UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Li Merila

Cross-Lingual Misinformation Detection: Aligning English and Estonian Fake Health News

Master's Thesis (15 ECTS)

Supervisors: Uku Kangur, MSc
Roshni Chakraborty, PhD

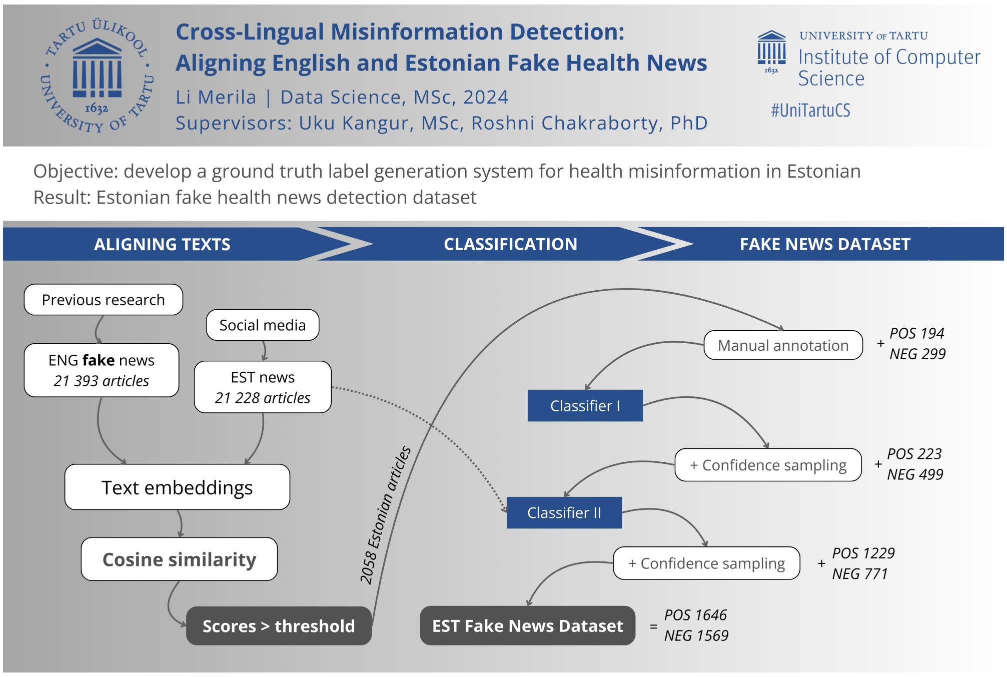
Tartu 2024

Cross-Lingual Misinformation Detection: Aligning English and Estonian Fake Health News

Abstract: Health misinformation poses a significant threat as it undermines trust in scientific expertise and reduces compliance with public health measures, ultimately decreasing community resilience against preventable diseases. This thesis focuses on identifying Estonian fake health news by leveraging a pre-labelled dataset in English. The primary objective is to develop a reliable system for generating ground truth labels for fake health news in Estonian, contributing to the field of fake news detection in low-resource settings. The proposed approach, namely Cross-Lingual Alignment and Confident Prediction Sampling (CAPS), employs a hybrid two-phase methodology involving semantic similarity measurements, manual annotation, classification, and confidence sampling to create a novel fake health news dataset in Estonian.

Keywords: Estonian Misinformation, Fake Health News, Confidence Sampling, Cross-Lingual Similarity Detection, Estonian Fake News Dataset

CERCS: P176 Artificial intelligence



Cross-Lingual Misinformation Detection:
Aligning English and Estonian Fake Health News
Li Merila | Data Science, MSc, 2024
Supervisors: Uku Kangur, MSc, Roshni Chakraborty, PhD

UNIVERSITY OF TARTU
Institute of Computer Science
#UniTartuCS

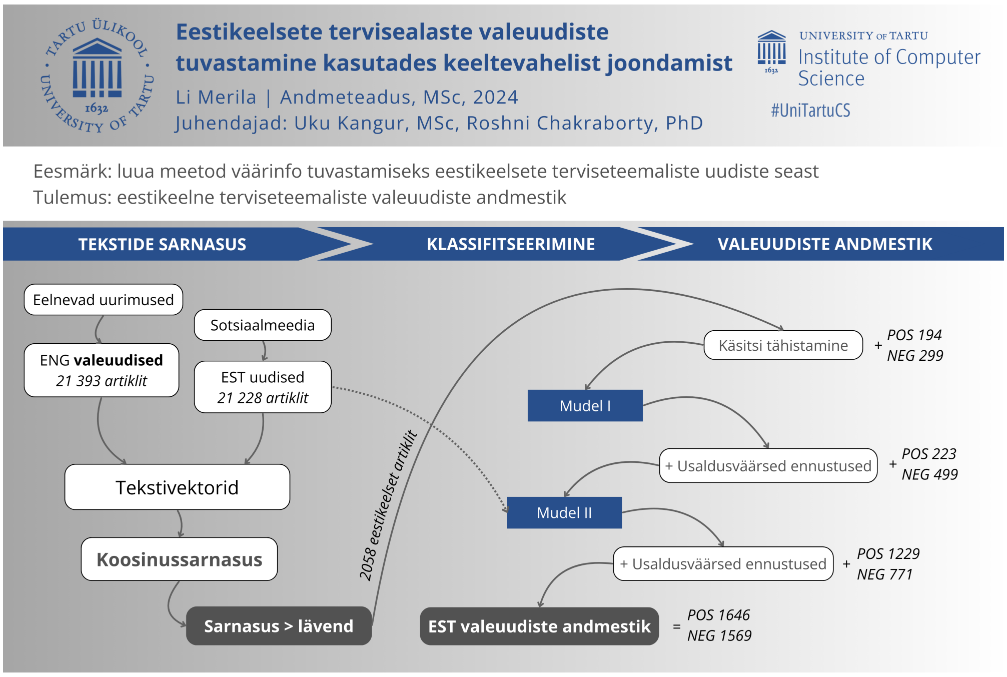
Objective: develop a ground truth label generation system for health misinformation in Estonian
Result: Estonian fake health news detection dataset

Eestikeelsete tervisealaste valeuudiste tuvastamine kasutades keeltevahelist joondamist

Lühikokkuvõte: Tervisealane väärinfo kujutab endast märkimisväärset ohtu, kuna see õõnestab usaldust teaduse vastu ja vähendab allumist riiklikele tervisemeetmetele, vähendades seeläbi ühiskonna vastupanuvõimet ennetatavatele haigustele. Käesolev magistritöö keskendub eesti keeles esinevate vale terviseuudiste tuvastamisele, kasutades selleks inglisekeelseid valeuudiste andmestikke. Põhieesmärk on välja töötada usaldusväärne süsteem eesti keeles terviseuudiste tõeväärtuse määramiseks, panustades seeläbi väheuuritud teadusvaldkonda. Loodud meetod, *Cross-Lingual Alignment and Confident Prediction Sampling* (CAPS), kasutab kahefaasilist hübriidmeetodit, mis hõlmab semantilise sarnasuse mõõtmist artiklite vahel, käsitsi märgendamist, klassifitseerimist masinõppe meetoditega ning usaldusväärsete ennustuste kogumist. Need tehnikad aitasid luua tavatu eestikeelse määratud tõeväärtusega terviseuudiste andmestiku.

Võtmesõnad: Eesti väärinfo, tervise valeuudis, usaldusväärne ennustamine, keeltevaheline sarnasuse joondamine, Eesti valeuudiste andmestik

CERCS: P176 Tehisintellekt



Eestikeelsete tervisealaste valeuudiste tuvastamine kasutades keeltevahelist joondamist

Li Merila | Andmeteadus, MSc, 2024

Juhendajad: Uku Kangur, MSc, Roshni Chakraborty, PhD

UNIVERSITY OF TARTU Institute of Computer Science

#UniTartuCS

Eesmärk: luua meetod väärinfo tuvastamiseks eestikeelsete terviseteemaliste uudiste seast

Tulemus: eestikeelne terviseteemaliste valeuudiste andmestik

Acknowledgments

I take this moment to express my gratitude to everyone who supported me through my master's journey.

I am deeply grateful to my supervisors, Uku Kangur and Roshni Chakraborty, for their supportive mentorship, fast responses, and innovative ideas. With a deep understanding of computational social science and research, they provided invaluable expertise and inspiration. Their relaxed approach and engaging teaching styles transformed complex research challenges into intriguing adventures and a vibrant learning experience.

I owe a heartfelt thank you to my partner, who motivated me daily with unwavering commitment and diligence towards academic excellence. Serving as my emotional anchor, they also brought their expertise in writing and languages to enrich my journey immeasurably. I cherish and am grateful for their support.

I want to thank my colleague, Kyrylo Medianovskyi, for his encouragement to explore diverse perspectives and creative solutions, which greatly enhanced my work. I appreciate how their sharp intellect, programming and data science knowledge significantly impacted my master's journey.

Lastly, a special nod to Caffeine, my steadfast companion through the long hours of study and writing. Their energizing presence was essential for getting me through my research and thesis. They have been my constant source of support throughout my academic life.

Contents

1 Introduction

In today's interconnected world, misinformation represents a pervasive threat that undermines informed decision-making and societal stability [1, 2]. Predominantly originating through social media, misinformation influences public health, elections, economic policies, and environmental strategies [3]. Social media platforms often prioritise engaging content over factual accuracy, leading to echo chambers and homogeneous communities [4]. This environment makes it difficult for individuals to distinguish authentic news from false reports [3]. To combat misinformation, a comprehensive strategy is essential, involving algorithmic refinements, stricter content moderation, and improved public education on media literacy [1]. The battle is not just technological but cultural, requiring a shift toward prioritising and verifying factual accuracy [4].

The COVID-19 pandemic has exacerbated the problem of misinformation, prompting the World Health Organization to describe this surge of health misinformation as an 'infodemic' [5]. This phenomenon characterises a parallel pandemic where misinformation spreads rapidly across digital platforms, with significant public health implications [6, 7, 8]. It has resulted in reduced compliance with public health measures and decreased vaccine uptake—critical factors for controlling disease outbreaks and achieving herd immunity. The spread of misinformation during this time has undermined trust in scientific expertise, negatively affecting public health initiatives and reducing community resilience against preventable diseases [9, 10].

Automatic detection of fake news is a formidable challenge in low-resource domains, such as Estonian health news. This is due to the complexity and length of the articles, which require sophisticated embeddings to capture the target language's semantics accurately. The task is further complicated by the absence of extensive, readily available datasets, necessitating the initiation of data creation through manual efforts. In high-resource languages like English, advanced detection mechanisms [11, 12, 13] have been developed leveraging extensive annotated datasets [14, 15], enabling more precise identification and mitigating health misinformation. However, such methodologies are not directly transferable to Estonian due to the lack of ground truth labels, making automatically detecting misinformation substantially more difficult.

