UNIVERSITY OF TARTU
Institute of Computer Science
Data Science Curriculum

Liina Vesilind

Understanding Toxicity of Estonian Politicians Facebook Posts Comments

Master Thesis (15 EAP)

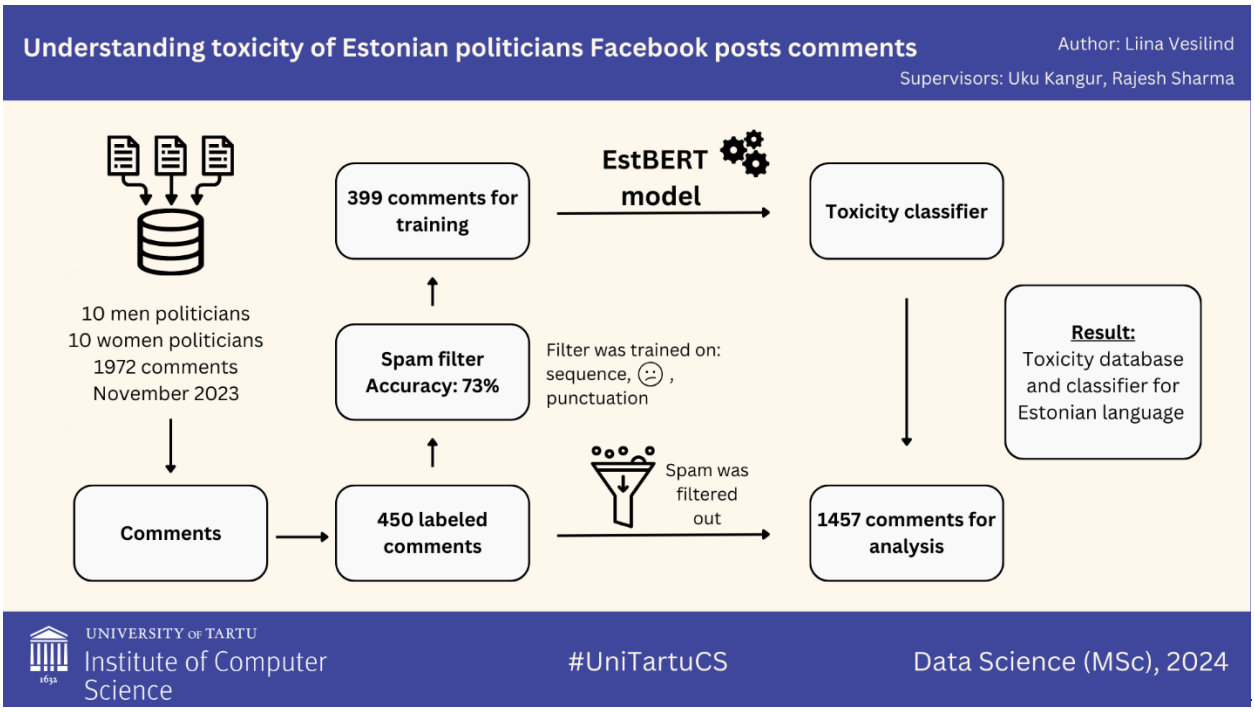
Supervisors: Uku Kangur MSc,
Rajesh Sharma PhD

Tartu 2024

Understanding Toxicity of Estonian Politicians Facebook Posts Comments

Abstract:

Estonian politicians share a lot of information in their social media platforms. This paper analyzes how toxic are the comments on Estonian politicians' official Facebook pages using trained toxicity classifier. 10 female and 10 male Estonian parliament politicians' Facebook page comments are collected, analyzed, and compared. Results have revealed that comments on female politicians contain a higher frequency of toxic language compared to those of male politicians. Additionally, this paper introduces an Estonian toxic word database and toxicity classifier to aid in further research on online discourse in Estonia.



Disclaimer: This thesis contains examples of harmful language used for illustration purposes only. These examples do not reflect the opinions or beliefs of the author, and they are not intended to promote or endorse any form of harmful language or behavior.

Keywords:

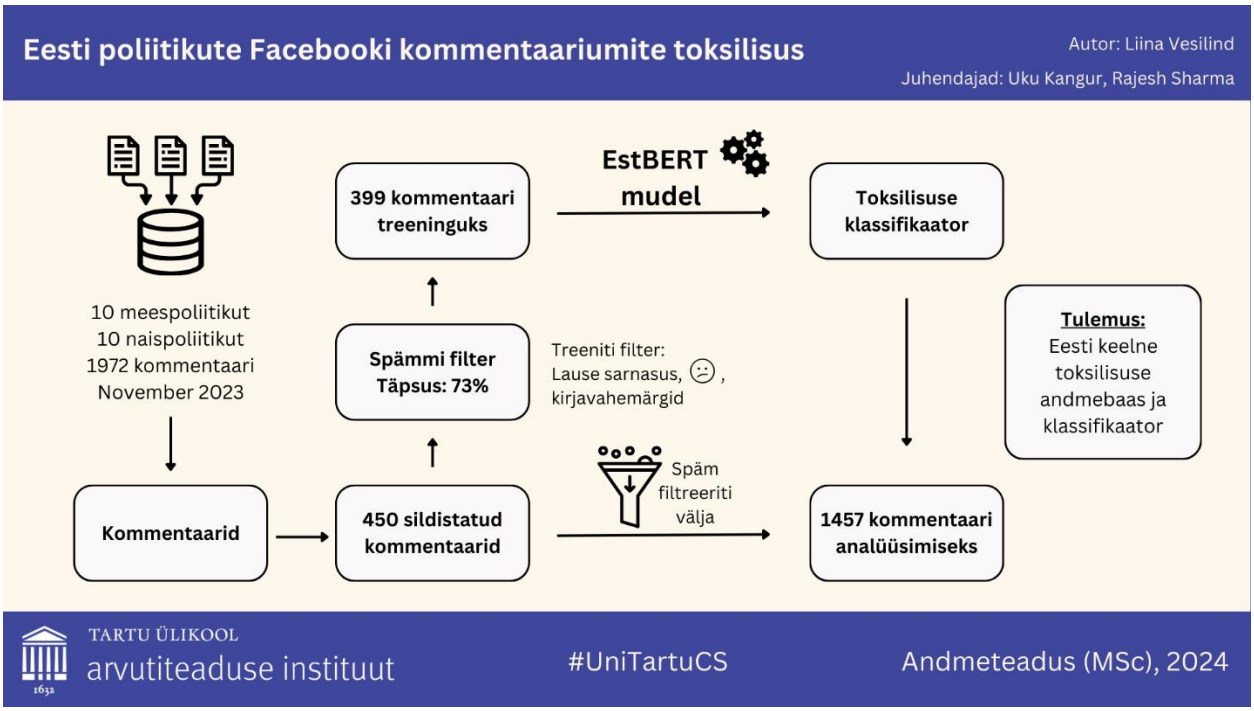
Toxicity Classification, Social Media Analysis

CERCS: P170 - Computer science, numerical analysis, systems, control

Eesti poliitikute Facebooki kommentaariumite toksilisus

Abstrakt

Eesti poliitikud jagavad palju teavet oma sotsiaalmeedia platvormidel. Käesolevas uurimuses analüüsitakse, kui toksilised on kommentaarid Eesti poliitikute ametlike Facebooki lehekülgede postituste all, kasutades toksilisuse klassifikaatorit. Uuringus koguti, analüüsiti ja võrreldi 10 naissoost ja 10 meessoost Eesti parlamendi poliitiku Facebooki postituste kommentaare. Tulemused näitavad, et kommentaarid naispoliitikute postituste all sisaldavad rohkem toksilist keelt võrreldes meespoliitikute postituste all olevate kommentaaridega. Lisaks tutvustatakse käesolevas uurimuses Eesti toksiliste sõnade andmebaasi ja toksilisuse klassifikaatorit, mis aitavad kaasa edasistele uurimustele Eesti veebisuhtluse valdkonnas.



