to perform topic modelling.

2. Pre-processing with NLP-PIPE: NLP-PIPE contains multiple modules useful for pre-processing text in Latvian:

- **A Tokenisation:** splitting the text into sentences and words.
- **B Morphological analysis:** NLP-PIPE can provide details like part of speech, case, gender, number, etc.
- **C Named Entity Recognition:** This step involves identifying and classifying key information in the text into predefined categories, such as names of people, places, organisations, etc. NLP-PIPE's capabilities in NER can be utilised here. It should be considered whether to include detected entities in the further analysis. NER for Latvian will be discussed further in the section.

The pre-processing step also usually includes text cleaning and normalisation (lowercasing, removing punctuation, numbers, etc.). Before performing topic modelling, it is usually beneficial to remove stopwords as well. There are collections of stopwords available, for example, on GitHub[77], or as libraries or packages for different programming languages, such as *stopwords* package for R.[78]

3. Topic Modelling: Topic Modelling capabilities are not included in NLP-PIPE, but there are plenty of packages available for different programming languages. The topic modelling methods previously used for Latvian include LDA and HDP (Baklāne and Saulespurēns, 2022[79], Viksna et al., 2020[80]), but there are other more complicated methods described in research, but not so widely used, as well.

4. Evaluation & Analysis: Both qualitative and quantitative methods are useful in the evaluation and analysis process. The generated topics should reflect the main themes of the text corpus and be interpretable. For the quantitative analysis, coherence metrics can be of use—a high coherence score means that the words in the topic are indeed related and make sense together, indicating a well-defined and meaningful topic.

The workflow presented is commonly used in text analysis and research, but there are other AI tools developed for Latvian that can potentially enhance the analysis.

With the rise of transformer-based language models, the capability of BERT models for NER was evaluated as well. Multilingual BERT models are widely used, but it was shown that a BERT model pre-trained on large Latvian corpora, achieves slightly better results, achieving a marginally better performance than multi-lingual BERT models.[81] For their experiments, authors used only two datasets available—proprietary TildeNER and a publicly available dataset which is a part (annotation layer) of FullStack-LV, a multilayer text corpus of Latvian.[82] There is also a publicly available LVBERT pretrained language model[83] that also achieves better performance than multilingual BERT models.[84]

It is also worth mentioning that TildeNER toolkit for NER recognition in Latvian and Lithuanian was developed more than a decade ago.[85] However, it is still available for download as a public deliverable of the ACCURAT project.[86]

Narrative Detection

Unfortunately, no literature regarding automatic narrative extraction was found. However, the previously described workflow can be used to provide the basis for initial research.

1. Named Entity Recognition: The available tools for NER in Latvian are TildeNER, NLP-PIPE and both multilingual BERT models and the LVBERT model trained specifically for Latvian.

2. Relationship Extraction: To the best of our knowledge, relationship ex- traction for Latvian was not explored at all. Hence, further research is needed to improve NLP capabilities for Latvian.

3. Plot discovery and story evolution: This area was not explored for Latvian as well; however, Latvian monolingual and multilingual language models and large language models can prove useful for the plot discovery tasks. However, as with the other Baltic languages, further research is needed.

Although these resources are not connected to the topic modelling and narrative detection directly, it is worth mentioning other resources that may be useful for NLP in Latvian:

1. Tēzaurs.lv: Tēzaurs.lv[87] is an extensive explanatory and synonym dictionary (Grasmanis et al., 2023[88]). The dataset is available as open data[89], hence it can potentially be used to enhance both topic modelling and narrative detection.

2. Latvian WordNet: Tēzaurs.lv is integrated with Latvian WordNet [90], which links words into synonym sets and other semantic relations (Paikens et al., [91]), which is an important aspect in both topic modelling and narrative detection. Latvian WordNet coverage was also expanded by linking it to the English Princeton WordNet (Strankale and Stāde, 2022[92]).

Present Scenario in the Baltics

This section upholds the capabilities of Baltic language technology as a whole with respect to the basic language processing, along with the topic of this study, i.e., topic modelling and narrative detection.

Pre-processing: In language technology, a standard pre-processing pipeline is necessary to prepare the textual content for any tasks. We found that, even if they have limited capabilities for all the language technology tasks, a well-established pre-processing pipeline is available for all the Baltic languages. For example, EstNLTK for Estonian, ltlangpack for Lithuanian and NLP-PIPE for Latvian.