The thesis aims to develop a robust method for generating ground truth labels for fake health news in Estonian, contributing to the broader field of misinformation detection in low-resource settings. This study utilizes multilingual modelling and practical annotation strategies to tackle misinformation effectively. The following points highlight the research contributions.

1. Proposed the Cross-Lingual Alignment and Confident Prediction Sampling (CAPS) methodology, which enhanced the quality of ground-truth labels and reduced annotation efforts. This approach combined cross-lingual alignment with confidence sampling, with further details and resources available at `https://github.com/goouthy/misinfo-eng-est/`.

2. Developed a methodology that used a pre-annotated English fake news dataset to filter potential fake health news in Estonian, establishing a precedent for cross-lingual information transfer. Manually annotated articles to create a gold standard dataset for Estonian health news and paired classification models with confidence sampling to effectively address health misinformation in a low-resource setting.

3. Achieved strong performance results from the Estonian fake health news dataset, with validation results showing an overall F1 score of 0.80 and accuracy of 0.81. Articles verified as genuinely health-related achieved higher performance, with an F1 score of 0.88 and an accuracy of 0.90.

4. Released a ground-truth dataset of 3,215 Estonian health-related articles, with 1,646 labelled as fake news and 1,569 as genuine. To the best of this thesis's knowledge, this dataset is the first for Estonian fake health news, serving as an essential resource for the research community and paving the way for future studies in misinformation detection.

The thesis is structured into key sections to provide an overview of the research conducted. Section 2 reviews techniques for detecting misinformation, focusing on low-resource settings. Section 3 details the collection and annotation of English and Estonian articles. Section 4 outlines the two-phase CAPS methodology for developing the Estonian fake health news dataset. The thesis concludes in Section 5, presenting and evaluating the results of the methodology.

2 Related works

This chapter provides an overview of prior research on the methods used to identify fake news and misleading information within monolingual frameworks as well as across multiple languages.

2.1 Dangers of Health Misinformation

Detecting misinformation is vital due to its significant impact on public perception and decision-making. Identifying falsehoods requires innovative techniques, as human judgment alone cannot navigate the intricate nature of deceptive content. Previous research lays the foundation for understanding how machine learning and natural language processing have emerged as essential tools in combatting the spread of misinformation. Exploring advancements in computational methods emphasizes the importance of harnessing cutting-edge technologies to tackle the challenges of misinformation detection and underscores the need for continued innovation in this field.

Identifying false claims from factual news articles is a prevalent topic, and in recent years, exaggerated by the COVID-19 pandemic, it has become a significant threat to public health. Unproven health beliefs and misconceptions have proliferated due to the massive spread of fake news flooding social media channels and the internet even before the pandemic's start. Prominent examples of such misconceptions include the MMS vaccine causing autism [16], and claims of Listerine as a cure for the common cold [17, 18]. However, these falsehoods have been magnified in the recent uncertain years [19].

Such misinformation aggravates the eroding trust in scientific expertise, undermining public health initiatives [6, 7, 8]. Misinformation related to vaccines, for instance, can reduce vaccination intent by approximately six percentage points among those initially willing to vaccinate, posing a direct threat to achieving herd immunity [9]. Moreover, projections indicate that, without intervention, anti-vaccination sentiments could dominate social media discourse within a decade [10].

COVID-19 misinformation has been linked to dangerous behaviours, including the consumption of harmful substances and an increased propensity for violence [20, 21]. A distressing incident in Iran, where false beliefs about curing COVID-19 with high-proof alcohol resulted in over 300 deaths due to methanol poisoning [22]. It is estimated that during the initial three months of 2020, over 800 lives worldwide were lost due to misinformation related to the coronavirus [19]. These examples underscore the lethal consequences of misinformation, and the World Health Organisation (WHO) has emphasized the danger posed by the infodemic, warning that it could lead to societal division and discord if not addressed [22].

Misinformation on the internet can be broadly categorized into fake news, rumours, and other forms, such as clickbait and social spam [23]. This thesis specifically focuses on fake news within the health domain. It adopts the definition of fake news provided by Allcott and Gentzkow [24] as 'news articles that are intentionally and verifiably false, and could mislead readers,' and health misinformation as defined by Chou et al. [25] as 'a health-related claim of fact that is currently false due to a lack of scientific evidence.'

2.2 Fake News Detection

Misinformation can significantly distort public perception, making it challenging to discern truth and falsehood. Research demonstrates that human judgment alone is often inadequate for identifying deceptive content, necessitating more advanced, automated detection methods. A meta-analysis by Bond and DePaulo [26] found that across 206 studies involving over 24,000 participants, individuals without additional resources could only correctly identify lies from truths 54% of the time—just 4% better than random guessing [27]. This finding highlights the limitations of relying solely on human judgment to discern fake news, emphasizing the critical need for automated detection methods to improve accuracy. Moreover, manual fact-checking approaches require significant human resources and are often limited in scale and efficiency [28].

The complexity of detecting fake news arises from the challenge of extracting features from natural language, which can involve elements like sarcasm and satire. This task becomes further complicated due to how misinformation manifests across social networks, hindering manual and automated detection efforts [29]. The lack of a comprehensive, theory-driven framework for detecting health-related fake news, particularly online, adds to these challenges [30]. This gap underscores the necessity for reliable features that machine learning techniques can utilise to identify misinformation effectively [30].

Researchers have developed datasets and models leveraging machine learning to counter these challenges to distinguish between fake and factual news [31]. Fake news detection can be tackled using natural language processing, machine learning, and deep learning approaches. These models can examine textual content, contextual details, source reliability, and other features to determine the veracity of news [19].

Detecting misinformation requires thoroughly analysing linguistic patterns and features rather than merely labelling news articles as fake or genuine. This field's groundwork has been laid through the manual selection of relevant linguistic features [19]. Based on these features, classification models have traditionally used statistical machine learning methods, with Support Vector Machine (SVM) recognised for its flexibility and efficacy across various datasets, including news articles and social media posts [32, 33, 34, 35]. Alongside SVM, models such as K-Nearest Neighbor, Naïve Bayes, Decision Trees, Random Forest, Gradient Boost, XGBoost, and Logistic Regression have also achieved high accuracy in detecting fake news [19], using textual features like n-grams, subjectivity, polarity markers, and keyword frequencies to distinguish false information [29, 30, 36].

A significant advancement in news article analysis has been adopting the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF refines model accuracy by emphasizing word frequency, allowing better separation between authentic and fabricated content. Its success has been particularly evident when used with Linear Support Vector Machines, which significantly improve classification outcomes [34, 32]. Various studies have supported the method's effectiveness in enhancing fake news detection precision, with research by Katsaros et al. [37] and Bojjireddy et al. [35] underscoring TF-IDF's effectiveness across multiple models and datasets.

Recently, a pivotal shift has been made toward exploring transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), which can harness deeper, contextualised information within news articles. BERT and its variants have shown superior performance in generating embeddings that enrich classification tasks, outperforming earlier models [11]. For instance, Samadi et al. [12] found that integrating BERT with Convolutional Neural Networks (CNN) signifies a promising approach for leveraging contextualised information embeddings in misinformation detection, achieving an impressive 91.4% accuracy in distinguishing fake news.

Pre-trained language models such as BERT have become mainstream in text classification [38] due to their transformer-based architecture [19]. Many studies have incorporated BERT and its variants into misinformation classification. A recent study by Alghamdi et al. [13] employed transfer learning using multilingual BERT to extract semantic knowledge, achieving 86% accuracy in a deep learning framework for multilingual fake news. Models like RoBERTa, evolved from BERT, have also been used to create semantic frameworks for further classification in detecting false information [39].

Numerous methodologies now utilise BERT embeddings for enhanced text classification [19]. Aggarwal et al. [40] highlight BERT's superior performance over traditional models like LSTM and gradient-boosted trees, even with minimal text preprocessing. In the context of fake news detection, BERT's efficacy is further evidenced in various studies [19, 41]. Veyseh et al. [42] used an ensemble model of CNN, LSTM, and BERT to achieve F1 scores between 0.976 and 0.981 by integrating content analysis with source credibility on a fake news dataset. The cased version of BERT also led to the highest reported accuracy of 98.41% in detecting COVID-19 misinformation [19, 42]. Further research by Alghamdi et al. [43] combined BERT embeddings with neural networks like CNN and BiGRU, resulting in a state-of-the-art F1 score of 0.98, confirming BERT's pivotal role in enhancing the precision of fake news detection frameworks during critical periods [19]. These insights underscore BERT's significant impact in advancing news article classification and detecting misinformation.

The field of misinformation detection employs a range of text-to-vector techniques, from traditional TF-IDF and n-grams to advanced transformer-based embeddings like those created by BERT. This array of methodologies, ranging from machine learning models to deep learning frameworks, provides a comprehensive toolkit for researchers and practitioners aiming to tackle the complex challenge of misinformation detection. The continuous evaluation and integration of these techniques reflect a commitment to improving the accuracy and reliability of detecting false information in the digital age.

2.2.1 Labelling in Low-Resource Settings

In the domain of health misinformation, Estonian is identified as a low-resource language due to the lack of an annotated dataset for fake news that could serve as a foundation for training machine learning models to detect falsehoods. While most research on automatic labelling fake news is conducted in English, this section examines various strategies for detecting fake news in non-English languages, exploring case studies from Amharic, Persian, Korean, Hindi, and other languages. The methodologies range from traditional machine learning approaches to more advanced deep learning and hybrid models.

For fake news detection in the Estonian language, no studies were found. However, a study [44] about fake news in Estonia examined the news site Telegram[1], which was found to play a significant role in disseminating misinformation and conspiracy theories, particularly during the COVID-19 pandemic. The study reveals that Telegram positions itself against mainstream media, spreading anti-government propaganda and encouraging defiance against the Estonian government. It is highlighted that fake news on Telegram is crafted to divert public loyalty away from established authorities.

In other low-resource settings, a foundational study by Abonizio et al. [45] created a dataset of news written in English, Brazilian Portuguese, and Spanish to identify language-independent features for fake news detection. Their analysis revealed that stylometric features, such as POS-tag diversity, the ratio of named entities to text size, quotation marks to text size, and out-of-vocabulary word frequency, significantly influenced prediction accuracy. The team improved detection rates by up to 85% by employing these features cross-lingually. This study underscores the potential of leveraging language-independent features to enhance the robustness of fake news detection across different languages.

Similarly, research by Gereme et al. [46] tackled Amharic fake news by creating a novel dataset and employing a CNN architecture, achieving 99% accuracy in identifying misinformation. This success highlights the effectiveness of deep learning models in processing and detecting fake news in less-resourced languages. Ghayoomi and Mousavian [39] made further advancements in Persian fake news detection when they used cross-lingual and cross-domain transfer learning with the XLM-RoBERTa model and CNN to address COVID-19-related fake news, achieving an accuracy of 94.46%.