Hoiatus: See lõputöö sisaldab näiteid toksilistest kommentaaridest, mis on esitatud vaid illustreerival eesmärgil. Need näited ei kajasta autori arvamusi ega tõekspidamisi ning nende eesmärk ei ole propageerida ega toetada mis tahes vormis kahjulikku keelekasutust või käitumist.

Võtmesõnad:

Toksilisuse klassifikaator, Sotsiaalmeedia analüüs

CERCS: P170 - Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Table of Contents

Introduction

Social media platforms such as Facebook, Instagram, X (previously known as Twitter) and other similar platforms are one of the most popular channels for people to connect with others, to share information, communicate, and build their social presence. With the social media benefits, there is also a downside such as harassment, cyberbullying, hate speech and online trolling. Disagreements and different opinions within online space may often lead to toxic environments, where discussion participants may experience discomfort or harm.

Defining the concept of toxicity is hard as it can vary in different situations and can also include different aspects. Oxford Learner's Dictionaries define toxicity as something that is very harmful or unpleasant (OxfordLearnersDictionaries.com, n.d.), meaning in online space it can be seen as inappropriate or harmful language usage. Fan et al. (2021) in their research paper describes toxic behavior as spreading negativity or hateful content to other people, who may find it affecting them negatively or badly. Usually, it is done in discussions or posts that other people can see and read. As social media provides anonymity, it makes it easier to troll and spread toxic and negative content.

Kim et al. (2021) in their research paper that studies online comments toxicity, has defined toxicity as being disrespectful toward others or specifically „*as those expressing disrespect for someone by using insulting language, profanity, or name-calling; by engaging in personal attacks; and/or by employing racist, sexist, and xenophobic terms*." This is also similar to Sydnor (2019) and Coe et al. (2014)'s definition of being disrespectful and not civil toward others.

Language toxicity can be determined by how many rude, disrespectful, hateful words a text itself contains and can be found using different Natural language processing tools. This, however, does not always mean that the comment itself is toxic (i.e. usage of swore words) as in some cases the toxic words are used to amplify the meaning of a text or show some characteristics or a feeling of something and is not considered toxic (Xia et al., 2020).

However, despite scholars' efforts on defining or measuring toxicity, it is still very nuanced and difficult to fully grasp. Sheth et al. (2021) introduces a framework for online toxicity detection and highlights the importance of including context, individual and community features to the analysis.

In this paper toxicity is primarily defined using Kim et al. (2021) definition, where **toxicity is defined as text containing disrespect, name-calling, or the use of terms that are racist, sexist, or xenophobic.** However, given the focus on the Estonian language, we have also included Estonian-specific swear words and other hurtful language.

This thesis uses comments from 10 Estonian male and female politicians Facebook posts collected in December 2023 to examine the level of toxicity in the comments and to identify potential differences. The comments are scraped starting from 03.03.2023 to 9.12.2023, however the majority of the posts were made in November and December 2023. To tackle the problem, we have proposed the following research questions:

RQ1: What differences exist in the word usage in Facebook comments on posts made by male and female politicians?

RQ2: How does the frequency of toxic language differ between comments on posts made by male and female politicians?

RQ3: To what extent do commenters exhibit gender bias, discriminatory language, or name-calling (labeling) in the male and female politician posts comments?

To tackle the proposed research questions this paper presents an extensive pipeline for the analysis of toxic comments. The pipeline includes several steps, beginning with collecting the comments then cleaning and preprocessing the comments to remove spam and any other irrelevant content (like comments in other languages, etc.). The pipeline also includes using natural language tools such as EstBERT to train a classifier to detect Estonian language specific toxic language patterns and extracting toxic words. Finally analyzing the word and bigram frequencies within the comment's datasets.

This paper also contributes further to Estonian language processing tools as it establishes:

1. New Estonian language toxicity database including a collection of toxic words from a social media platform such as Facebook.
2. Estonian language toxicity detection classifier, leveraging state-of-the-art natural language processing techniques to identify toxic comments.
3. Toxicity analysis pipeline, proposing a systematic identification and classification of toxic language within Estonian textual data.

This thesis has used AI for the purpose of idea generation, fixing code syntax errors and wording suggestions.

The thesis is structured around four main parts. Firstly, an overview of how to determine and classify toxicity using natural language methods are presented in the related works section (section 1). Following this, the data collection, cleaning, and preprocessing is introduced in the data section (section 2). In the methodology section (section 3), the focus shifts to the utilization of the EstBERT language model for classifier training. Lastly, the results are showcased, accompanied by answering the proposed research questions.

1. Related Works

In the related works section, previous works on toxic comment and content detection tools are given. In the second part, tools detecting spam are described. As this thesis focuses also on differences in word usage in comments made under male and female posts, gender bias in online discussions is also provided.

1.1 Toxic comments and content detection

Comments are textual content and are viewed as a natural language processing task. There have been many different approaches to tackle the challenging task of detecting toxic comments or content in the online environment.

One of the easiest methods and most used in online channels are moderators, which are widely used to moderate the communication of one channel with the help of different text classification tools (Nobata et al., 2016).

One of the widely used approaches to eliminate or find toxicity in text is to compare the text using word or text matching or dictionary-based toxic text detection systems or blacklists of commonly used expressions, where the text is compared against the list of words/expressions that are deemed to have toxic traits. This method, however, can be easily bypassed by writing the words incorrectly or differently without losing the meaning. This is also relevant in current online presence as people want to express themselves quickly and shorten, misspell, or modify the texts, without losing the meaning (Wang et al., 2021).

Another widely used approach to detect toxic text is sentiment analysis where the models identify the core opinions, topics and other information and classify whether the textual content presents toxic traits or not. In the beginning sentiment analysis models used statistical (naïve bayas, Fleiss' Kappa) and grammatical tools (Bag of Words) to detect different sentiments, mainly to categorize texts into positive, negative, and neutral (Kwok & Wang, 2013). These approaches, however, were not as sufficient as Kwok and Wang (2013) found in their research on classifying anti-black tweets and stated already then, that algorithms need to include more features to accurately detect toxicity.

Diego Reforgiato Recupero, Consoli, S., Gangemi, A., Andrea Giovanni Nuzzolese, & Spampinato, D. (2014) in their paper included semantic web technologies in their sentiment analysis models, an addition to previously presented models, to include language specific semantics to better the algorithms. Their model was additionally able to identify topics, subtopics, and opinions as well as semantic sentiment. Diego Reforgiato Recupero and Mauro Dragani (2016) explored the semantic features usage further and proved that semantic features in combination with Machine Learning, Natural Language Processing approaches such as Semantic Computing have higher performance compared to simple sentiment analysis.

Nobata, et al (2016) used a supervised classification method using Natural Language Processing features such as N-gram with 3-5 characters (inc. spaces), linguistic features such as the average length of the comment or word, number of letters, punctuation, etc. They also used syntactic features such as parent/grandparent of node, POS (part of speech) to catch the dependencies that other features may miss. They also combined embedding features with the abovementioned NLP

features, which showed good results in noisy data sets. As data is becoming more noisy and more specific, manual feature engineering or finding unique triggers to detect toxic comments or toxic sentiment in the natural language is becoming more challenging, hence neural network methods present to be more accurate and effective in feature learning (Zhang & Luo, 2019).