Topic modelling: Despite the realm of neural approaches since the last decade, for all the Baltic languages, we could not find any neural approach to address the topic modelling

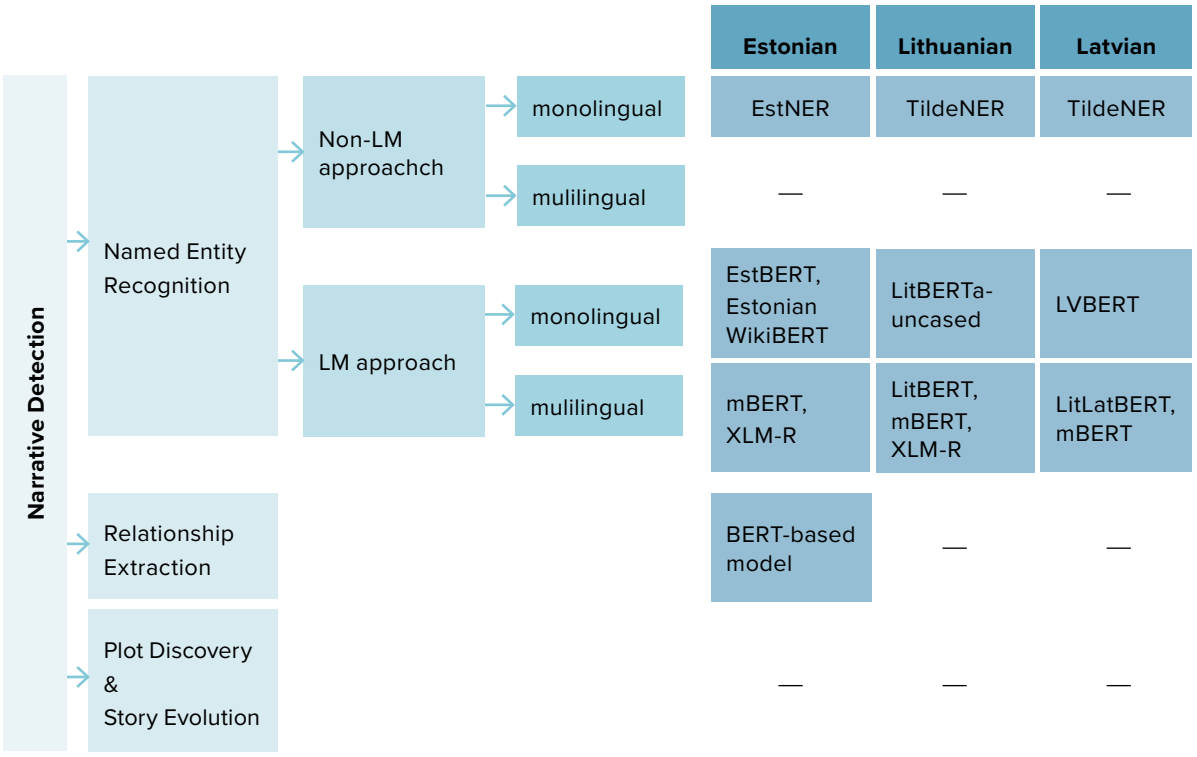

Figure 1: Present Scenario for Narrative Detection in the Baltics

task. Efforts have been made to apply topic modelling to all Baltic languages, utilising traditional machine learning methods such as LDA, Correlated Topic Model, and HDP, among others. Traditional probabilistic ML approaches were used in both Lithuanian and Latvian. Hence, the topic modelling task for Baltic languages is still in the nascent stage, and it requires serious attention to employ state-of-the-art LM-based approaches. As a future direction with respect to the topic modelling task, all the Baltic languages may leverage the state-of-the- art BERTopic[93] that can act as a multilingual model. This can be achieved by generating Baltic language-specific embeddings by employing either a Baltic language-specific monolingual LM or an LM of type multilingual, which supports the specific Baltic language.

Narrative Detection: The exploration of narrative detection in Baltic languages remains uncharted. Following the procedure outlined in Section 3 for narrative detection in these languages, their current ability to undertake this task is illustrated in Figure 1. This figure reveals the availability of conventional and sophisticated LM-based methods for the NER task. For the Baltic NER task, there are transformer-based monolingual models (such as EstBERT and LVBERT) and multilingual models (like LitLat-BERT and mBERT). Therefore, in the realm of narrative detection, it is apparent that NER can be effectively addressed given the current capabilities of Baltic languages.

Nevertheless, the subsequent two stages in the narrative detection process face significant limitations. The task of relationship extraction has been explored only in Estonian, and the resulting model is not publicly accessible due to its development within a military project. Regarding the discovery of plots and story evolution, no current models exist to address this aspect. Consequently, substantial development is required for these final two workflow phases in the context of Baltic languages.

Large Language Model annotation exercise

In this section, we explore the annotation exercise conducted to assess the effectiveness of GPT-4, a state-of-the-art Large Language Model (LLM), in performing narrative detection for the Estonian language.