[1] https://www.telegram.ee/

These studies collectively emphasise the importance of cross-lingual and transfer learning techniques in enhancing the accuracy of fake news detection models in various linguistic context articles.

Korean fake news detection saw improvements by integrating user engagement metrics with traditional textual analysis using advanced models such as BERT, ELECTRA, and RoBERTa. As demonstrated by Kang et al. [47], RoBERTa, in particular, delivered the best performance with a score of 0.709. This approach underscores the value of combining user interaction data with textual features to improve fake news detection.

Sharma and Arya [48] made significant progress in detecting fake news in Hindi by leveraging linguistic feature-based word embeddings designed to capture the unique nuances of the language. Their model, trained on a large corpus of manually annotated texts, achieved an impressive accuracy of 98.49%. This study highlights the potential of using language-specific features and large annotated datasets to enhance fake news detection accuracy in non-English languages.

Further, Chu et al. [49] explored the feasibility of detecting Chinese fake news by training a pre-trained language model on English misinformation datasets. Their research highlights the scarcity of fake news datasets in other languages, demonstrating that BERT achieved an F1 score of 0.69 when trained on English and tested on Chinese, and 0.79 when trained on Chinese and tested on English. This study underscores the potential for cross-lingual transfer learning in fake news detection.

De et al. [11] proposed a neural network trained on English and four languages under-resourced in misinformation detection—Hindi, Swahili, Indonesian, and Vietnamese—aiming to create a language-independent and domain-agnostic multilingual fake news classification model. Although their training dataset was limited, zero-shot experiments revealed that the model could identify fake news without seeing any examples in a particular language. This approach showcases the potential of multilingual models in addressing the challenges of fake news detection in low-resource settings.

A recent study by Alghamdi et al. [13] addressed the issue of transformer models' maximum sequence length and text truncation by implementing a hybrid summarization technique. They extracted only the most relevant content from texts, reducing data length while preserving crucial information. Their approach, when compared with multilingual BERT, XLM-RoBERTa, and semantic graph-based topic modelling, demonstrated better accuracy for most languages except English.

These methodologies and case studies illustrate diverse approaches to fake news detection across multiple linguistic settings. They showcase innovative uses of language-dependent and language-independent features to enhance the reliability of misinformation detection in low-resource environments.

3 Data

This section provides an overview of the data collection process for this thesis, which aims to generate ground truth labels for health-related fake news in Estonian. It discusses the procedures for source English datasets already annotated as fake news and the strategy for collecting unlabelled Estonian articles from health-related Facebook groups and news websites. The challenges faced during data collection for both languages are outlined, forming the foundation for the CAPS methodology used in this thesis and the approach taken to classify and analyse health-related misinformation.

3.1 English Article Data Collection

The CAPS approach for this thesis necessitated English-language news articles specifically annotated as fake news, yet sourcing pre-labelled data that was freely available online posed significant challenges. The primary issue lay in the limited availability of comprehensive, article-style texts labelled as containing fake news, in contrast to the more abundant datasets composed of tweets and brief social media posts, such as those in the COVID-19 Rumor [50], CoAID [31], ANTi-Vax [51], and Truthseeker [52]. This scarcity primarily results from earlier misinformation classification research focusing on tweet-based data before X's API access was restricted in March 2023 [53]. Since this thesis required article-format data, these sources were deemed incompatible.

Additionally, the prevalence of datasets focusing solely on COVID-19 misinformation posed another challenge. To ensure a comprehensive analysis, covering a broad range of medical misinformation themes beyond COVID-19 was necessary. Consequently, the CAPS methodology excluded datasets strictly balanced between fake and genuine news, as only half of their labelled content was relevant to this study.

The English data used in this thesis was sourced from the datasets listed in Table 1, which contain articles specifically formatted as fake news. Health-related keywords filtered sources that were not explicitly health-related to obtain relevant records for this thesis.

Table 1. Previously Annotated English Fake News Datasets

Dataset	Articles	Date Range
Med-MMHL [54]	6059	Jan 2017 – May 2023*
Monant Medical Misinformation Dataset [14]	5680	Apr 2001 – Jan 2022
FNID: Fake News Inference Dataset [55]	2988	Aug 2007 – Apr 2020
ISOT Fake News Dataset [56]	4756	Apr 2015 – Feb 2018
ReCOVery [15]	1910	Jan 2020 – May 2020
Total	**21393**	**Apr 2001 – May 2023**

*Claimed by authors, date variable not included in the dataset.

3.2 Estonian Article Collection

The data collection strategy aimed to gather a comprehensive range of Estonian news sources to identify health-related misinformation on a broad selection of news sites referenced on social media. Recognizing the significant role of Facebook in the exchange and debate of health information and misinformation among the Estonian-speaking community, it was selected as the primary focus. Facebook groups dedicated to medicine, alternative medicine, and other health topics were explicitly targeted. According to data from GS.StatCounter, in 2023, Facebook accounted for 69% of social media usage in Estonia, making it the most popular platform by a significant margin compared to other social media channels [57]. This makes these groups potential key locations for Estonian internet users to seek health information and places where health misinformation can quickly spread due to the mix of accurate information with unverified or false claims.

Within these identified Facebook groups, both posts and comments were systematically collected. The aim was to filter out URLs linked to news articles, which were presumed to be potential carriers of misinformation. This filtering process yielded a collection of websites and their subsites, which were then scrutinized for article content relevant to our study. Notably, the websites included in our study were among the most referenced within the groups, such as Delfi[2], Eesti Päevaleht[3], Telegram,

[2] https://www.delfi.ee/

[3] https://epl.delfi.ee/

Geenius[4], Uueduudised[5], and Objektiiv[6], among others. It is important to note that the methodology applied faced limitations regarding technological constraints, particularly with websites featuring automatic pagination. This aspect restricted our ability to extract data seamlessly across all platforms, underscoring the need for adaptable data collection strategies in future research.

Collected articles were put through strict pre-processing to ensure the best results when embedding into text representations.

- All duplicate entries were removed to ensure that the dataset contained only unique text samples.
- Sentences that were too short were dropped from the dataset, as they often lack sufficient contextual information for effective analysis.
- All hyperlinks were removed to focus purely on textual content.
- Any entries with missing values (NAs) were excluded from the dataset.
- All text entries were converted to lowercase to maintain consistency.
- Non-alphanumeric characters, including punctuation, were removed.

The scraping of news articles was carried out using two Python packages—BeautifulSoup4[7] and Newspaper3k[8]. These packages played a crucial role in web-scraping and text retrieval, significantly reducing the need for data cleaning. The collected information included article links, body text, publishing dates, authors, tags, and the language of each article.

The focus was explicitly on articles categorized under 'health' or similar themes across different news sites. To refine this selection further, articles were filtered using health-related keywords like 'infection,' 'medicine,' and 'virus.' The timeframe of the collected data spanned from November 2010 to January 2024. However, the publishing date was unavailable for 3,918 articles due to scraping limitations. While these articles were retained in the dataset, they are not shown in Figure 1, which displays the distribution of scraped articles by year. After data cleaning and processing, the Estonian dataset contained 21,228 articles.

[4] https://geenius.ee/

[6] https://objektiiv.ee/

[7] https://pypi.org/project/beautifulsoup4/

[8] https://newspaper.readthedocs.io/

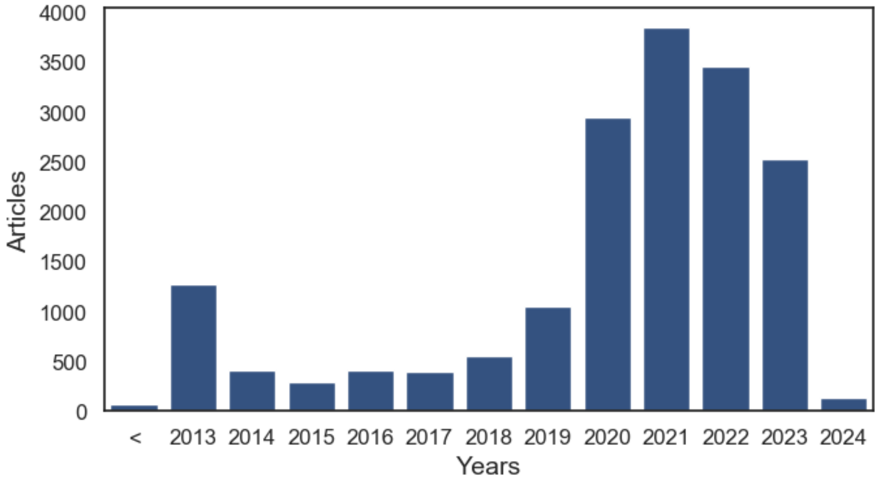


Figure 1. Distribution of Collected Estonian Articles by Year

3.3 Ground Truth

In this thesis, samples of the datasets are manually annotated to estimate the effectiveness and provide an addition to the fake news dataset. Annotation could involve four categories, according to the need of the annotation round. This overview details the specific meanings and implications of the various categories annotators could select when labelling the articles, providing clear guidelines for consistent data classification.

- **Veracity**: This category detects the truthfulness of claims in the article.
  - *Misinformation*: The content is clearly false and misleading, contrary to verified facts.
  - *Not Misinformation*: The content is accurate and aligns with verified information.
  - *Not Sure*: The content's truthfulness cannot be conclusively determined due to ambiguous evidence or lack of information.

- **Similarity**: This category measures the thematic and topical resemblance between articles.
  - *Similar*: Articles possess nearly identical talking points and subject matter.
  - *Sharing a Similar Subtopic*: Articles cover closely related subtopics within the broader main topic.
  - *Somewhat Similar*: Articles discuss broadly related topics but lack specific, point-to-point concordance.
  - *Dissimilar*: Articles discuss the same general topic but do not share specific points or refer to the same details within the same timeframe.
- **Stance**: This category identifies the perspective or attitude that the content takes regarding the topic.
  - *Same Stance*: Both articles present the same viewpoint on the issue discussed.
  - *Different Stance*: Despite discussing similar topics, the articles have opposing viewpoints.
- **Health Theme**: This category checks whether the content is relevant to health topics, crucial for focusing the dataset on health misinformation.
  - *Health-Related*: Directly pertains to health topics or medical information.
  - *Not Health-Related*: Does not pertain to health or medical topics.

For the annotation process, two manual annotators were provided with articles and specific criteria for annotation. Both annotators were proficient in English and Estonian, ensuring they could accurately interpret the nuances of the content. To support their analysis, annotators were encouraged to consult reliable web sources, such as Wikipedia, whenever they encountered ambiguous or unclear information. The consensus between the two annotators was considered the ground truth label for each article. In cases of disagreement, further discussions and verifications were undertaken to reach a consensus.