Deep learning approaches need a lot of labeled data to train and learn, hence it is broadened to have multiple tasks simultaneously. Elnaggar, A., Bernhard Waltl, Glaser, I., Jörg Landthaler, Scepankova, E., & Matthes, F in 2018 used a deep learning model to translate and classify comments at the same time. They also identified the imbalanced toxic comment datasets to be most challenging in classifying toxic comments.

Many models have their fair share of errors and challenges hence, van Aken et al., (2018) proposed a toxic comment learning algorithm with a goal of minimizing those errors. They used logistic regression with n-grams (character and word) and a variety of deep learning algorithms such as recurrent and convolutional neural networks. Long-Short-Term-Memory Network (LSTM) model, that uses the sequence of words as an input and makes a prediction for multi-class classification and convolutional neural network model that detects the combination of features in a text to find information. They train word embeddings on large datasets to capture the information that has been missed. They use popular tools for word and sub word embedding such as Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), to eliminate the influence of misspelled, modified, or compendious words. The proposed learning algorithm is an ensemble that decides, which of the abovementioned classifiers is the most correct one for an individual text, considering the classifiers individual strengths and weaknesses and improve F1-Score when compared to individual classifiers alone. This also shows that easy statistical approaches such as logistic regression with a combination of deep learning approaches work very effectively.

As with the growth of social media platforms, automated toxicity identification has become more and more crucial, the previous sentiment analysis models cannot compete with it. Pavlopoulos, J., Prodromos Malakasiotis, and Ion Androutsopoulos (2017) researched automatic and semi-automatic moderation tools to moderate online content. They used 1.6 million user comments from a news portal and 115 thousand comments from Wikipedia. Their results showed that deep learning approaches such as RNN (recurrent neural networks) outperformed previous moderation tools that employed logistic regression with n-gram features or word character features. The proposed model also highlighted words that can be considered toxic or suspicious for manual revision. These findings are in line with previously stated research that easy approaches do not perform well compared to new unsupervised models.

Zaheri et al. (2020) used LSTM and Naïve Bayes methods as benchmark to classify texts into either toxic or non-toxic. LSTM is usually preferred in NLP tasks compared to for example RNNs (recurrent neural network) as textual tasks are longer, LSTM is more designed to have longer memory. Their research showed that LSTM does very well in classification tasks compared to benchmark methods.

The widely known BERT (Bidirectional Encoder Representations from Transformers) model was firstly introduced in 2018 by Google researchers Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. The pre-trained model uses unlabeled text and text context, which can be used in many tasks including classifying toxicity with the fine-tuned model. Pavlopoulos et al (2020) used LSTM context-insensitive classifier with feed-forward neural network and BERT in their research and found that there is no evidence that the context in which the comment is written has any effect in the performance of identifying the toxicity classifiers. BERT model is also trained for Estonian language on Estonian cased corpus and is called EstBERT and is finetuned on natural language processing tasks such as part of speech (POS), text classification and much more (Tanvir et al., 2021).

Most of the state-of-the-art models are meant for English language, however with the introduction of BERT, the model is also trained now for many languages, such as CamemBERT for French (Martin et al., 2020) or Finnish BERT for Finnish language (Antti Virtanen et al., 2019).

There have also been many additions to the BERT model apart from training it for different languages. RoBERTa, an optimized BERT model was introduced in 2019 (Liu et al., 2019) with improved design choices, which resulted in state-of-the-are results compared to other models at the time.

There are now language models designed for different languages and are capable of doing many language specific and natural language processing tasks. The next research indicated that toxicity is better determined with language specific models than translating text to English as it does not give that accurate results.

Kobellarz and Silva (2022) researched how Google Perspective API detected toxicity on a comment dataset that was automatically translated from Portuguese to English and then compared to baseline and human annotated datasets. The results showed that original text is superior as high toxic comment can be reduced to lower toxic levels as translations usually capture only generic language characteristics in classification tasks (Chen et al., 2019).

Eskelinen et al. (2023) studied how Finnish toxic language can be detected. They used an English Jigsaw and a Finnish dataset, which were manually annotated. The Jigsaw dataset was machine translated to the English language. The results showed that machine translations are usable in translating texts to detect toxicity when there is no annotated data available.

For the English language one state-of-the-art toxic content classifier is Google's Perspective API, that uses a character-based Transformer (soft gradient-based sub word tokenization module) (Tay et al., 2021) framework to detect toxic content. In their paper Lees et al., (2022) compared multiple models (Perspective API, customized BERT, Multilingual T5 (mT5) and Unified Toxic Content Classification (UTC)) for multilingual toxic content classification. They experimented on 12 different languages and proposed a new toxic classifier model that is developed for Perspective API.

The natural language processing area is rapidly developing with new state-of-the-art models being introduced for various language processing tasks in various languages very quickly. However, there is not one size fits all approach for NLP tasks and the appropriate method for a specific task

must be found therefore in this paper's analysis part the focus is on tools made for Estonian specific language.

1.2 Spam filtering

Text and image-based spam filtering became a problem with the rise of emails. People used different techniques to specially tailor filters on emails for spam. In their paper they used similarity matching techniques and IP address matching as an example (Goodman et al., 2007). This method however required a knowledge of how the spam email should look beforehand, so the filter can identify the emails as spam.

In essence spam filtering is a text categorization problem which can be tackled with algorithms. One way the filter can work is to separate the email to words, then tokenize the email body, lemmatize the words and remove stop words. Then use a spam classifier to classify it as spam or not. Statistical approaches like Naive Bayes initially introduced by Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. in 1998, Support Vector Machines (SVM), Logistic regression and hybrid methods were used for spam categorization tasks on emails. All these methods, however, worked if the spam stayed the same. When new terms or additional features were to be added, the model had to be rebuilt again (Guzella & Caminhas, 2009).

As time moved forward spam content moved from emails to the internet. Platforms like YouToube, Twitch, Instagram, etc., online shops with places for public reviews, all have a section (comment, post a review, etc.), where people can write whatever they want, hence a perfect place for posting spam content. In their 2015 paper about the comment section on YouTube, the authors compared multiple state-of-the-art classification techniques (logistic regression, decision tree, Naïve Bayes, random forest, and Gaussian and linear Support Vector Machines) at the time on automatically classifying YouTube comments. Their results showed that compared to statistical methods their method achieved 90% accuracy. Automatic comment section moderation, however, remained to be a challenge (Alberto et al., 2015).

Enaitz Ezpeleta et al, in their 2018 paper improved YouTube spam filtering by adding a mood analysis (new feature) on top of Alberto et al. (2015) dataset to improve statistical analysis techniques. The results showed improved accuracy scores and lower false positives. In different paper by Enaitz Ezpeleta et al, in the same year (2018) also added sentiment analysis and personality recognition on another YouTube datasets (O'Callaghan et al., 2021). The results were also improved in both accuracy and false positives. This shows that mixing multiple techniques improve the results as they capture different nuances in textual content.

Regarding spam content on Facebook, previous research has been more focused on eliminating or identifying the spam accounts rather than filtering or tagging spam comments on Facebook posts. Fahim and Naseem (2015) in their paper proposed a methodology where posts are analyzed by Artificial Neural Network to look at posting patterns, keywords related to spam and previous Facebook activity to detect spamming users, however this method does not only include hurtful spam content but also other spam like links, requests, messages, game promotions, etc.