Annotation procedure

GPT-4 is an advanced artificial intelligence language model developed by OpenAI. It represents a significant leap in AI capabilities, featuring the transformer neural network architecture. It was trained on a diverse range of internet text. This training allows GPT-4 to generate human-like text, understand context, and handle various language-based tasks with remarkable proficiency.

While it excels in applications ranging from creative writing to business automation, GPT-4 has limitations, such as potential biases and inaccuracies reflective of its training data. The LMSYS Chatbot Arena, hosted on HuggingFace[94], assesses LLMs such as chatbots. It involves human users interacting with two undisclosed models and then voting for their preference. Distinctive in its application of the Elo rating system used for ranking chess players, this platform effectively ranks a broad array of models, supporting gradual evaluation. Incorporating open-source and proprietary models, including GPT-4, the Arena sheds light on their practical performance capabilities. The GPT-4 LLM family consistently maintains the top position in the platform's rankings.

The annotation task entails selecting 20 news articles in Estonian[95], focusing on a specific situation in the Baltic countries: the unauthorised mass migration from Belarus to Latvia. The unauthorised mass migration from Belarus to Latvia, involving individuals from the Middle East and Africa, has led to a crisis at the Latvian border. This influx, influenced by political and economic turmoil in the migrants' countries and Belarus's relaxed border policies, has heightened security in Latvia and caused diplomatic strains. The situation presents a multifaceted challenge, balancing humanitarian needs with national security. Two human annotators were assigned to annotate these articles using the narrative detection layered method described in Section 3. The annotators are two native speakers of Estonian who possess prior experience in NLP annotation and a solid grasp of linguistics. The annotation process involved the following stages:

Collaborative Annotation of an Article. The annotators reviewed the relevant section from this article related to the narrative detection task. They independently annotated an article employing the layered approach outlined in the previously mentioned section. This article was translated from Estonian to English, and one of the study's authors also performed annotations. Subsequently, all three individuals agreed on a unified version of the annotation and established basic guidelines for annotating the remaining articles in the pool.

Individual Annotation of Articles. Following the preliminary guidelines established in the initial step, each annotator independently annotated all the articles.

Consensus on Human Annotation. Each annotator reviewed the annotations made by their colleague. A final, mutually agreed-upon human annotation was then created by reconciling discrepancies through discussion.

Example 5.1 below displays a segment of the annotation created by the annotators for one of the news articles.[96] It includes only partial representations of the first two layers, while the final two are fully shown.

Named Entity Recognition (NER):

- Location: Latvian-Belarusian border
- Location: Latvian-Russian border
- Actor: Latvian border guards
- Actor: Belarusian authorities

Relationship Extraction (RE):

- Detainment: (Latvian border guards, African and Asian immigrants)
- Support: (Belarusian authorities, African and Asian immigrants)

Plot Discovery:

- Exposition: [Stabilised number of illegal immigrants] Event at the [Latvian-Belarusian border] Location.
- Rising Action: [Installation of wire fences] Event along the [Latvian-Belarusian border] Location.
- Climax: Detention of [144 illegal immigrants] Event at the [Latvian-Belarusian border] Location.
- Falling Action: Discussion of [enhanced border security] Event and [military needs] by [Latvian government] Actor
- Resolution: Plans for [complete border fencing] Event and [electronic monitoring device installation].

Story Evolution:

- Shift Detection: Shift from [border control] Event to broader geopolitical issues and strategic military considerations for [border defence] Event.
- Causal Relation Extraction: [Support from Belarusian authorities] cause leads to [increased illegal crossings] effect; [Increased illegal crossings] cause leads to [enhanced border security measures] effect by [Latvian government] Actor. [enhanced border security measures] cause lead to [a week-long queue in a legal border crossing point] effect.

We have been experimenting with various templates to automate narrative detection. A template is a predefined and structured format used to guide the creation of prompts or shape responses. It serves as a blueprint with placeholders or specific formatting rules, ensuring consistency and standardisation in the input given to or output received from an LLM. The best template choice is presented in Template 5.1 below.

template = Perform narrative detection on an article you receive as input. The text is in a language (language of the article) you will also receive as input. The text is about illegal immigrants on the Latvian-Belarusian border who wish to reach Western Europe. The narrative detection should have the following layers:

Named Entity Recognition (NER): Identify entities like actors, events, and timelines that significantly extract semantic insights from narratives. The names of the entities' tags should be in English. The entities are in the language of the article. (Example for an arbitrary named entity recognition operation: Location [Lati-Valgevene piir], etc)

Relationship Extraction (RE): Extract binary semantic relationships between the entities identified in the first layer. The name of the relations should be in English. The entities linked by the relations should be in the language of the article (Example for an arbitrary

relationship: is detained [Lati-Valgevene piiril, 100 inimest], etc)