4 Proposed Approach

This chapter outlines the methodology designed to generate ground-truth labels of misinformation for Estonian health news articles. The developed approach, namely Cross-Lingual Alignment and Confident Prediction Sampling (CAPS) methodology, was used to create an Estonian fake health news dataset from a collection of unlabelled health news articles. By predominantly automating the process of ground-truth label generation, this proposed method significantly reduces the annotator's efforts without compromising the quality of the ground truth.

CAPS was proposed to achieve the thesis's objectives: Phase-I identified the most similar Estonian news articles to annotated English health news articles based on linguistic and semantic characteristics. In Phase-II, a confidence sampling approach integrated manual annotation to establish a gold standard dataset for sequentially generating additional ground-truth labels. This phase enabled effective fine-tuning of classification models and the extraction of the most confident predictions with minimal human intervention.

The chapter emphasises the strategic selection of the best-performing pre-trained multilingual model in each phase to ensure accurate and effective fake news detection, providing a framework for subsequent analysis.

4.1 Phase-I: Similarity Estimation

Phase-I aims to identify Estonian news articles likely containing health misinformation by finding those similar to verified English fake news. The thesis hypothesises that if Estonian news demonstrate significant similarity to known English fake news, they likely contain similar false claims. The input consists of English labelled news articles and unlabelled Estonian news articles, with the expected output being a set of Estonian articles flagged as potentially containing health misinformation.

To achieve this objective, text embeddings are created from both English and Estonian articles, and cosine similarity is calculated between their respective word vectors. As illustrated in Figure 2, Phase-I involves converting collected texts into vector representations using a pre-trained language model and computing cosine similarity between the vectors. This similarity score offers insights into the relationship between languages, which helps identify Estonian news articles with high similarity scores for further analysis.

The conversion from articles to text embeddings forms a vector space model, placing texts in multidimensional positions. The distances or similarities between these positions can then be assessed using cosine similarity [58]. Cosine similarity measures the angle between two vectors, quantifying their similarity based on the cosine of that angle. A value of 1 indicates identical direction (high similarity), while a value of 0 indicates orthogonality (complete dissimilarity) [58].

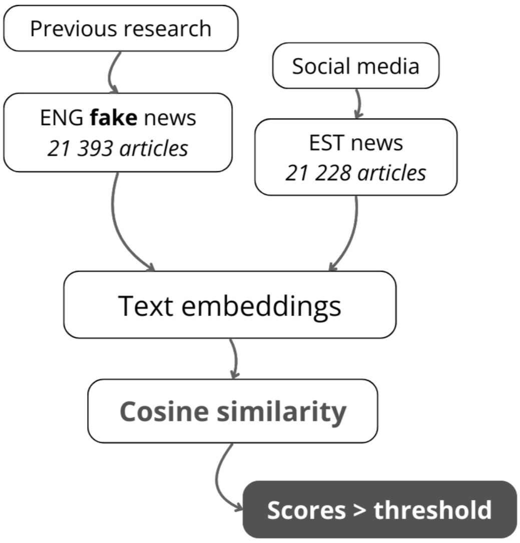


Figure 2. Workflow Process of Phase-I Methodology

Although there are several traditional mechanisms of similarity calculation in text, embedding-based similarity calculation can ensure better effectiveness [59], providing vector-based representations that transform text into numeric formats. This conversion bridges the gap between human linguistic comprehension and machine processing [19]. These embeddings are usually generated by language models trained on extensive text corpora in one or multiple languages [19].

In recent years, the creation of text embeddings has shifted towards language models that generate context-aware text representations, particularly those based on transformer architectures like BERT [60]. BERT and its derivatives have demonstrated substantial improvements across various NLP tasks, delivering remarkable results and setting new benchmarks for performance [60, 19]. These advancements mark significant progress in enabling machines to process and understand text in a contextually meaningful manner.

Various multilingual embedding models are considered and tested by comparing samples in both languages to ensure that the similarity scores accurately reflect the actual relationship between texts. Based on previous research and their support for the Estonian language, LASER [61], multilingual BERT [60, 62], and Sentence-BERT [63] were chosen for evaluation.

- Language-Agnostic SEntence Representations (**LASER**): Developed by the Facebook Research team, LASER supports over 200 languages, including Estonian [61], making it a suitable choice for this research. It is specifically designed for efficient cross-lingual information retrieval and translation tasks, leveraging a shared multilingual embedding space to achieve this goal. Using this shared space, LASER can handle semantic representations across multiple languages, making it effective for multilingual tasks such as translation and cross-lingual search [61].

- Multilingual BERT (**mBERT**): Widely used for generating multilingual embeddings, mBERT is trained on a large multilingual corpus covering the top 104 languages in a self-supervised manner [60]. As previous studies have indicated [12, 11], mBERT provides a robust baseline for multilingual natural language understanding tasks. It offers valuable embeddings that facilitate many forms of natural language processing across different languages [60].

- Sentence-BERT (**SBERT**): Originating from the Sentence Transformers library, SBERT is known for efficiently generating embeddings by modifying the BERT architecture for sentence pair tasks [63]. Employing Siamese or triplet networks enables highly accurate and computationally efficient sentence embeddings for semantic search and similarity assessment. Its optimised architecture allows SBERT to handle sentence pairs swiftly, making it ideal for semantic similarity tasks and multilingual embeddings [63].

This thesis recognises a known limitation of pre-trained language models related to their restricted maximum sequence length [13]. This limit determines the number of words or tokens that can be embedded, leading to the truncation of any content beyond that length. Nevertheless, the thesis assumes that the core message of an article typically appears at the beginning, which makes text truncation during embedding creation less significant.

Phase-I involves testing and comparing Estonian-to-Estonian and Estonian-to-English texts to calculate similarities and manually verify their contextual alignment, ensuring that the embeddings accurately represent Estonian. This step is crucial due to the limited availability of annotated fake health news resources in Estonian. Additionally, compared to English, Estonian is less represented in the training data of these language models. Each model is evaluated for its ability to identify similar sentences within the same language and cross-lingually and recognise sentences with similar or opposing stances across individual sentences and longer paragraphs. The subsequent chapter will offer a subjective comparative analysis of the performance of various embedding models for Estonian.

4.2 Phase-II: Dataset Creation

Phase-II aims to create a comprehensive dataset of Estonian fake health news by establishing a classification and confidence sampling pipeline, with each step adding additional labelled data points. This will be achieved by creating a gold standard for Estonian fake news, which will support the classification models and confidence sampling processes used to expand the resulting dataset. The initial input for this gold standard is the dataset referenced at the end of Phase-I, from which a representative sample undergoes manual annotation. Figure 3 depicts the complete pipeline of Phase-II.

Creating an automated process for flagging fake news is essential since manually annotating all collected articles is not feasible due to the time and effort required. Nevertheless, a subset was randomly selected for annotation for the initial gold standard dataset. This annotation followed the rules outlined in Section 3.3. This annotated dataset is the basis for fine-tuning the pipeline's first classifier.

The chosen model is trained exclusively on the manually annotated gold standard dataset in the first classification step. The classifier's predictions focus on the subset of potential fake news from Phase-I that was not selected for annotation, as these articles are potentially more likely to contain misinformation among all collected Estonian articles. The second classification model is then fine-tuned on the combined gold standard dataset and high-confidence predictions from the first step.

The classification pipeline required selecting an appropriate model to label Estonian articles. This selection was made by validating several BERT-based models, including mBERT [60], XLM-RoBERTa [64, 65], and ELECTRA [66], to find the most suitable one for generating text embeddings and making predictions. Transformer-based models were chosen due to their proven effectiveness in fake news classification [12, 43, 13]. BERT-based models have set a foundational standard for the architecture and capabilities expected in language processing tasks.

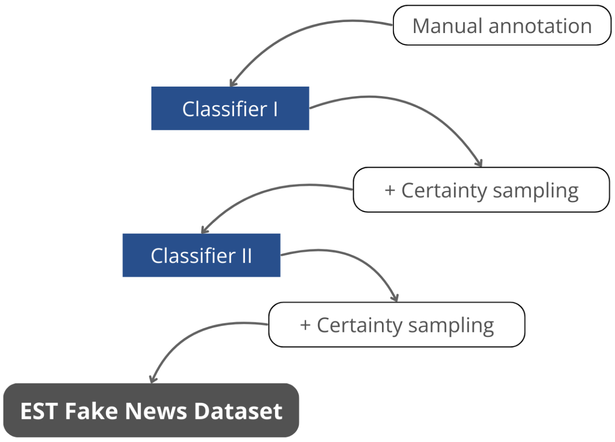


Figure 3. Workflow Process of Phase-II Methodology

BERT (Bidirectional Encoder Representations from Transformers), initially introduced by Devlin et al. [60], is the first deeply bidirectional, unsupervised language representation model [19, 60]. Its bidirectional structure enables the model to apply self-attention in both forward and backward directions, understanding each word's context based on surrounding words [19, 60]. BERT employs masked language modelling to enhance further comprehension, wherein random words are hidden, and the model predicts them based on the surrounding context. This technique facilitates deeper contextual learning [60]. Combined with the bidirectional structure, BERT outperforms traditional models that process text unidirectionally in understanding complex sentence relationships. This comprehensive contextual awareness significantly improves performance across tasks like text classification and sentiment analysis [60].

Multilingual BERT models are highly effective at learning deep, context-aware representations from multilingual, unannotated corpora through self-supervised training [19]. Since Estonian is underrepresented in training data and lacks dedicated fake news datasets, multilingual models offer distinct advantages. Despite their high computational cost due to the large number of parameters [19], this thesis focused on increasing accuracy and reducing annotation time rather than minimising training complexity. Thus, the computational expense was not considered a major limitation.

4.2.1 Classifier Selection

As potential models for the classification task, three pre-trained models were considered: mBERT [60], XLM-RoBERTa [64, 65], and ELECTRA [66], all of which evolved from the original BERT framework.

- Multilingual BERT (**mBERT**): A variant of BERT designed to support multiple languages, mBERT is trained on Wikipedia articles from 104 languages, focusing on the largest corpora, making it ideal for cross-lingual tasks [62]. This model employs a shared vocabulary of subword tokens that captures multilingual texts' semantics and builds effective representations across different languages. The shared vocabulary and the model's parameters facilitate efficient zero-shot learning across languages, enhancing its performance in low-resource settings where annotated data is limited [62].

- Cross-Lingual Model - Robustly Optimised BERT Pretraining Approach (**XLM-RoBERTa**): An extension of the RoBERTa model, XLM-RoBERTa is specifically designed to handle multilingual tasks [65]. RoBERTa, introduced by Yinhan Liu et al. [64], was trained on a dataset ten times larger than BERT's original corpus. It uses dynamic masking, larger batch sizes, and extended training iterations to develop a more powerful version of BERT [64, 19]. XLM-RoBERTa builds upon this foundation by training on a vast multilingual corpus, offering the same robust features across different languages. This model is particularly effective in low-resource settings due to its ability to handle diverse languages and tasks. Its architecture has proven efficient in studies focusing on misinformation classification, consistently delivering impressive results [67, 47, 39].