In the 2022 paper by Hakim Azri, Hafida Belbachir and Fatiha Guerroudji Meddah investigated identifying spammers, who interact with Facebook public pages. Their focus was not only

identifying individual spammers but also identifying groups of spammers that may use similar URLs. They used a scoring method by manually identifying relevant textual features and characteristics (i.e. overuse of uppercase letters, comments in another language, using URLs and email addresses in comments, special characters and blacklisted words and expressions) on selected Facebook spam comments in addition to features from previous literature. They found that potential spam accounts activity patterns differed from non-spam accounts. Spam accounts appear to be more active at night (from 00:00 to 04:00) and less active in normal working hours (from 12:00 to 16:00). They also stated that potential spammers do not interact with other spammers.

1.3 Gender bias towards politicians

From previous research on word usage and descriptions of male and female in textual format have found that textual representation does have gender and ethnical stereotypical bias, where men are usually described more with words related to skills or work and females usually with emotions, appearances, or labor (Rudinger et al., 2017).

Gender bias on comments directed towards female politicians and public figures were researched by Field A. and Tsvetkov Y. in 2020 using a classifier that can predict gender of the addressee of the comment. Papers results are in line with Rudinger et al., (2017) results with female public figures are described using appearance and sexualized words. For female politicians comments results however have a mix of domestic words such as family, love, and spouse, but also words related to power and strength. Mertens et al. (2019) looked at Tweets during German elections in 2017 and found textual semantic evidence that Tweets toward female politicians were more targeted with gender related language rather than competence or profession related language.

Marjanović et al. in 2022 researched gender bias in politics by analyzing 10 million comments on Reddit. They used linguistic analysis to determine gender bias in comments directed towards politicians. They found that comments directed toward female politicians tend to be shorter and use more given or full name compared to usage of surname, which was more prominent in comments towards male politicians.

Looking at troll or spam comments on Twitter (now known as X) Pnina Fichman and Maanvi Rathi in their 2022 paper found that troll comments are more likely on tweets made by female than by male. In their research they compared American and Indian troll comments.

Based on the previous research it is evident that gender bias in textual format, which can be comments, articles, or texts, still exists toward politicians. Female politicians are usually described with words related to appearances and traditionally domestic words. It is also worth mentioning that they are referred to with a given name or full name, which can be considered as a more familiar reference compared to men, who are referred to mostly by surname and described with words related to occupation or competence.

2. Data

This section includes information about the data used in the thesis. Data collection, cleaning and preprocessing is described. In the end a small overview of the data is given in the general statistics of the data section, including the average lengths of the comments, most frequent words in both men and female political posts comments and comments overall sentiment overviews are given.

This thesis analyses Estonian male and female politicians, focusing on 10 male and 10 female politicians and were chosen from XV Parliament composition as of 30.03.2023 (Riigikogu, 2023). Preliminary analysis was conducted on the entire parliament composition list to check the frequency of people's postings on Facebook.

Following the initial analysis, a finalized list included 10 male and 10 female parliament politicians, covering all 6 parliament represented parties, post at least 1-2 times per week and own a Political Facebook page (Table 1).

Name	Party	FB profile name	Minister
Yana Toom	Eesti Keskerakond	yanatoom.ee	
Ester Karuse	Eesti Keskerakond	karuseester	
Kert Kingo	Eesti Konservatiivne Rahvaerakond	KertKingoEKRE	
Kaja Kallas	Eesti Reformierakond	kallaskaja	Prime Minister
Signe Riisalo	Eesti Reformierakond	signe.riisalo.1	Minister of Social Protection
Annely Akkermann	Eesti Reformierakond	akkermannannely	
Liisa-Ly Pakosta	Erakond Eesti 200	liisapakosta	
Riina Solman	ISAMAA Erakond	solmanriina	
Marina Kaljurand	Sotsiaaldemokraatlik Erakond	marinakaljurand	
Riina Sikkut	Sotsiaaldemokraatlik Erakond	riinasikkut.ee	Minister of Health
Mihhail Kõlvart	Eesti Keskerakond	MihhailKolvart	
Jüri Ratas	Eesti Keskerakond	ratasjuri	
Jaak Valge	Eesti Konservatiivne Rahvaerakond	ekrejaakvalge	
Henn Põlluaas	Eesti Konservatiivne Rahvaerakond	polluaashenn	
Urmas Paet	Eesti Reformierakond	urmaspaet1	
Jürgen Ligi	Eesti Reformierakond	JurgenLigi	
Lauri Hussar	Erakond Eesti 200	LauriHussarEesti200	
Hendrik Johannes Terras	Erakond Eesti 200	hendrikjohannesterras	
Priit Sibul	ISAMAA Erakond	ristteesibul	
Lauri Läänemets	Sotsiaaldemokraatlik Erakond	l22nemets	Minister of the Interior

Table 1. Politicians included in the research.

The six parliament parties are: Estonian Reform Party (Eesti Reformierakond), Conservative People's Party of Estonia (Eesti Konservatiivne Rahvaerakond), Estonia 200 (Eesti 200), Social

Democratic Party (Sotsiaaldemokraatlik Erakond), Fatherland (ISAMAA Erakond) and Estonian Centre Party (Eesti Keskerakond).

Some popular or well-known politicians did not make the list due to following reasons: they post from their personal profile (Helle-Moonika Helme), repost everything from their personal profile (Urmas Reinsalu), Facebook page does not have a distinct name (Martin Helme-eestiesikonservatiiv) and therefore cannot be scraped with the facebook_page_scraper.

2.1 Data Collection

Data was collected from scraping Facebook pages using facebook_page_scraper packages over the period of 1 week in the beginning of December 2023. The posts and comments scraped start from 03.03.2023 to 9.12.2023, however the majority of posts were made in November and December 2023, with few posts also scraped from spring as some politicians did not post during the summer (Henrik Johannes Terras) and early autumn. Data was analyzed using Python Jupyter Notebook and Google Collab environments.

2.2 Data preprocessing

The data scraping process involved the extraction of individual Facebook posts along with associated metadata, such as commenter username, comment timestamp, reactions, post content (text), and all the comments under the post.

Initially all comments were in politicians' individual files, however, were later divided into two separate datasets. One including comments on posts made by female politicians and other including comments on posts made by male politicians. These datasets included only a unique comment id number (comment_id) and the comment text (comment_text). Both excel files were manually examined and comments in languages other than Estonian, particularly Russian/English, as well as website links/gifs/picture links, were removed. Later the two datasets were also merged for further analysis.

From the women and men dataset a sample of 450 comments (200 women and 250 men post comments) were extracted and manually labeled for toxicity (-1 positive, 0 neutral, 1 lightly toxic, 2 toxic, 3 very toxic, X for incoherent and S for spam) and stance (1 agreeable, 0 neutral, 1 difference, X for incoherent and S for spam). Labeling was done by 9 different individuals to ensure more objective labels on the data. The description of the labels can be found in Appendix 1.

2.3 General statistic on the data

General statistics on the whole data set are provided in Table 2. The table includes the number of comments on men and female politician posts separately and together. Additionally, it includes an overview of the average number of words and average length of the sentences. Analysis of most common words and are firstly shown without eliminating any information and secondly with eliminating stop words, punctuation, and words are lemmatized using EstNLTK package tools. For stopwords Kristel Uiboaed Estonian Stopword (2018) stopword collection was used. Additionally, most popular verbs, nouns and adjectives are listed to establish a baseline for comparison with toxic language later in the research. Short sentiment analysis is also provided,

where comments were categorized into 3 categories (neutral, positive, and negative). Sentiment analysis was done using tartuNLP/EstBERT128_sentiment tool.