Plot Discovery: Identify the main components of the plot using Freytag's Pyramid with the following stages: Exposition, Rising Action, Climax, Falling Action, and Resolution. The entities in the components of the plot should be identified in the first layer whenever possible. The names of the stages should be in English. The content of the stages should be in the language of the article. Example for the first stage: Exposition: Läti-Valgevene piiril tegeletakse suure hulga ebaseaduslike immigrantidega, päevas peetakse kinni umbes 100 inimest, etc]

Story Evolution: This layer should have two components: Shift Detection, which recognises significant changes or deviations in the narrative, and Causal Relation Extraction, which unveils the reasons behind every significant event or decision in the story. Whenever possible, the entities in the components of the story should be identified in the first layer. (Example for the first component: Shift Detection: [Narratiivis toimub oluline muutus, kui peaminister lubab paigaldada jälgimisseadmed kogu idapiirile.]

Language of the article: Estonian article: {text}

The template is designed for narrative detection with an LLM. It outlines a structured approach to analyse an article's content, presented in Estonian, focusing on illegal immigration at the Latvian-Belarusian border. The template details the discussed analysis layers: NER for identifying key entities like actors and events, RE for uncovering relationships between these entities, Plot Discovery using Freytag's Pyramid to outline the story's structure, and Story Evolution for detecting narrative shifts and causal relations. For each layer, the template specifies the output language for entity tags and relation names (English), while the content remains in the article's language (Estonian). It also includes examples for each layer. The template's output is formatted to mirror the annotators' document annotations, facilitating an easy comparison.

Evaluation

Two types of evaluations were conducted on the narrative detection process: quantitative and qualitative. For the quantitative evaluation, we counted the number of annotations a human and the LLM agreed on. The qualitative evaluation looked at how effectively and accurately the system could understand and analyse stories. This was not just about counting data but also seeing how well the system grasped the heart of the story, its flow, and how all the parts fit into the context.

Table 1 presents the quantitative outcomes for the NER layer. In this table, the first column lists the types of entities. The following columns detail the count of entities annotated by our annotators and by the LLM for each entity type. The last column lists the number of common entities annotated by humans

Entities	Human	LLM	Common
Location	**92**	81	60
Actor	**184**	119	87
Event	**138**	70	51
Timeline	69	**71**	45
Other	10	**19**	1

Table 1: Quantitative Evaluation of the NER layer (Highest value for each entity in bold)

and the machine. Overall, human annotators identified more entities than the LLM across each category. However, considering that the LLM operated in an unsupervised manner, its performance is very good.

In the second layer of Relationship Extraction, human annotators identified 160 relationships, while the LLM annotated 134. There were 61 relationships commonly identified by both, with human annotators annotating an additional 99 relationships that the LLM did not. Conversely, the LLM identified 73 semantic relationships that the human annotators did not annotate. The performance at the second level is not as high as at the first. This is primarily because the LLM tends to annotate additional relations that, while present in the text, are not directly relevant to the narrative's theme. Additionally, human annotators often identify abstract relations not explicitly stated in the text.

The outcomes for the elements of plot discovery identified by the LLM are as follows: 10 instances of Exposition (out of 19), 6 instances of Rising Action, 4 Climaxes, 7 Following Actions, and 9 Resolutions. Concerning the story revolution, the annotators noted 21 transitions, while the LLM recognised 19. Out of these, 10 shifts were identified by the human annotators and the LLM. The statistics for causal relations are as follows: annotators identified 48 causal relations, while the LLM identified 26, with 14 of these relations being common between both.

The qualitative evaluation provides specific insights for each layer. The observations for the NER layer are as follows.

There were occasions where the system did not recognise 'immigrants' as Actors. However, immigrants were referenced in the Events category, and they were included as entities in the Relationship extraction.

The system sometimes inaccurately tagged geopolitical entities as Locations. Notably, there was an instance of dual tagging where NATO was marked both as a Location and as an Organisation (Actor).

Durative adverbs, such as "kahe aasta jooksul" (translated as 'during two years'), were incorrectly identified by the system as Timelines.

Therefore, while the NER layer demonstrates robust annotation, there are some notable inconsistencies within this layer and its integration with the Relationship Extraction layer.

The following observations have been made concerning the Relationship Extraction layer:

There were instances where the LLM overlooked certain relationships identified by human annotators. An example is when the LLM failed to report the delay in border fence construction. Overall, the LLM struggles to deduce relationships that are not explicitly mentioned in the text of the article.

The LLM pinpointed certain relationships that human annotators overlooked, especially in cases involving individual statements. However, many of these identified relationships are not pertinent to the main theme of the narrative.

A notable tactic employed by the LLM involved forming relationships from noun phrases that included premodifiers.