- Efficiently Learning an Encoder that Classifies Token Replacements Accurately (**ELECTRA**): Introduced by Clark et al. [66], ELECTRA differs from BERT and RoBERTa by using replaced token detection instead of traditional masking techniques. In this approach, the model learns to differentiate between genuine input tokens and plausible yet synthetically generated ones [66]. This innovative method allows ELECTRA to achieve performance comparable to RoBERTa while requiring only a quarter of the computational resources [19].

The model used for the classification task is selected through training and testing on gold standard data. Although the pre-trained model choice remains the same throughout both stages of the pipeline, it is trained on slightly different datasets and hyperparameters each time to ensure peak performance in detecting fake news at each step.

4.2.2 Confidence Sampling

This thesis applies confidence sampling to enhance the classification process by including only the most confident predictions in the dataset. Although predictions are made on all unlabelled Estonian articles at each step, only those with the highest confidence are incorporated into the new training dataset for the subsequent stage.

BERT-based classification models rely on calculated classification scores or output logits, which are converted into probabilities using a sigmoid function. This process creates a vector of probabilities for each class, and the predicted class is the one with the highest probability, representing the model's confidence in that prediction. This probability score ranges between 0 and 1, with higher values indicating greater confidence in the prediction [68].

A confidence sampling technique ensures that the new training dataset includes only highly confident predictions. This method, inspired by Palakodety et al. [69], employs a confidence sampling strategy in conjunction with sequential classification models. This innovative approach expanded their seed dataset from 11 YouTube comments to 2,790 comments about the Rohingya refugee crisis. Given the results, this technique was deemed suitable for this thesis.

Confidence sampling incorporates only predictions that exceed a predetermined confidence threshold into the training data for the next stage. This approach measures the model's confidence in its classifications and refines its accuracy and precision by filtering out more complex samples. As a result, the model can train for more epochs. While overfitting may occur, it does not impact the final results because only the most confident samples are extracted, ensuring accurate identification of fake news.

This sampling process required establishing a confidence threshold for predictions. The threshold was set by analysing the validation metrics and remaining prediction dataset sizes for the selected classifier across various thresholds, ensuring the best balance moving forward.

5 Results

This chapter presents the results of a CAPS methodology developed to generate ground-truth labels for misinformation in Estonian health news articles, resulting in an Estonian fake health news dataset from an initially unlabelled collection.

A two-phase CAPS approach was employed to achieve this objective. In Phase-I, a gold standard dataset was created using manual annotation combined with cosine similarity analysis to identify potential misinformation. Phase-II involved a two-step process to refine the dataset further: classification models filtered the content, followed by certainty sampling to expand the dataset.

The results section examines the impact of these methods on refining predictions and expanding the dataset. It concludes with a comprehensive evaluation and discussion, highlighting how the strategic implementation of multilingual pre-trained models and certainty sampling techniques in each phase ensures precise and effective fake news detection.

5.1 Phase-I: Text Similarity Calculation

Phase-I of the CAPS methodology aim to identify Estonian health news articles that exhibited thematic similarity to English fake health news articles. This required finding an effective embedding model for cross-lingual comparison of English and Estonian articles. For this, three models were tested: LASER, mBERT, and SBERT. These models have been introduced in more detail in Section 4.1.

5.1.1 Embedding Model Comparison

The process involved embedding a selected sample from collected Estonian and English datasets and then manually assessing the semantic accuracy of the articles. The cosine similarity scores were evaluated to determine how accurately they reflected the degree of similarity between English and Estonian articles. This evaluation was based on comparing these scores to the manual assessments, ensuring the chosen embedding model could effectively capture the semantic relationships across the two languages.

During the manual assessment of articles embedded with different models, it became evident that LASER and mBERT struggled with textual negations, leading to overly optimistic similarity scores. Furthermore, LASER demonstrated significantly lower performance when embedding Estonian articles than English ones, highlighting issues in accurately assessing semantic similarity for Estonian content. mBERT's accuracy in assessing similarity also decreased with longer texts. Conversely, SBERT provided clear insight into cross-lingual similarity and maintained accuracy when comparing Estonian articles. Therefore, SBERT was selected to generate the embeddings for further analysis in this thesis.

5.1.2 Cosine Similarity Scoring

The next step in the CAPS methodology was to select Estonian articles similar to the English fake news. This required establishing a threshold for similarity, with cosine similarity ranging from 0 to 1, where 1 indicates complete similarity.

To achieve this, samples from Estonian and English article pairs were manually analysed at various similarity levels. SBERT was utilised to create the embedding and cosine similarity scores were calculated to quantify textual similarities. This quantitative analysis enabled the establishment of specific thresholds to categorise news article pairs based on their degree of similarity.

- texts with similarity scores below 0.5 were considered dissimilar.
- Scores between 0.5 and 0.6 indicated that texts were not precisely similar but shared a broad theme, such as medicine.
- Scores from 0.6 to 0.7 denoted a closer thematic similarity, for example, texts discussing vaccines or cancer.
- texts scoring above 0.7 were found to share more precisely defined themes, such as specific diseases and related opinions.

This approach selected a threshold of 0.75 to define the similarity between articles in the two languages. Articles meeting this criterion were considered thematically similar in content and conclusions.

As the next step in this phase, text embedding were created from 21,393 pre-annotated English fake news articles and 21,228 unlabelled Estonian health articles. The total time taken to generate embedding was 27 minutes, and the calculation of similarity scores took approximately 10 hours and 8 minutes.

The dataset containing all computed scores was vast, with over 454 million potential combinations, approximately 5.6 million of which had a similarity score above the 0.5 threshold. After calculating similarity, the subset of Estonian articles with scores over 0.75 reduced the dataset to 14,267 unique article pairs or 2,058 unique Estonian articles.

The total distribution of Estonian articles with a similarity higher than the 0.75 threshold compared to English fake news articles is shown in Figure 4. This distribution highlights that the count of articles with high similarity to fake news decreases exponentially, with only a minimal number exceeding a similarity score of 0.8 compared to the 0.75 threshold.

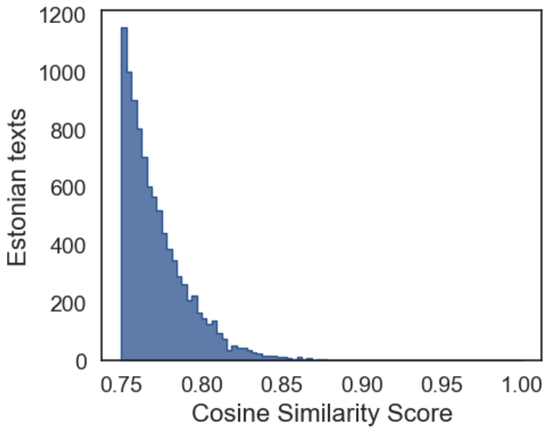


Figure 4. Estonian Articles by Cosine Similarity Scores (Threshold > 0.75)

When evaluating the similarity matches for Estonian news articles, it was observed that the majority of articles corresponded with only one English article, as illustrated in Figure 5. This outcome is interpreted as a positive indication of the methodology's effectiveness, demonstrating that Estonian news articles do not arbitrarily align with unrelated news articles but rather with those exhibiting substantial similarity. For example, it is postulated that an article generically discussing COVID-19 will not have high similarity with all other articles on COVID-19 but will specifically need to address the same topics and views.

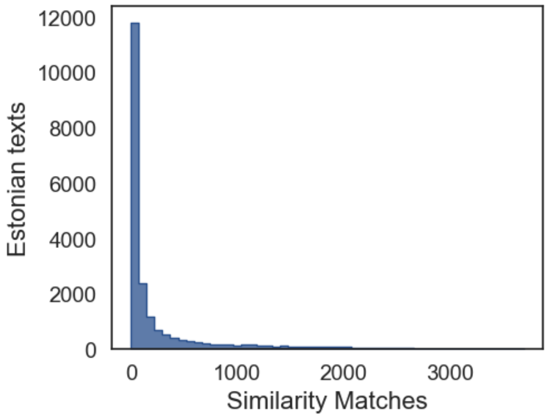


Figure 5. Similarity Matches of Estonian and English Articles (Threshold > 0.5)

5.2 Phase-II: Generation of Ground Truth Labels

Phase-II of this study focuses on generating ground truth labels for unlabelled Estonian health news articles using a confidence sampling technique integrated with multiple classification models. Initially, a subset of articles is manually annotated to create a gold standard dataset. The selected multilingual classifier is applied in a two-step classification process with confidence thresholds to ensure high-quality predictions. This approach aims to enhance the model's accuracy and generalisability, progressively expanding the labelled dataset of Estonian health articles.

5.2.1 Gold Standard Data Annotation

From the pool of 2,058 Estonian articles identified as potential fake news, a subset of 500 articles was randomly selected for manual annotation. During the annotation process, the Estonian news article and the corresponding English news article—identified as having the highest similarity score with the selected Estonian article—were chosen for further examination. This annotation assessed three categories: similarity, stance, and veracity. Similarity evaluates whether the two articles share common themes and content. If the similarity is established, the stance is then assessed to determine whether the perspectives presented are aligned or divergent; for instance, one article might advocate for a vaccine while the other opposes it. Finally, veracity examines whether the Estonian article can be classified as misinformation. Section 3.3 describes the detailed annotation method.

Figure 6 illustrates the distribution of the manual annotations. Regarding similarity, the evaluation indicated that 47% of the news article pairs between Estonian and English were classified as dissimilar. The exact similarity was observed in 12.2% of the news article pairs. Regarding stance, 78.5% of the similar articles shared the same stance. This annotation process took approximately 30 hours for two annotators to complete.

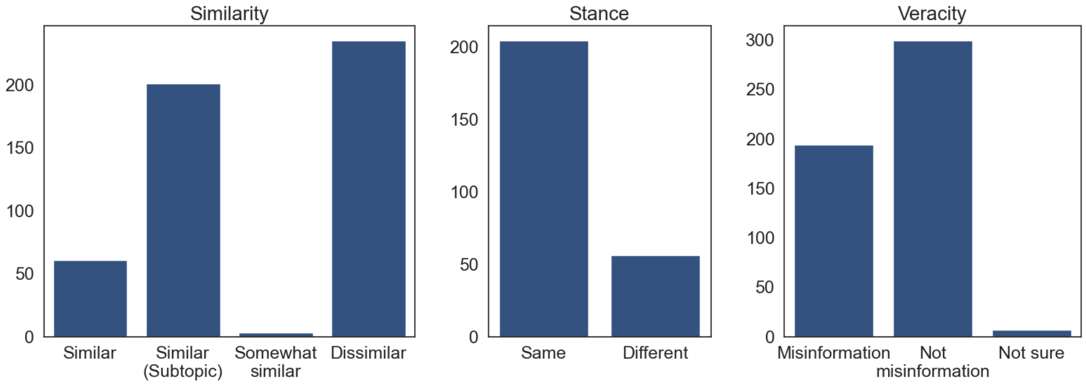


Figure 6. Gold Standard Annotation: Similarity, Stance, and Veracity

The distribution of veracity revealed that this methodology component identified a significant portion of articles as misinformation, with 38.8% of the randomly sampled articles being annotated as such. This annotated dataset consists of Estonian articles, of which 299 were classified as not fake news, and 194 were identified as fake news, establishing a gold standard for the study.