Men		**Women**		**Total**	
Comments	803	Comments	1169	Comments:	1972
Average number:		Average number:		Average number:	
Words	24.4	Words	37.6	Words	32.14
Sentences	2.27	Sentences	2.97	Sentences	2.69
Most common words		**Most common words**		**Most common words**	
Inc. all	w/o stopwords/punctation and lemmatized	Inc. all	w/o stopwords/punctation and lemmatized	Inc. all	w/o stopwords/punctation and lemmatized
.: 1676	Eesti (Estonia): 91	,: 3057	Tahtma (to want): 230	,: 4091	Tahtma (to want): 263
,: 1034	Inimene (person): 65	.: 2236	Väga (very): 178	.: 3912	Eesti (Estonia): 257
on: 430	Palk (salary): 619	on: 1169	Või* (or): 171	on: 1599	Või* (or): 222
?: 187	Kõik (all): 60	?: 291	Võtma (to take): 170	?: 478	Väga (very): 211
!: 173	Õpetaja (teacher): 59	Teie (you): 266	Andma (to give): 168	!: 405	Võtma (to take): 210
": 73	Riik (country): 57	Mul (I have): 240	Eesti (Estonia): 166	Teie (you): 291	Kõik (all): 208
Pole (not): 73	Pere (family): 53	!: 232	Ära (don't): 151	Pole (not): 276	Ära (don't): 198
Minu (mine): 62	Või* (or): 51	Seda (that): 211	Kõik (all): 148	Minu (mine): 255	Andma (to give): 196
Eesti (Estonia: 57	Raha (money): 51	Pole (not): 203	Ühendus (association): 125	Mul (I have): 252	Inimene (person): 185
Oli (was): 50	Teadma (to know): 48	Mind (me): 20	Laps (child): 124	Seda (that): 248	Riik (country): 166

Table 2. General statistics on the data, *Või* can mean either 'or' or butter.*

The dataset includes 1972 comments. with women's post comments accounting for approximately 60% (1169 comments) and male post comments constituting 40% (803 comments). Women's post comments include on average more words and more sentences. A preliminary assessment suggests that the comments are longer as they include more spam/phishing content. The data was collected during the period where teachers' salary discussions were in the center of attention (Rohemäe, 2023), which is also seen from the most popular words in the comments as they include words like Eesti (Estonia), õpetaja (teacher), palk (salary), laps (child), pere (family).

From the Table 3 below the same themes emerge. The words presented are in lemmatized form.

Men			**Women**		
Noun	**Verb**	**Adjective**	**Noun**	**Verb**	**Adjective**
Inimene (person)	Teadma (to know)	Hea (good)	Inimene (person)	Tahtma (to want)	Hea (good)
Õpetaja (teacher)	Võima (can)	Ilus (beautiful)	Laps (child)	Võtma (to take)	Raske (hard)
Palk (salary)	Võtma (to take)	Tubli (good)	Ühendus (association)	Andma (to give)	Järgmine (next)
Riik (country)	Nägema (to see)	Suur (big)	Summa (sum)	Saatma (to send)	Haige (sick)
Raha (money)	Elama (to live)	Kinnitatud* (confirmed)	Sõnum (message)	Elama (to live)	Elav (lively)

Table 3 Most popular nouns, verbs and adjectives in men and women post comments.

The sentiment distribution among comments on posts by male and female politicians exhibits a similar pattern. Approximately 23% of the comments are classified as positive (22% for women

and 26% form men posts). Some of the positive sentiment examples are “Edu ja jõudu kõigile!” (“Good luck and strength to all!”), “Teie perele ka ilusat jõuluootust!” (“For your family, too, have a nice Christmas!”).

Around 11% of the comments are categorized as neutral with 12% for women and 10% for men posts. Neutral sentiment was given to comments like “Ma isiklikult ei toeta ka erakorralisi valimisi. Pigem peaministri umbusalduses tuleks hääled kokku saada” (“Personally, I do not support extraordinary elections either. Rather, the prime minister's censure should get the votes together.”) and “kas siis palga nr oli oluline?” (“So did the salary number matter?”).

The majority making 66% of the comments are classified as negative and accounts for 67% for women and 65% for the men comments. Some of the comments that are classified as negative are “Käi vittu värdjas, vaata peeglisse paks siga” (“Fuck off bastard, look in the mirror you thick pig”) and “No aga kui FB võtab tõesti aluseks meie seadused, siis on ju kõik hästi, lihtsalt Teie jutt võibki olla vastuolus sõnavabadusega?” (“Well, if FB would actually be based on our laws, then all would be well, it's just that what you're saying might be contrary to freedom of speech?”).

Automated sentiment tools served the purpose of providing a general overview of sentiment trends within the dataset. We use it later as a baseline in comparing toxic classifications.

Sentiment distribution (Figure 1) between women and men politician comments are similar with around two thirds of the comments being negative, around 20%-25% being positive with around 10% of the comments being neutral.

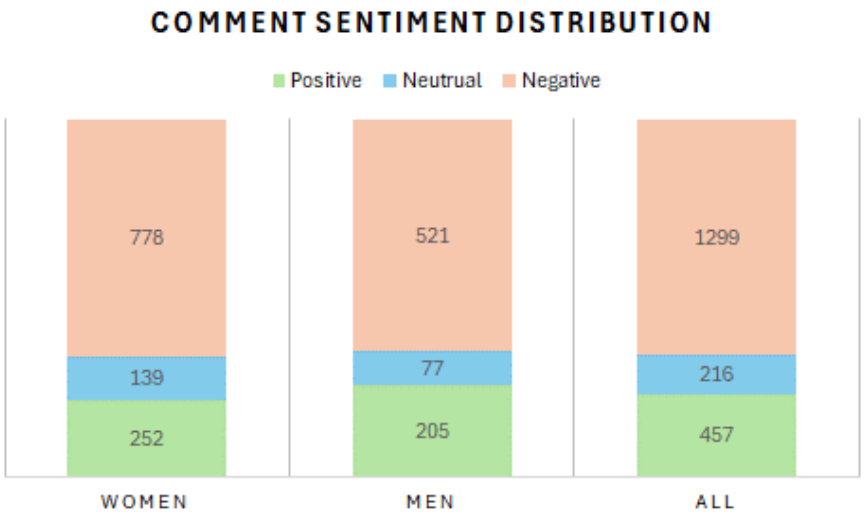


Figure1. Sentiment distribution on the women, men, and whole dataset

3. Methodology

In the methodology section, analysis pipeline is introduced. Firstly, spam elimination methods are described with the results of applied 3 methods. Secondly, EstBERT model is introduced with a toxicity classifier.

3.1 Analysis pipeline

Eliminating spam comments

Based on spam filtering techniques used in other papers like similarity matching techniques (Goodman et al., 2007) or features related to spam content (i.e overuse of uppercase letters, comments in another language, using URLs and email addresses in comments, special characters and blacklisted words and expressions) (Aziri et al., 2022), 3 methods (sequence matching, number of emojis and same sequence of punctuation) were tried on manually annotated comments (450 comments in total). It is worth mentioning that in the preprocessing stage of data cleaning comments in other languages, URLs and email addresses as well as URLs to pictures and GIFs were eliminated, hence these spam filtering methods were not considered.

The first method for spam filtering involved sequence matching using the SequenceMatcher from the difflib module. In essence it compares two sequences and identifies similarities between them. It calculates a similarity ratio indicating how alike the sequences are. A ratio of 1 signifies identical sequences, while lower ratios denote less similarity.