The LLM often recast Events as Relationships. Typically, the names of these relationships were derived from finite verbs in English, with the entities acting as verb arguments.

In examining the Plot Discovery layer, distinct variations were observed in the effectiveness of plot detection. It appeared more straightforward to identify Expositions than Climaxes, as evidenced by the higher number of matches in Expositions. However, it is somewhat paradoxical that there were relatively few matches in the Climax category. This

discrepancy might be attributed to the articles' lack of a clear narrative structure. Excessive information, or 'noise,' in the articles could have complicated the detection process. Of note, a significantly higher match rate was observed in a much shorter article.

Upon comparing the annotations, it was evident that the LLM selections were superior in some instances, while in others, the reverse was true. The GPT-4 results were particularly peculiar in scenarios where the Rising Action appeared to be placed after the Climax or the Resolution seemed to occur before the Rising Action.

In the Story Evolution layer, humans identified various shifts, such as thematic changes or state alterations. In contrast, the LLM focused on pinpointing the locations of these shifts within the article. As a result, the findings are difficult to compare. However, in some instances, it was apparent that the shifts detected by the LLM led to changes identified by the human annotators. Significantly, several shifts identified by the machine corresponded with Rising Actions or Climaxes in Plot Discovery.

Regarding causal relations, the machine's interpretations were generally accurate. The machine's identified causal relations primarily pertained to the main storyline. In contrast, human annotators identified a broader range of relations, some more specific and detailed.

Conclusion

This report has examined the tools and datasets predominantly accessible in academic settings for conducting topic modelling and narrative detection in Baltic languages. Traditional topic modelling is well-supported by existing tools for all Baltic languages. When opting for neural topic modelling, its largely language- independent nature might prove advantageous for some applications.

In narrative detection, it is apparent that this task remains largely unexplored for Baltic languages. However, our layered approach reveals that NER is sufficiently supported across these languages, while Relationship Extraction shows limited support. This discrepancy underscores a gap in the current NLP capabilities in the Baltic languages, highlighting the need for further research and development in Relationship Extraction methodologies.

We have also evaluated the capabilities of GPT-4, one of the most advanced large language models, in narrative detection for Estonian news articles. The findings indicate that GPT-4 demonstrates reasonable effectiveness in NER. However, its performance in Relationship Extraction is moderate, as many relationships identified by the model are not central to the main narrative. Similarly, Plot Discovery and Story Evolution show moderate effectiveness. This observation suggests that while GPT-4 possesses a robust capacity for identifying entities and their interactions, its ability to discern the narrative's core thematic elements and plot progression requires refinement.

The qualitative and quantitative evaluations conducted provide a nuanced understanding of GPT-4's current limitations and strengths in narrative detection. Specifically, the model's performance in accurately capturing the essence of complex narratives and its occasional misinterpretation of entity relationships point towards the necessity for ongoing adjustments and training with more targeted datasets.

As a result, while GPT-4 can assist human annotators in narrative detection, its performance level is not yet optimal for standalone use. It could be feasible to deploy GPT-4 with minimal supervision in contexts where preliminary narrative analysis is beneficial, but a comprehensive understanding and interpretation of narratives would still rely significantly on human expertise. Assessing its effectiveness in such a setup was outside the scope of this report, but it presents an intriguing avenue for future research. This exploration could potentially lead to the development of more sophisticated models that are capable of handling the intricacies of narrative detection in less commonly studied languages, thereby broadening the scope of NLP applications and making them more inclusive.

Endnotes

1 European Language Grid. live.europe-an-language-grid.eu

2 Greene, D., O'Sullivan, J., & O'Reilly, D. (2024). *Topic modelling literary interviews from The Paris Review*. Digital Scholarship in the Humanities. Advance online publication. doi.org/10.1093/llc/fqad098

3 AJP-10: Allied Joint Doctrine for Strategic Communications.

4 Bergmanis-Korāts, G., Bertolin, G., Pužule, A., & Zeng, Y. (2024). *AI in Support of StratCom Capabilities*. NATO Strategic Communications Centre of Excellence. Retrieved from stratcomcoe.org/publications/ai-in-support-of-stratcom-capabilities/296

5 Ibid.

6 As of 4 Apr 2024.

7 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ..., & Amodei, D. (2020b). Language models are few-shot learners. *CoRR*, abs/2005.14165. Retrieved from arxiv.org/abs/2005.14165

8 Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. doi.org/10.1162/jmlr.2003.3.4-5.993

9 Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian non-parametric inference of topic hierarchies. *Journal of the ACM*, 57(2), 7:1–7:30. org/10.1145/1667053.1667056

10 Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17–35.