5.2.2 Classifier Comparison

A gradual classification and sampling method was employed to gather labels for the unlabelled Estonian health articles, necessitating the identification of a suitable multilingual language model. Given the limited size of the gold standard dataset—comprising 500 samples—and the linguistic nuances specific to Estonian, the classification task posed substantial challenges. Three models—mBERT, XLM-RoBERTa, and ELECTRA—were compared to determine the best performance on this data.

Each model was evaluated using a dataset created from manual annotations over 12 epochs, incorporating a learning rate 1e-5 and a batch size of 64. The training dataset was split, with 20% used for validation, and then upsampled for the minority class to maintain balanced labels. The optimiser used was AdamW[9] with a regularisation factor of 0.1, and the sequence length for all models was 512 tokens. Comprehensive fine-tuning was performed on each model, including both the transformer blocks and the classification head, to maximise accuracy and precision with the available data and adapt to the specific linguistic characteristics of the Estonian language.

The results, shown by the training loss, validation loss, and validation accuracy for each model in Figures 7, 8, and 9, respectively indicated significant overfitting. Since these language models were extensively pre-trained on large textual datasets, it was hypothesised that the observed overfitting in this initial phase was due to the limited training data available. However, as explained in Section 4.2.1, overfitting was not considered an issue because it was counterbalanced by the confidence sampling technique.

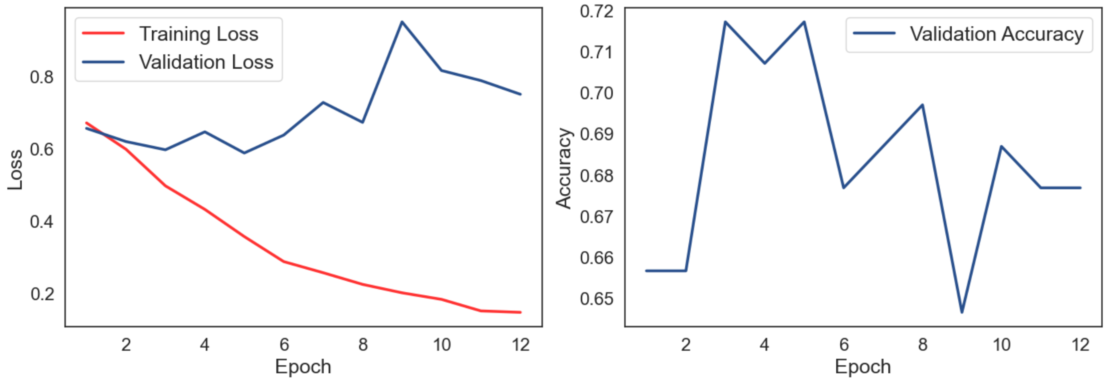


Figure 7. Training and Validation Metrics of mBERT

[9] https://keras.io/api/optimizers/adamw/

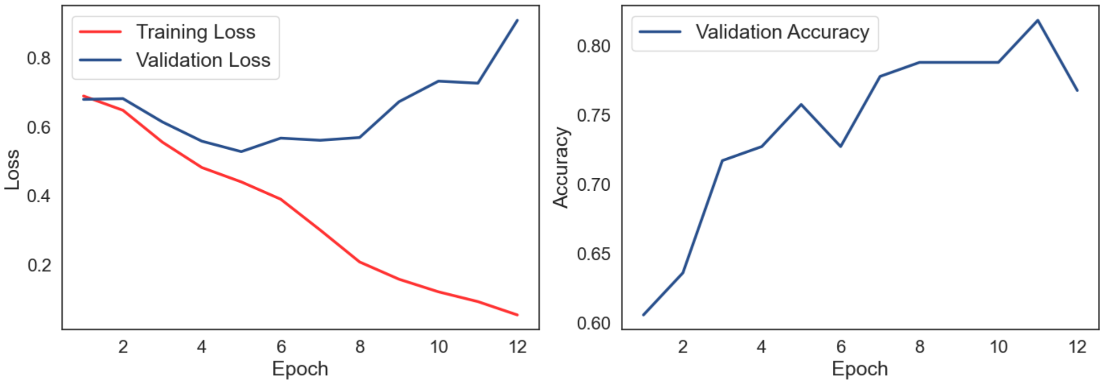


Figure 8. Training and Validation Metrics of XLM-RoBERTa

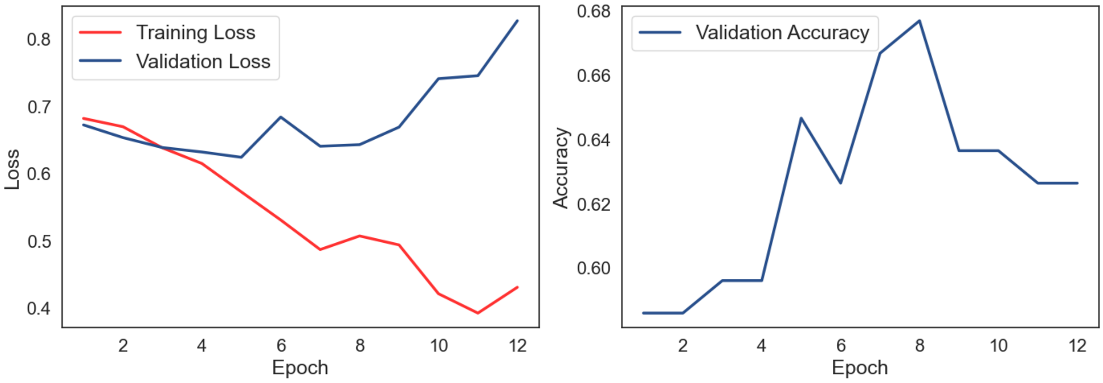


Figure 9. Training and Validation Metrics of ELECTRA

Comparing the validation accuracies of the fine-tuned models revealed that XLM-RoBERTa achieved the highest accuracy at 0.768, followed by mBERT at 0.677 and ELECTRA at 0.626. Despite its decent accuracy, ELECTRA exhibited a notably low F1 score, suggesting it struggles with class imbalance or distinguishing between specific classes. Furthermore, when comparing precision and F1 scores, as shown in Table 2, XLM-RoBERTa outperformed both mBERT and ELECTRA.

Table 2. Validation Results

Model	F1 Score	Precision	Accuracy
mBERT-base	0.6364 ± 0.0948	0.5600 ± 0.0978	0.6768 ± 0.0921
XLM-RoBERTa-large	**0.6462** ± 0.0942	**0.7778** ± 0.0819	**0.7677** ± 0.0832
ELECTRA-base	0.1778 ± 0.0753	0.5714 ± 0.0975	0.6263 ± 0.0953

The superior accuracy and precision of XLM-RoBERTa's predictions are clearly shown in Figures 10, 11, and 12, which compare validation predictions for misinformation and non-misinformation categories.

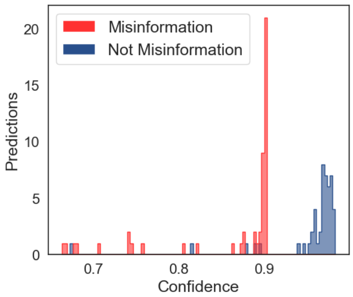


Figure 10. mBERT

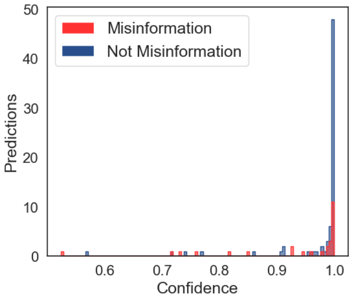


Figure 11. XLM-RoBERTa

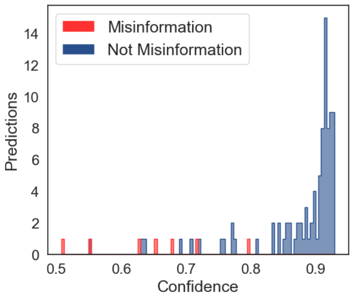


Figure 12. ELECTRA

The mBERT model shows uncertainty, failing to classify any articles as misinformation with a confidence level above 0.9. Conversely, the ELECTRA model demonstrates tunability but tends to classify all articles as fake news. Only XLM-RoBERTa provides a balanced distribution between potential labels with significant confidence. Therefore, it was chosen as the model for both steps of the Phase-II.

5.2.3 Confidence Sampling Results

As outlined in Section 4.2.2, CAPS involves applying confidence sampling to the predictions from classification models. A threshold was established for this, meaning a prediction needed to exceed this threshold to be included in the next step's training dataset. Various thresholds were tested using XLM-RoBERTa validation predictions, and the corresponding F1 scores, precision, and accuracy values for these thresholds are presented in Table 3.

Table 3. 1st Step Model (XLM-RoBERTa) Validation Results for Different Thresholds

Threshold	**Size**	**F1 Score**	**Precision**	**Accuracy**
0.8	91.9%	0.6545 ± 0.0977	0.7826 ± 0.0847	0.7912 ± 0.0835
0.9	88.9%	0.6538 ± 0.0994	0.8095 ± 0.0820	0.7955 ± 0.0843
0.95	82.8%	0.6512 ± 0.1032	0.7778 ± 0.0900	0.8171 ± 0.0837
0.99	68.7%	0.7500 ± 0.1029	0.8571 ± 0.0832	0.8824 ± 0.0766
0.995	57.6%	**0.8000** ± 0.1038	**1.0000** ± 0.0000	0.9123 ± 0.0734
0.996	52.5%	0.7778 ± 0.1130	1.0000 ± 0.0000	**0.9231** ± 0.0724
0.997	50.5%	0.7778 ± 0.1152	1.0000 ± 0.0000	0.9200 ± 0.0752
0.998	35.4%	0.6667 ± 0.1562	1.0000 ± 0.0000	0.9143 ± 0.0927

It was observed that the optimal results were obtained with a confidence threshold set at 0.995, based on the comparative analysis of F1 scores and precision. At this threshold, 57.6% of predictions were incorporated into the dataset, achieving an accuracy of 0.9231. This threshold was subsequently adopted as the standard for all stages of confidence sampling throughout the classification process. The first-step model produced predictions on the dataset from Phase-I, excluding data points randomly sampled for manual annotation. This process resulted in 722 predictions exceeding the confidence threshold of 0.995: 499 negative and 223 positive predictions, with positive predictions characterising an article as fake news. The distribution of predictions by their confidence levels can be seen in Figure 13. This step expanded the dataset from 493 articles to 1,215, increasing the overall dataset size about two times.