In this analysis a similar sequence indicates spam as phishing and spam comments were very similar in manual examination. This method compares the similarity between sentences and marks those exceeding a set threshold as spam. Initially, a similarity threshold of 0.7 was used, indicating that sentences with a similarity ratio above this threshold were deemed potential spam and filtered out. Additional thresholds of 0.9, 0.8, and 0.75 were also experimented with, but after manual examination of the results, the threshold of 0.7 appeared to be the most effective with the most logical spam filtered out. Different experiment results are not provided.

The second method used the number of emojis used in a comment. This method used Regular Expression module in Python. When the comment included 3 or more emojis then it was to be filtered out as spam. This approach assumed that spam comments have a lot of emojis. Normally people use neither or 1 to 3 emojis in a comment to express themselves.

The third approach used the same sequence of punctuation in a comment as spam text is usually the same hence has the same sequence of punctuation. Comments that have more than 4 same punctuation sequences were potentially filtered out.

The results of all 3 spam filters (Table 4) were compared and the results are presented in the table below. The sequence spam filter demonstrated the most accurate results compared to human labeling. Additionally, the sequence filter outcome also covered the results of emoji and punctuation filters hence did not offer any further improvement of the results. After manually evaluating the sequence filter results it deemed acceptable to use the filter on all the data and filter out spam comments.

	Sequence	Emoji	Punctuation	Human
# labeled as spam	68	10	13	51
# labeled as X (irrelevant)				45
same as human (s)	37	4	9	
same as human (x)	13	2	1	
Accuracy (s)	73%	8%	18%	
Accuracy (x)	29%	4%	2%	

Table 4. Method accuracy compared to human labeling.

Clean comments dataset includes 1457 comments ca 74% of the original dataset (515 comments were filtered out). Final dataset included 697 men posts comments and 760 female posts comments.

3.2 Toxicity classifier using EstBERT

For toxicity classifier, first EstBERT (TartuNLP/EstBERT · Hugging Face, n.d.) was used as it is one of the most state-of-the-art natural language models specifically trained for Estonian language.

From manually labeled data (450 comments), comments labeled as spam (51 comments) were removed, leaving 399 comments for classifier training (Table 5).

	Comments	Percentage
-1 (positive)	52	13%
0 neutral	139	35%
1 (lightly toxic)	119	30%
2 (toxic)	38	10%
3 (very toxic)	6	2%
5 (old X or irrelevant)	45	11%
Total	**399**	**100%**

Table 5. Results of manually labeled data divided into different classes based on toxicity.

For the first binary classifier comments labeled as 1, 2 ja 3 were all taken as toxic in training (hereinafter referred to as Broad Model). As the proportion of the comments labeled as toxic is small, we encountered a similar problem to Elnaggar et al. (2018) with imbalance data, however instead of using multi-task approach, the toxic comments in this thesis were oversampled during the training process.

Training was done using the following parameters: learning rate 0.0001, 15 epochs, batch size of 128 for both training and evaluation per device, with results saved to the "./results" directory. The output directory was overwritten, and models were saved after each epoch. Additionally, mixed precision training (fp16) was utilized.

The training process obtained metrics as follows: precision: 0.745, recall: 0.864, F1-score: 0.8 and accuracy: 0.8.

Running the classifier on all the other comments resulted in 488 comments (33% of the total comments) as toxic (235 comments from men politician posts (48%) and 253 comments from female politician comments (52%) and 969 comments (67% of the total comments) as non-toxic (462 (47%) comments from men politician posts and 507 (53%) comments from female politician comments. The distribution of men and female posts comments are roughly the same.

For the second variance the same process was done, but comments labeled as 2 and 3 were labeled as toxic, so the classifier would learn to better classify relatively toxic comments (hereinafter referred to as Strict Model). This was done to decrease the probability of false negatives. The second variant of the model had the same preprocessing i.e oversampling in the toxic class and training model had the same parameters as previously. Due to the high number of epoch and heavy oversampling the model became very overfitted as all 4 metrics: precision, recall, F1-score, and accuracy were 1.0. To solve this problem a random sample of 100 comments were taken out and manually labeled to compare the predicted labels and this method adjusted the training metrics as follows: precision: 0.5, recall: 0.33, F1-score: 0.6 and accuracy: 0.96.

Running the second classifier on the comments resulted in 75 comments (5% of the comments) classified as toxic (37 comments from male and 38 comments from female politician posts) and 1382 (95% of the comments) as non-toxic (660 (48%) comments from male and 722 (52%) comments from female politician posts). Also, here the distribution of men and female posts comments stays roughly the same.

3.3 Ngram and sentiment analysis

Ngram analysis is done using EstNLTK tools for ngram extractions (Estnltk, n.d.). For the analysis 2-word ngrams (bigram) is used as some Estonian toxic language have 2-word phrases, hence bigrams are used for the analysis. 3 or more ngrams were not considered as the dataset is small and toxic comments distribution was small only 33% for Broad Model and 5% for Strict Model.

For the sentiment analysis TartuNLP/EstBERT128_sentiment analysis tool was used (TartuNLP/EstBERT128_sentiment · Hugging Face, n.d.), which is finetuned EstBERT model. No additional adjustments were made to the sentiment model.

3.4 Estonian language toxicity database

Estonian toxic language list was done by filtering out words from nontoxic comments that, when present in toxic comments, contributed to their toxicity (i.e only unique words to toxic comments remained) This is done for both models. After the elimination both lists were manually overlooked and all words that are not directly toxic or swear words were eliminated, leaving only words that are as itself toxic or swear words.

Some examples of words that in the right context could be toxic, but as itself are not ‘pankurid’ bankers, words related to nationality or racial words like ‘slaavlased’ Slavs, ‘mustlased’ gypsies, ‘venelased’ Russians, ‘moslemid’ Muslims, etc. Also, some animals like pig or mole ‘mutt’, that in the right context can mean many things including mean and toxic meaning, however as a word itself is not associated with negative meaning were also eliminated from the toxic dataset.

4. Results and discussion

In the results and discussion part, the results of the 2 trained toxicity classifier models are presented in form of most frequent words between models and comments from male and female politician comments. The same comparison is made also for bigrams and most used verbs, nouns, and adjectives. A result of small sentiment analysis is also provided for the comments classified as toxic by the models. At the end of each section results discussion is provided.

4.1 Word frequencies in toxic and non-toxic classes

Firstly, word frequencies in toxic and non-toxic classes using both trained classifiers (Broad Model and Strict Model) are presented and compared. All comments classified as toxic by Strict Model were also classified as toxic by Broad Model.

From Table 6, the words distribution between Broad Model and Strict Model shows distinct word patterns in both toxic and non-toxic classes. Several words appear frequently in both toxic and non-toxic categories. For instance, "Eesti" (Estonia), "Inimene" (person), and "Riik" (country) are among the most common words in both categories. This shows that these words are in a wide range of comments and are expected in the Estonian political discussions.

Interestingly in Strict Model words such as "Rahvas" (nation) and "Valitsus" (government) are presented in the toxic word category. The government being the target of toxicity is expected as the parties presented in the government hold lower popularity ratings out of the Estonian parties in the period reviewed (Erakondade Reitingud - Iganädalane Erakondliku Eelistuse Küsitlus, n.d.).

When looking at the non-toxic word distribution, Broad Model has frequent occurrences of words like "Hea" (good) and "Õpetaja" (teacher). As the data was collected from the time when teachers' salaries were a hot topic of discussion (Rohemäe, 2023), it shows that the commenters were using those words rather frequently. The same topic theme continues with the further analysis of word frequences between female and male posts comments. In general, at this point toxic class word frequency does not show any toxic words.