11 Miao, Y., Yu, L., & Blunsom, P. (2015). Neural variational inference for text processing. *CoRR, abs/1511.06038*. Retrieved from arxiv.org/abs/1511.06038

12 "The Middle East Region Is Quieter Today Than It Has Been in Two Decades." (*The Atlantic*). (2023, October 7). Retrieved from www.theatlantic.com/international/archive/2023/10/israel-war-middle-east-jake-sullivan/675580

13 Ranade, P., Dey, S., Joshi, A., & Finin, T. (2022). Computational understanding of narratives: A survey. *IEEE Access, 10*, 101575–101594. doi.org/10.1109/ACCESS.2022.3205314

14 Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers*) (pp. 4171–4186). Association for Computational Linguistics. doi.org/10.18653/V1/N19-1423

15 Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: deep contextualized entity representations with entity-aware self-attention. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020* (pp. 6442–6454). Association for Computational Linguistics. doi.org/10.18653/V1/2020.EMNLP-MAIN.523

16 Chambers, N., & Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In K. R. McKeown, J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), ACL 2008, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20,*

2008, Columbus, Ohio, USA (pp. 789–797). The Association for Computational Linguistics. aclanthology.org/P08-1090

17 Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992* (pp. 539–545). Retrieved from aclanthology.org/C92-2082

18 U.S. Department of State. (2022, January 20). Russia's Top Five Persistent Disinformation Narratives. Retrieved from www.state.gov/russias-top-five-persistent-disinformation-narratives/

19 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR, abs/1907.11692*. Retrieved from arxiv.org/abs/1907.11692

20 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877-1901.

21 Zhang, Y., Qi, P., & Manning, C. D. (2018). Graph convolution over pruned dependency trees improves relation extraction. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31-November 4, 2018* (pp. 2205–2215). Association for Computational Linguistics. doi.org/10.18653/V1/D18-1244

22 Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In K. Su, J. Su, & J. Wiebe (Eds.), *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore* (pp. 1003–1011). The Association for Computer Linguistics. aclanthology.org/P09-1113

23 Egan, K. (1978). What is a plot. *New Literary History, 9*, 455. api.semanticscholar.org/CorpusID:147002098
Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J. L., & III, H. D. (2016). Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (pp. 1534–1544). The Association for Computational Linguistics. doi.org/10.18653/V1/N16-1180

24 Mani, I. (2012). *Computational modelling of narrative*. Morgan & Claypool Publishers. doi.org/10.2200/S00459ED1V01Y201212HLT018

25 Laur, S., Orasmaa, S., Särg, D., & Tammo, P. (2020). Estnltk 1.6: Remastered Estonian NLP pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 7154–7162). Retrieved from www.aclweb.org/anthology/2020.lrec-1.884

26 Barbu, E., Orav, H., & Vare, K. (2018). Topic interpretation using WordNet. In K. Muischnek & K. Müürisep (Eds.), *Human Language Technologies – The Baltic Perspective - Proceedings of the Eighth International Conference Baltic HLT 2018*, Tartu, Estonia, 27-29 September 2018 (Vol. 307, pp. 9–17). IOS Press. doi.org/10.3233/978-1-61499-912-6-9

27 Lingsoft. (n.d.). Microservices at your service - bridging the gap between NLP research and industry. Retrieved from www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry

28 Industry here refers to sectors involved in the development, application, or utilization of NLP and language technology services.

29 TartuNLP. (2024). Tartu NLP Microservices. Retrieved March 18, 2024, from tartunlp.ai/mikroteenuste-projekt

30 Sirts, K. (2023, May). Estonian named entity recognition: New datasets and models. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 752–761). University of Tartu Library. aclanthology.org/2023.nodalida-1.76

31 Orasmaa, S. (2015). Event coreference detection in Estonian news articles. *Eesti Rakenduslingvistika Uhingu Aastaraamat, (11)*, 189–203.

32 Orasmaa, S. (2014, May). Towards an integration of syntactic and temporal annotations in Estonian. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1259–1266). European Language Resources Association (ELRA). www.lrec-conf.org/proceedings/lrec2014/pdf/530 Paper.pdf

33 Tanvir, H., Kittask, C., Eiche, S., & Sirts, K. (2021, May). EstBERT: A pre-trained language-specific BERT for Estonian. In S. Dobnik & L. Øvrelid (Eds.), *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 11–19). Linköping University Electronic Press, Sweden. aclanthology.org/2021.nodalida-main.2

34 Barbu, E., Muischnek, K., & Freienthal, L. (2020). A study in Estonian pronominal coreference resolution. In A. Utka, J. Vaicenoniene, J. Kovalevskaite, & D. Kalinauskaite (Eds.), *Human Language Technologies - The Baltic Perspective – Proceedings of the Ninth International Conference Baltic HLT 2020*, Kaunas, Lithuania, September 22-23, 2020 (Vol. 328, pp. 3–10). IOS Press. doi.org/10.3233/FAIA200595

35 github.com/SoimulPatriei/EstonianCoreferenceSystem

36 Vaišnienė, D., Zabarskaitė, J., Rehm, G., & Uszkoreit, H. (2012). The Lithuanian language in the digital age (Vol. 385). META-NET White Paper Series "Europe's Languages in the Digital Age". G. Rehm & H. Uszkoreit (Eds.). Springer.