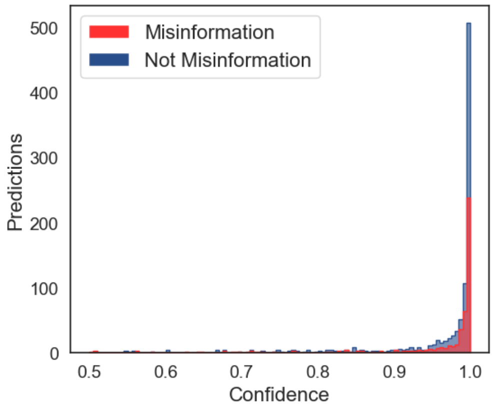


Figure 13. Prediction distribution of XLM-RoBERTa

The second step model of Phase-II was trained on a combined dataset consisting of 1,215 articles: 798 non-fake articles and 417 deemed fake news. This iteration spanned five epochs with a learning rate of 1e-5, a batch size of 64, and a sequence length of 512 tokens. The dataset was split, with 80% used for training and 20% for validation. The unbalanced training dataset was upsampled to ensure balanced labels for improved performance and generalisability. The optimiser used was AdamW, with a regularisation factor of 0.1. The training and validation losses, along with the validation accuracy, are illustrated in Figure 14.

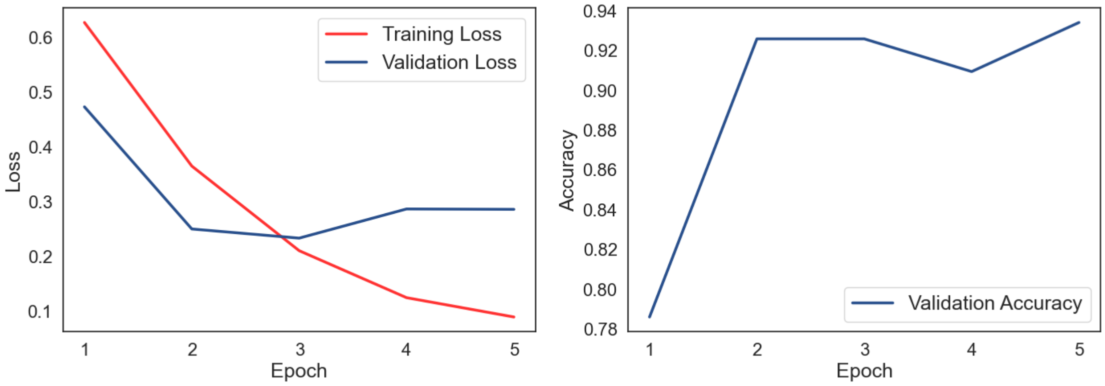


Figure 14. Training and Validation Metrics of 2nd Step Model (XLM-RoBERTa)

Validation of the model indicates that, although confidence sampling results in the loss of some samples, the performance metrics demonstrate significant improvements. As seen from Table 4, this step illustrates a substantial increase in accuracy, precision, and F1 score, likely due to the more extensive training dataset. The F1 score of 0.96 and accuracy of 0.96 indicate that the model has learned effectively.

Table 4. 2nd Step Model (XLM-RoBERTa) for Different Thresholds

Condition	**Size**	**F1 Score**	**Precision**	**Accuracy**
No Threshold	100%	0.9252 ± 0.0331	0.8919 ± 0.0390	0.9342 ± 0.0312
Threshold 0.995	67.9%	**0.9643** ± 0.0283	**0.9419** ± 0.0357	**0.9636** ± 0.0286

With the second-step classifier, predictions are made on all collected Estonian articles that the model has not been trained on. Predictions surpassing the confidence threshold of 0.995 classify 771 articles as fake news and 1,229 as not fake news. Consequently, the entire dataset expanded from 1,215 to 3,215 articles, doubling in size. Table 5 illustrates this progressive dataset growth across steps.

Table 5. Dataset Growth of Phase-II Steps

Step	**Positive**	**Negative**	**Total**	**Dataset Size**	**Increase**
Gold	194	299	493	493	-
Classifier I	223	499	722	1215	146.45
Classifier II	771	1229	2000	3215	164.61

5.3 Evaluation

This section evaluates the performance of CAPS approach to generate ground truth labels. The final dataset is assessed using essential metrics—accuracy, precision, and F1 score. These metrics ensure that the labels gathered in Phase II reflect the articles' genuine semantic and thematic properties. Additionally, the dataset created through CAPS methodology undergoes further analysis, including additional manual annotation, to validate the classification and sampling techniques used.

5.3.1 Evaluation Metrics

This thesis employs standard classification metrics to evaluate the performance of classifiers against manually annotated labels.

- *True Positive* (TP): The model accurately identifies an instance as positive when it is genuinely positive. In this thesis, this means correctly detecting fake news when an article is indeed fake.
- *True Negative* (TN): The model accurately classifies a non-fake article as genuine.
- *False Positive* (FP): The model incorrectly classifies a genuine article as fake news.
- *False Negative* (FN): The model mistakenly classifies a positive instance as negative, meaning it labels a fake article as genuine.

The accuracy of a model is defined by its ability to classify news articles as false or truthful correctly, as quantitatively expressed in Equation 1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is chosen for this task to highlight the classifier's exactness. It measures how effectively the model avoids falsely labelling genuine articles as fake, as delineated in Equation 2.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Finally, the F1 score, as described in Equation 3, is the weighted harmonic mean of the classifier's ability to correctly identify true positive predictions and to avoid incorrectly labelling negative instances as positive. For this thesis, a high F1 score indicates that the model efficiently identifies fake news articles while minimizing the number of genuine articles wrongly classified as fake.

$$\text{F1 Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3}$$

5.3.2 Dataset Analysis

The fake news dataset created through the CAPS methodology, finalized in Phase-II, underwent further analysis to evaluate the efficacy of this approach. This assessment involved additional manual annotation of a sample of 300 articles in Estonian to determine whether the dataset was suitable for research purposes and to test the CAPS's effectiveness. As with previous annotation rounds, the procedure followed the guidelines detailed in the Ground Truth subsection of the Data section. However, unlike the previous round that assessed the similarity of Estonian and English news articles, their stance, and veracity, this annotation was limited to two categories, focusing solely on verifying the veracity of the news articles and whether their themes were genuinely health-related. This focus was crucial for calculating the F1 score, precision, and accuracy needed for further analysis.

The annotation process took a total of 4 hours to complete for two annotators. In contrast, the previous manual annotation of 500 samples across three categories and two languages required approximately 30 hours. As illustrated in Table 6, the validation results demonstrated an overall F1 score of 0.80 and an accuracy of 0.81 for the entire annotated dataset. For articles verified as genuinely health-related, the performance metrics increased, achieving an F1 score of 0.88 and an accuracy of 0.90. However, the model's effectiveness was considerably lower in detecting fake news within non-health-related articles, primarily due to its tendency to mistakenly classify content as misinformation more frequently than expected. This resulted in reduced F1 scores and accuracy for this subset of the dataset.

Table 6. Validaton Results by Category

Category	**Size**	**F1 Score**	**Precision**	**Accuracy**
All Articles	100%	0.7971 ± 0.0455	0.9322 ± 0.0284	0.8133 ± 0.0441
Health Articles	76%	**0.8791** ± 0.0423	0.9091 ± 0.0373	**0.9035** ± 0.0383
Non-Health Articles	24%	0.6383 ± 0.1110	1.0000 ± 0.0000	0.5278 ± 0.1153

The distribution of fake and truthful news across different news sites is presented in Figure 15. The analysis reveals that some sites predominantly disseminate misinformation, while others, like Geenius, exhibit no fake news. This outcome aligns with previous research findings, which suggest that Telegram has been found to host misinformation and conspiracy theories, many of which are thematically linked to the COVID-19 pandemic. In addition to Telegram, both Objektiiv and Uueduudised have most articles labelled as fake news.

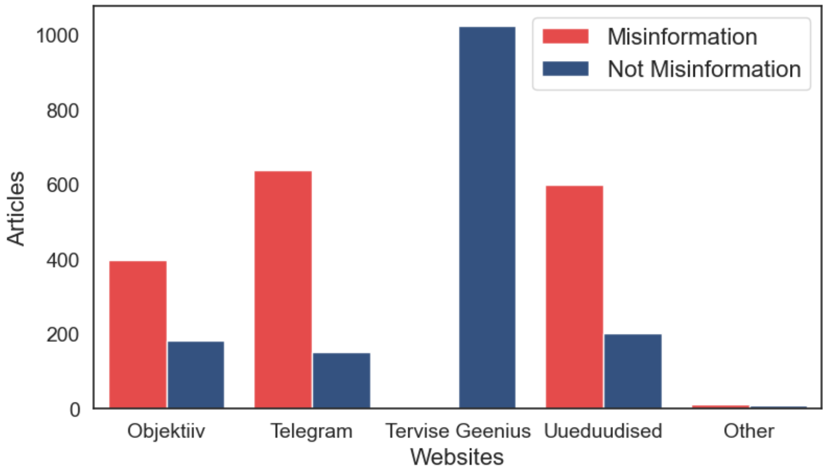


Figure 15. Distribution of Articles from Estonian News Sites

The dataset generated through the CAPS methodology is balanced and comprehensive, containing a total of 3,215 articles, with 1,646 labelled as fake news and 1,569 identified as genuine. This thesis contributes three pivotal resources: the final balanced dataset, the gold standard dataset, and the final manual annotation dataset, all established through rigorous annotation and applied methodology.

To the best of the author’s knowledge, this constitutes the first dataset focused on Estonian fake health news. It stands as a significant asset for the research community, providing a foundational resource for advancing the study of misinformation detection in Estonian and other low-resource settings.

5.4 Discussion

The Cross-Lingual Alignment and Confident Prediction Sampling (CAPS) methodology has demonstrated its ability to generate fake news labels for health news in low-resource settings through its robust, two-phased application. Integrating similarity estimation with SBERT during Phase-I was not merely a procedural step; it was pivotal for establishing a dependable framework for cross-lingual analysis. This was particularly crucial for setting the groundwork where traditional methods might falter due to linguistic variations between English and Estonian. SBERT excelled in identifying thematic relationships across languages, proving essential for accurate initial filtering of potential misinformation.

In Phase-II, the innovative use of confidence sampling with advanced language models like XLM-RoBERTa significantly enhanced the dataset's quality and scope. XLM-RoBERTa's superior classification performance underlined CAPS's efficacy in handling complex, low-resource tasks. This phase did more than expand the dataset; it refined the data quality, ensuring that only high-confidence predictions were retained. By effectively reducing noise and errors typically present in unlabelled datasets, CAPS refined over 20,000 news articles down to a more manageable and accurate dataset of 3,000, achieving a high F1 score of 0.8.