Broad Model		Strict Model	
Full toxic word frequency:	Full non-toxic word frequency:	Full toxic word frequency:	Full non-toxic word frequency:
Eesti (Estonia): 103	Riik (country): 88	Eesti (Estonia): 26	Eesti (Estonia): 162
Inimene (person): 91	Kõik (all): 85	Rahvas (nation): 16	Inimene (person): 153
Õpetaja (teacher): 79	Eesti (Estonia): 85	Riik (county): 14	Riik (country): 150
Palk (salary): 76	Inimene (person): 71	Üks (one): 12	Kõik (all): 148
Riik (country): 76	Hea (good): 62	Raha (money): 12	Õpetaja (teacher): 132
Kõik (all): 75	Õpetaja (teacher): 59	Ära (away): 12	Raha (money): 113
Raha (money): 72	Väga (very): 58	Kõik (all): 12	Või (or): 110
Või (or): 63	Või (or): 56	Inimene (person): 9	Palk (salary): 103
Ära (away): 53	Võima (may): 54	Palk (salary): 9	Võima (may): 90
Üks (one): 47	Raha (money): 53	Valitsus (government): 9	Ära (away): 89

Table 6. Top 10 words in toxic and non-toxic classes for both models. ‘Või’ (or) in this case was not labeled as a stop word but as a noun ‘butter’, which in Estonian language is the same word. However knowing the context, it is the word for ‘or’

Results of Broad Model and Strict Model

Next the word frequencies for both models and both toxic and non-toxic classes were separated into comments under men and female politician posts to examine whether there is any difference. The first table has the words for Broad Model (Table 7) and the second for Strict Model (Table 8).

Analyzing the word frequencies for Broad Model (Table 7), we can observe differences in the word frequencies between people’s comments on posts made by male and female politicians.

For example, for male politician’s posts comments for toxic class have word frequencies higher, with terms like "Eesti" (Estonia), "Palk" (salary), and "Inimene" (person) being among the most frequent compared to female politician comments for toxic class, where "Inimene" (person) and "Eesti" (Estonia) are also presented. Interestingly female posts comments have more words related to money “Raha” (money) is only presented there, however “Palk” (salary) is presented for both.

The most frequent non-toxic words remain consistent across both male and female politicians' posts, including terms like "Eesti" (Estonia), "Kõik" (all), "Inimene" (person), and "Õpetaja" (teacher), but for female politician post comments also Kaja and “Peaminister” (prime minister) is mentioned. Comments were collected during the time Estonian Prime minister was female Kaja Kallas from Estonian Reform Party (Eesti Reformierakond). As Kaja was more mentioned than her surname Kallas, it is in line with Marjanović et al. (2022) study that found that comments directed toward female politicians use more politician’s given name compared to usage of surname.

Men politician posts		Women politician posts	
Full toxic word frequency:	Full non-toxic word frequency:	Full toxic word frequency:	Full non-toxic word frequency:
Eesti (Estonia): 53 Palk (salary): 38 Inimene (person): 34 Õpetaja (teacher): 34 Või (or): 30 Riik (country): 28 Üks (one): 27 Teadma (to know): 25 Kõik (all): 24 Ära (away): 23	Eesti (Estonia): 38 Kõik (all): 35 Hea (good): 35 🙂: 32 Inimene (person): 31 Raha (money): 27 Riik (country): 27 Pere (family): 26 Aasta (year): 25 Õpetaja (teacher): 25	Inimene (person): 57 Kõik (all): 51 Eesti (Estonia): 50 Raha (money): 49 Riik (country): 48 Õpetaja (teacher): 45 Palk (salary): 38 Või (or): 33 Rahvas (nation): 30 Ära (away): 30	Riik (country): 61 Kõik (all): 50 Eesti (Estonia): 47 Inimene (person): 40 Väga (very): 39 Või (or): 35 Õpetaja (teacher): 34 Kaja: 33 Võima (may): 31 Peaminister (prime minister): 31

Table 7. Gender separated word frequencies for Broad Model.

The same analysis was also done for the Strict Model results (Tabel 8).

Analyzing the word frequencies for Strict Model, we can observe differences in the word frequencies between people's comments on posts made by male and female politicians.

Male politicians' posts comments most frequent toxic words include "Eesti" (Estonia), "Üks" (one), and "RE" (Reformierakond). This can be considered a little bit toxic when the context is related to people from the prime minister's party, however the comments also include male posts from that party so it can mean both things.

When looking at the toxic word frequency for female politician's post comments "Kurat" (Devil) in Estonian can be considered as a swear word and as the comments also include people names that are tagged, then after manual evaluation, we removed peoples named from the results, however, we kept names of the politicians.

For non-toxic words the top word frequency remains like first model and do not differ much between male and female politician post comments and include words like "Eesti" (Estonia), "Inimene" (person), "Õpetaja" (teacher), "Palk" (salary) and "Kõik" (all). Here also Marjanović et al. (2022) findings hold true.

Men politician posts		Female politician posts	
Full toxic word frequency:	Full non-toxic word frequency:	Full toxic word frequency:	Full non-toxic word frequency:
Eesti (Estonia): 17	Eesti (Estonia): 74	Rahvas (nation): 10	Riik (country): 101
Üks (one): 10	Inimene (person): 60	Eesti (Estonia): 9	Kõik (all): 97
RE (Reformierakond): 8	Õpetaja (teacher): 55	Riik (country): 8	Inimene (person): 93
Kõik (all): 8	Kõik (all): 51	Raha (money): 7	Eesti (Estonia): 88
Aasta (year): 7	Palk (salary): 50	Ära (away): 5	Õpetaja (teacher): 77
Ära (away): 7	Riik (country): 49	Inimene (person): 4	Raha (money): 68
Palk (salary): 7	Või (or): 45	**Kurat (devil): 4**	Või (or): 65
Aeg (time): 7	Raha (money): 45	Kõik (all): 4	Väga (very): 59
Rahvas (nation): 6	Teadma (to know): 42		Kaja: 54
Riik (country): 6	Võima (may): 41		Palk (salary): 53

Table 8. Gender separated word frequencies for Strict Model.

Distinct words for toxic class

A lot of words match between toxic and non-toxic classes. Further analysis was done for both men and female politicians posts comments where all words that are present in toxic and non-toxic datasets were eliminated and only words that are distinct to only toxic class are shown in the table 9.

From the results the distinct topics of discussion can be seen with the words like “Kirik” (church). Also, words related to Russia and Putin can suggest discussions in topics like political actions, references to individuals, and societal issues. In bold the words that could be related to toxicity are shown like “ori” (slave), “möla” (roar or word describing pointless text or speech) also exaggerations in words related to god as in Estonian there is a common phrase “Issand Jumal…..” (“Dear god, …”) that can be frequently used in comments.

From the female side, worlds that are related to political structures such as "resignation", "leaving", "power", "political party", and "taxpayer" in toxic comments may suggest dissatisfaction with governance and expressions of disagreement.

The usage of these words suggest that toxic comments might come from people who aren't happy with how politicians handle social issues (like raising teachers’ salaries). It can imply commenters want the government to change things or they're expressing different opinions. This kind of talk might include feelings of frustration, criticism, or disagreement with political choices, actions, or rules and the word usage supports that.