37 Utka, A., Amilevičius, D., Krilavičius, T., & Vitkutė-Adžgauskienė, D. (2016). Overview of the development of language resources and technologies in Lithuania (2012–2015). *In Human Language Technologies – The Baltic Perspective* (pp. 12–19). IOS Press.

38 META-SHARE. (n.d.). Retrieved from www.meta-share.eu

39 CLARIN. (n.d.). Retrieved from clarin.eu

40 Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3, 993–1022. doi.org/10.1162/jmlr.2003.3.4-5.993

41 Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

42 Mandravickaite, J., Briediene, M., Uus, J., & Krilavicius, T. (2020). What's in the news? Identification of trending topics in alternative and mainstream Lithuanian media. *TWSDetection, 3–17.*

43 spaCy. (n.d.). Retrieved from spacy.io

44 Retrieved from github.com/tokenmill/ltlangpack

45 Rabitz, F., Telešienė, A., & Zolubienė, E. (2021). Topic modeling the news media representation of climate change. *Environmental Sociology*, 7(3), 214–224.

46 Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17–35.

47 Mandravickaite, J., Briediene, M., Uus, J., & Krilavicius, T. (2020). What's in the news? Identification of trending topics in

alternative and mainstream Lithuanian media. TWSDetection, 3–17.

48 spaCy. (n.d.). Retrieved from spacy.io

49 Retrieved from github.com/tokenmill/ltlangpack

50 Zitkus, V., Butkienė, R., Butleris, R., Maskeliūnas, R., Damaševičius, R., Woźniak, M., et al. (2019). Minimalistic approach to coreference resolution in Lithuanian medical records. *Computational and Mathematical Methods in Medicine, 2019.*; Zitkus, V., & Nemuraitė, L. (2015). First steps in automatic anaphora resolution in Lithuanian language based on morphological annotations and named entity recognition. *Information and Software Technologies: 21st International Conference, ICIST 2015, Druskininkai, Lithuania, October 15-16, 2015, Proceedings 21*, 480–490.

51 Vitkutė-Adžgauskienė, D., Utka, A., Amilevičius, D., & Krilavičius, T. (2016). NLP infrastructure for the Lithuanian language. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2539–2542.

52 Mandravickaite, J., Briediene, M., Uus, J., & Krilavicius, T. (2020). What's in the news? Identification of trending topics in alternative and mainstream Lithuanian media. *TWSDetection, 3–17.*

53 Rabitz, F., Telešienė, A., & Zolubienė, E. (2021). Topic modeling the news media representation of climate change. *Environmental Sociology*, 7(3), 214–224.

54 Grootendorst, M. (2022). Bertopic: Neural topic modelling with a class-based TF-IDF procedure. arXiv preprint *arXiv:2203.05794*.

55 Pinnis, M. (2012). Latvian and Lithuanian named entity recognition with TildeNER. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1258–1265).

56 ACCURAT Project. (n.d.). ACCURAT Toolkit. Retrieved March 18, 2024, from www.accurat-project.eu/index.php?p=toolkit

57 Ulčar, M., & Robnik-Sikonja, M. (2021). Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 162–172).

58 LINDAT/CLARIAH-CZ. (2018, February 25). Plaintext Wikipedia dump 2018 [Data file]. hdl.handle.net/11234/1-2735

59 LINDAT/CLARIAH-CZ. (2018, August 15). JRC EU DGT Translation Memory Parsebank DGT-UD 1.0 [Data file]. hdl.handle.net/11356/1197

60 Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL* (pp. 125–127).

61 Viksna, R., & Skadina, I. (2007). Multilingual transformers for named entity recognition. *The Annals of Applied Statistics, 1*(1), 17–35.

62 Regulation (EU) 2016/679 of the European Parliament and of the Council, Official Journal of the European Union L 119 (2016).

63 Vileniškis, T., Sukys, A., & Butkienė, R. (2015). An approach for semantic search over Lithuanian news website corpus. *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, 1, 57–66.