CAPS provided a scalable and efficient solution for detecting health misinformation across language barriers. The ability of CAPS to triple the dataset size at each classification stage demonstrated its reliability and showcased its potential as a model for similar challenges in other linguistic contexts. This approach has not just advanced the state of misinformation detection in Estonian; it has set a new standard for how such challenges can be approached with nuanced, language-sensitive technologies.

Despite these achievements, the study encountered limitations. The data collection, primarily sourced from Facebook group posts, might not fully represent the diversity of misinformation circulating across Estonian online spaces. While enabling efficient annotation and processing, this focused data source also introduced some noise, as not all articles were directly health-related or relevant to the study's goals. Consequently, the models faced challenges distinguishing health misinformation from other themes and struggled with non-health-related articles. These limitations could be addressed in future works.

6 Conclusion

The pervasive spread of health fake news has become a significant public health risk, particularly during the COVID-19 pandemic, which highlighted the phenomenon known as the 'infodemic'. This term, coined to describe the overwhelming surge of misinformation, exacerbates the challenges to public health by undermining efforts like vaccine uptake and compliance with health guidelines. The thesis aims to develop a robust method for generating ground truth labels for fake health news in Estonian. This work contributes to the broader field of fake news detection in low-resource settings.

The Cross-Lingual Alignment and Confident Prediction Sampling (CAPS) methodology was introduced to enhance detecting and annotating health fake news. Phase-I employed Sentence-BERT to establish thematic relationships between English and Estonian news articles, setting the groundwork for accurate cross-lingual information transfer using a pre-existing English fake news dataset. Phase-II built upon this foundation with a focus on manual annotation, XLM-RoBERTa classification, and confident prediction sampling, all pivotal in refining and expanding the dataset.

CAPS methodology effectively tripled the dataset from 500 to over 3,000 articles, achieving impressive metrics such as an overall F1 score of 0.80 and an accuracy of 0.81. Performance was even higher for verified health-related news articles, with F1 scores of 0.88 and an accuracy of 0.90. These results underscore the substantial efficacy of the CAPS methodology in a low-resource linguistic context. Additionally, this research led to the creation of a ground-truth dataset comprising 3,215 Estonian health-related news articles. To the best of this thesis's knowledge, this dataset is the first of its kind and serves as a vital resource for the research community, laying the groundwork for future studies in fake news detection.

Future work could explore a broader spectrum of news sources and themes, enhancing the understanding and detecting fake news. Incorporating images and applying text summarisation techniques could address multi-modal fake news and maintain context in longer articles, respectively. Exploring alternative classification models, including Large Language Models, could enhance the scalability and generalisability of the CAPS methodology. Additionally, future work can utilise the established ground truth labelled dataset to develop a more effective automated fake health news detection model tailored for Estonian.

References

[1] Miriam Fernandez and Harith Alani. Online misinformation: Challenges and future directions. In *Companion proceedings of the the web conference 2018*, pages 595–602, 2018.

[2] Christian Scheibenzuber, Laurentiu-Marian Neagu, Stefan Ruseti, Benedikt Artmann, Carolin Bartsch, Montgomery Kubik, Mihai Dascalu, Stefan Trausan-Matu, and Nicolae Nistor. Dialog in the echo chamber: Fake news framing predicts emotion, argumentation and dialogic social knowledge building in subsequent online discussions. *Computers in Human Behavior*, 140:107587, 2023.

[3] Brian G Southwell, Jeff Niederdeppe, Joseph N Cappella, Anna Gaysynsky, Dannielle E Kelley, April Oh, Emily B Peterson, and Wen-Ying Sylvia Chou. Misinformation as a misunderstood challenge to public health. *American journal of preventive medicine*, 57(2):282–285, 2019.

[4] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[5] World Health Organization. Managing the covid-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. `https://www.who.int/health-topics/infodemic#tab=tab_1`, 2020. Accessed: 2023-05-10.

[6] Sander Van Der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine*, 28(3):460–467, 2022.

[7] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199, 2020.

[8] Roland Imhoff and Pia Lamberty. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118, 2020.

[9] Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[10] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro-and anti-vaccination views. *Nature*, 582(7811):230–233, 2020.

[11] Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20, 2021.

[12] Mohammadreza Samadi, Maryam Mousavian, and Saeedeh Momtazi. Persian fake news detection: Neural representation and classification at word and text levels. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–11, 2021.

[13] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Fake news detection in low-resource languages: A novel hybrid summarization approach. *Knowledge-Based Systems*, page 111884, 2024.

[14] Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2949–2959, 2022.

[15] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212, 2020.

[16] Daniel Freeman, Felicity Waite, Laina Rosebrock, Ariane Petit, Chiara Causier, Anna East, Lucy Jenner, Ashley-Louise Teale, Lydia Carr, Sophie Mulhall, et al. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in england. *Psychological medicine*, 52(2):251–263, 2022.

[17] Eve Dubé, Maryline Vivion, and Noni E MacDonald. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert review of vaccines*, 14(1):99–117, 2015.

[18] Gary M Armstrong, Metin N Gurol, and Frederick A Russ. A longitudinal evaluation of the listerine corrective advertising campaign. *Journal of Public Policy & Marketing*, 2(1):16–28, 1983.

[19] Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, pages 1–59, 2023.

[20] Hamidreza Aghababaeian, Lara Hamdanieh, and Abbas Ostadtaghizadeh. Alcohol intake in an attempt to fight covid-19: A medical myth in iran. *Alcohol*, 88:29–32, 2020.

[21] Daniel Jolley and Jenny L Paterson. Pylons ablaze: Examining the role of 5g covid-19 conspiracy beliefs and support for violence. *British journal of social psychology*, 59(3):628–640, 2020.

[22] Darrin Baines, Robert JR Elliott, et al. Defining misinformation, disinformation and malinformation: An urgent need for clarity during the covid-19 infodemic. *Discussion papers*, 20(06):20–06, 2020.

[23] Victoria L Rubin, Yimin Chen, and Nadia K Conroy. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[24] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[25] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. Addressing health-related misinformation on social media. *Jama*, 320(23):2417–2418, 2018.

[26] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

[27] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.

[28] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *Plos one*, 13(4):e0196087, 2018.

[29] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55, 2019.

[30] Yuehua Zhao, Jingwei Da, and Jiaqi Yan. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, 58(1):102390, 2021.

[31] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.

[32] Hu Zhang, Zhuohua Fan, Jiaheng Zheng, and Quanming Liu. An improving deception detection method in computer-mediated communication. *Journal of Networks*, 7(11):1811, 2012.

[33] Sanket Mhatre and Akhil Masurkar. A hybrid method for fake news detection using cosine similarity scores. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.

[34] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer, 2017.

[35] Sirisha Bojjireddy, Soon Ae Chun, and James Geller. Machine learning approach to detect fake news, misinformation in covid-19 pandemic. In *DG. O2021: The 22nd*

Annual International Conference on Digital Government Research, pages 575–578, 2021.

[36] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big data & society*, 6(1):2053951719843310, 2019.

[37] Dimitrios Katsaros, George Stavropoulos, and Dimitrios Papakostas. Which machine learning paradigm for fake news detection? In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 383–387, 2019.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[39] Masood Ghayoomi and Maryam Mousavian. Deep transfer learning for covid-19 fake news detection in persian. *Expert Systems*, 39(8):e13008, 2022.

[40] A Aggarwal, A Chauhan, D Kumar, M Mittal, and S Verma. Classification of fake news by fine-tuning deep bidirectional transformers based language model. eai endorsed trans. scalable inf. syst. *Online First*, 2020.

[41] Arup Baruah, Kaushik Amar Das, Ferdous A Barbhuiya, and Kuntal Dey. Automatic detection of fake news spreaders using bert. In *CLEF (working notes)*, 2020.

[42] Amir Pouran Ben Veyseh, My T Thai, Thien Huu Nguyen, and Dejing Dou. Rumor detection in social networks via deep contextual modeling. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 113–120, 2019.

[43] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Towards covid-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274:110642, 2023.

[44] Oksana Belova-Dalton. *Spread of Fake News and Conspiracy Theories Leading to Potential Radicalisation during COVID-19 Pandemic: The Case of Telegram. ee*. PhD thesis, Estonian Academy of Security Sciences, 2021.

[45] Hugo Queiroz Abonizio, Janaina Ignacio De Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87, 2020.

[46] Fantahun Gereme, William Zhu, Tewodros Ayall, and Dagmawi Alemu. Combating fake news in “low-resource” languages: Amharic fake news detection accompanied by resource crafting. *Information*, 12(1):20, 2021.

[47] Myunghoon Kang, Jaehyung Seo, Chanjun Park, and Heuiseok Lim. Utilization strategy of user engagements in korean fake news detection. *IEEE Access*, 10:79516–79525, 2022.

[48] Richa Sharma and Arti Arya. Lfwe: Linguistic feature based word embedding for hindi fake news detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24, 2023.

[49] Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. Cross-language fake news detection. *Data and Information Management*, 5(1):100–109, 2021.

[50] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. A covid-19 rumor dataset. *Frontiers in Psychology*, 12:644801, 2021.

[51] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30, 2022.

[52] Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A Ghorbani. The largest social media ground-truth dataset for real/fake content: Truthseeker. *IEEE Transactions on Computational Social Systems*, 2023.

[53] X Staff. Announcing new access tiers for the twitter api - announcements - x developers, Mar 2023.

[54] Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*, 2023.

[55] Fariba Sadeghi, Amir Jalaly Bidgoly, and Hossein Amirkhani. Fake news detection on social media using a natural language inference approach. *Multimedia Tools and Applications*, 81(23):33801–33821, 2022.

[56] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9, 2018.

[57] StatCounter. Social media stats estonia. `https://gs.statcounter.com/social-media-stats/all/estonia`, 2023. Accessed: 2024-05-11.

[58] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[59] Shrutika Chawla, Preeti Aggarwal, and Ravreet Kaur. Comparative analysis of semantic similarity word embedding techniques for paraphrase detection. In *Emerging Technologies for Computing, Communication and Smart Cities: Proceedings of ETCCS 2021*, pages 15–29. Springer, 2022.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[61] Facebook Research. LASER: Language-Agnostic SEntence Representations. `https://github.com/facebookresearch/LASER`, 2023. Accessed: 2023-05-08.

[62] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

[63] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[65] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[66] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[67] Junaed Younus Khan, Md Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, 2021.

[68] Hugging Face. BERT - hugging face transformers documentation, 2024. Accessed: 2024-05-11.

[69] Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 454–462, 2020.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Li Merila**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, **Cross-Lingual Misinformation Detection: Aligning English and Estonian Fake Health News**,
   (title of thesis)

   supervised by Uku Kangur and Roshni Chakraborty.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Li Merila
15/05/2024