64 Skadina, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I., & Rudzīte, A. (2012). Latviešu valoda digitālajā laikmetā – The Latvian Language in the Digital Age [White paper]. Springer. Retrieved from www.meta-net.eu/whitepapers

65 Ibid.

66 Znotins, A., & Cirule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. Human Language Technologies – The Baltic Perspective, 307, 183–189. doi.org/10.3233/978-1-61499-912-6-183

67 Latvijas Zinātņu akadēmija. (n.d.). Mūsdienu latviešu valodas izpēte un valodas tehnoloģiju attīstība [Research on contemporary Latvian language and the development of language technologies]. Latvijas Zinātņu akadēmija. Retrieved from www.lzp.gov.lv/project/musdienu-latviesu-valodas-izpete-un-valodas-tehnologiju-attistiba

68 Balsu Talka. (n.d.). balsutalka.lv

69 Viksna, R., Kirikova, M., & Kiopa, D. (2020). Exploring the Use of Topic Analysis in Latvian Legal Documents. COUrT@CAiSE. api.semanticscholar.org/CorpusID:229357043

70 18HDP is also probabilistic model that can be seen as an extension or a more advanced version of LDA. The main advantage of HDP is that the number of topics is inferred automatically from the data, while in LDA there is a need to pre-specify the number of topics.

71 Baklāne, A., & Saulespurēns, V. (2022). The application of latent Dirichlet allocation for the analysis of Latvian historical newspapers: Oskars Kalpaks' case study. *Science, Technologies, Innovation, 1(21)*, 29–37. Publisher: State Scientific Institution– Ukrainian Institute of Scientific and Technical Expertise and Info. doi.org/10.35668/2520-6524-2022-1-05

72 Znotins, A., & Cirule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies – The Baltic Perspective*, 307, 183–189. doi.org/10.3233/978-1-61499-912-6-183

73 LUMII-AILab. (n.d.). NLP-PIPE. GitHub. github.com/LUMII-AILab/nlp-pipe

74 AILab. (n.d.). NLP Tools. nlp.ailab.lv

75 Latvijas literatūra. www.literatura.lv

76 Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., Jong, F. D., Kamock P., König, A., Lindén, K., Navarretta, C., et al. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. Language Resources and Evaluation, 1–32. Publisher: Springer.

77 github.com/stopwords-iso/stopwords-iso

78 cran.r-project.org/web/packages/stopwords/readme/README.html

79 Baklāne, A., & Saulespurēns, V. (2022). The application of latent Dirichlet allocation for the analysis of Latvian historical newspapers: Oskars Kalpaks' case study. Science, Technologies, Innovation, 1(21), 29–37. doi.org/10.35668/2520-6524-2022-1-05

80 Viksna, R., Kirikova, M., & Kiopa, D. (2020). Exploring the Use of Topic Analysis in Latvian Legal Documents. *COUrT@CAiSE*. Retrieved from api.semanticscholar.org/CorpusID:229357043

81 Ibid.

82 github.com/LUMII-AILab/FullStack/tree/master/NamedEntities

83 github.com/LUMII-AILab/LVBERT

84 Znotins, A., & Barzdins, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. Baltic HLT, 111–115.

85 Pinnis, M. (2012). Latvian and Lithuanian Named Entity Recognition with TildeNER. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), 1258–1265.

86 ACCURAT Project. Retrieved on 18 March 2024, www.accurat-project.eu

87 Tezaurs Dictionary. *Tezaurs*, Retrieved on 18 March 2024, tezaurs.lv

88 Grasmanis, M., Paikens, P., Pretkalnina, L., Rituma, L., Strankale, L., Znotins, A., & Gruzitis, N. (2023). Tezaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, 400–418. Retrieved from elex.link/elex2023/wp-content/uploads/89.pdf

89 Tēzaurs.lv 2023 (Autumn Edition). (2023, September 01) *LINDAT/CLARIAH-CZ*. Retrieved from repository.clarin.lv/repository/xmlui/handle/20.500.12574/92

90 Latvian WordNet, Artificial Intelligence Laboratory of the Institute of Mathematics and Computer Science, University of Latvia, Retrieved 18 March 2024, wordnet.ailab.lv

91 Paikens, P., Klints, A., Lokmane, I., Pretkalnina, L., Rituma, L., Stade, M., & Strankale, L. (2023). Latvian WordNet. *Global WordNet Conference*. Retrieved from api.semanticscholar.org/CorpusID:265034438

92 Strankale, L., & Stāde, M. (2022). Automatic Word Sense Mapping from Princeton WordNet to Latvian WordNet. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence* (pp. 478–485). doi.org/10.5220/0011006000003116

93 Grootendorst, M. (2022). Bertopic: Neural topic modelling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

94 "Chatbot Arena Leaderboard," Hugging Face Spaces, Retrieved on 18 March 2024, huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

95 One article was removed from the selection due to its near-identical content with another article in the pool.

96 The news article was translated into English.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org