Men politician posts		Female politician posts	
Top words in toxic but not in non-toxic:		Top words in toxic but not in non-toxic:	
Näitama (to show): 8	Lammutama (demolition): 4	Vastaja (respondent): 9	Seisukoht (position): 5
Putin: 6	Toiduaine (food): 4	Maksumaksja (taxpayer): 8	Kolm (three): 4
Artikkel (article): 5	Valesti (wrong): 4	**Kurat (Devil): 7**	Igalpool (everywhere): 4
Kirik (church): 5	Tühi (empty): 4	Toetaja (supporter): 6	Koostöö (cooperation): 4
Vana (old): 5	**Jumal (God): 4**	Taas (again): 6	Põhi (bottom): 4
Lisa (additional): 5	Reaalsus (reality): 4	**Tagasiastumine (resignation): 5**	Võim (power): 4
Postimees (news outlet): 5	Amet (occupation): 4	Kondade (political party): 5	Valija (voter): 4
Ots (point): 5	Tõestama (to prove): 4	**Lahkuma (leave): 5**	Aegne (old): 4
Põhjustama (to cause): 5	Opositsioon (opposition): 4		Kuskile (somewhere): 4
Ori (slave): 5	**Möla (roar): 4**		

Table 9. Gender separated word frequencies specific to toxic class for Broad Model.

As Strict Model was trained to classify only relatively or very toxic comments as toxic, the results concur with it. In Table 10 most words can be classified as toxic for example words relating to disrespect and name calling: “joodik” (drunkard), “narkar” (junkie), “vitt” and “värdjas”, which both can be translated as bastard. Also, words like “paks” (fat), “siga” (pig), “päkapikk” (elf) can be put into the name calling category.

Words like “luud” (manipulative word used to describe a women) can be considered sexist. Words like “toppima” (stuffing like in a sentence: Don’t stuff your nose into other people’s business), “kaebama” (complain), “nõiajutud” (witch stories), “olevustega” (creatures) this are the words that can negatively amplify the sentence to be more hurtful or disrespectful to the readers.

The results also contain Estonian specific swear words like “puts”, “sita” (probably from word “sitaks” or grammatically falsely written word for “sita hunnik”) and “värdjas” (bastard) that also can be viewed as toxic. Words like ‘luud’ is usually a negative word for women. These results between genders are similar to Mertens et al. (2019) results that stated that female politicians were more described or talked with using gender related language especially in a negative way as the examples previously mentioned support.

Men politician posts		Female politician posts	
Top words in toxic but not in non-toxic:		Top words in toxic but not in non-toxic:	
Joodik (drunkard): 3	Positiivne (positive): 2	Tsaar (czar): 3	Ekre: 1
Tarkus (knowledge): 2	Kehv (bad): 2	Toppima (stuffing): 2	**putš:** 1
Alatu (sneaky): 2	**Narkar (junkie): 2**	**Isand (lord): 2**	päkapikk (elf): 1
EV (Estonia): 2	Põgenik (refugee): 2	**Käi (go): 1**	**sittuma (shit): 1**
Kaebama (complain): 2	**Sita: 1**	**Vitt (bastard): 1**	vähegi (little): 1
Kohus (judge): 2	rataske: 1	**Värdjas (bastard): 1**	sedasi (so): 1
Delfi (news outlet): 2	toiduahel (food chain): 1	**Paks (fat): 1**	**nõidajutud (witch**
Maksegraafik (payment		**Siga (pig): 1**	**stories): 1**
schedule): 2		**Kuradima (devilish): 1**	**olevustega (creature): 1**
			Luud: 1

Table 10. Gender separated word frequencies specific to toxic class for Strict Model. Some words do not have a direct translation to English and are described after the table.

The analysis of word frequencies in toxic and non-toxic classes using trained classifiers (Broad Model and Strict Model) revealed different patterns, however at the surface level the most frequent words surround the topics that were popular at the time. Notably, certain words such as "Eesti" (Estonia), "Inimene" (person), and "Riik" (country) feature prominently across toxic and non-toxic categories, reflecting that political posts comments are also very nation and country related.

Toxic comment word frequencies emerge when the comments are split between male and female politicians' posts. Male politician posts tend to have more toxic words related to topics like salaries and political affiliations, while female politician posts elicit discussions on political structures and governance dissatisfaction. This gender-based difference in word usage aligns with previous studies highlighting the tendency to use gender-related language, especially in a negative context (Mertens et al., 2019).

Words related to disrespect, name-calling, and sexist language are also included in toxic comments, suggesting that commenters are frustrated or in a disagreement with political decisions and actions. Additionally, the presence of Estonian-specific swear words shows the cultural influence in online discourse in Estonia.

4.2 Bigram analysis

In this section most used bigrams for both models are analyzed. Given the nuances of the Estonian language, bigrams, or pairs of words, can provide valuable insight into the potential indications of toxicity in online comments. In the bigram results, the names that are not politicians (i.e tagged persons) are manually removed from the presented results.

Looking at the results of the whole dataset (Table 11) for toxic and nontoxic comments for both models, there is clear distinction that prime minister at the time Kaja Kallas (in form Kaja Kallas (name) and Kaja Kallase (genitive case) (Estonian Cases: Introduction to Basic Estonian Grammar, n.d.) are among the most popular bigrams. This observation shows her important presence in public discussions during the period under observation.

For the Broad Model results for the toxic dataset, there is nonsingular bad or toxic bigram among the most popular bigrams. Topics are around money and teachers' salaries, which was the most prominent discussion point during the time data was collected (Rohemäe, 2023). Frequent bigrams for non-toxic class includes a lot of bigrams regularly used in every conversation (i.e I am/have, we have, he/she is, etc.).

Results of the Strict Model (trained on strongly toxic comments) already give negative and toxic bigrams that can be seen as offensive. Some examples are ('narkarid', 'joodikud') drug addicts and alcoholics/junkies, ('rahva', 'raha') people's money and ('vene', 'impeeriumi') Russian Empire's. Presented are also words in genitive case, which show belonging to whom or what, some examples are ('eesti', 'rahva') Estonian people's, ('rahva', 'raha') people's money and ('kaja', 'kallase') Kaja Kallas in genitive case. The presence of bigrams in the genitive case can potentially show generalizations or manipulative language, thus contributing to a toxic or offensive tone. The bigrams in non-toxic class are bigrams regularly used in every conversation.

Broad Model		Strict Model	
Bigrams in toxic dataset	Bigrams in non-toxic dataset	Bigrams in toxic dataset	Bigrams in non-toxic dataset
('on', 'väga') (is, very) 13	('ma', 'olen') (I am) 16	('eesti', 'rahva') (Estonian people's) 3	('kaja', 'kallas') 24
('kaja', 'kallas') 12	('kaja', 'kallas') 13	('ma', 'olen') (I am) 3	('ma', 'olen') (I am) 21
('on', 'meie') (is, our) 10	('on', 'väga') (is very) 11	('on', 'väga') (is very) 3	('on', 'väga') (is very) 21
('ma', 'olen') (I, am) 8	('teie', 'perele') (to your family) 9	('rahva', 'raha') (people's money) 3	('meil', 'on') (we have) 16
('meil', 'on') (we, have) 8	('on', 'kõik') (is all) 9	('ühte', 'paati') (in the same boat) 2	('on', 'meie') (is our) 13
('on', 'üks') (is, one) 8	('meil', 'on') (we have) 8	('aru', 'saada') (to understand) 2	('mul', 'on') (I have) 13
('saa', 'aru') (to, understand) 6	('mul', 'on') (I have) 8	('eestis', 'pole') (there is not in Estonia) 2	('on', 'kõik') (is all) 13
('raha', 'on') (money, is) 6	('kõik', 'on') (all is) 8	('kehva', 'seisu') (poor condition) 2	('kõik', 'on') (all is) 13
('inimesed', 'on') (people, are) 6	('on', 'vaja') (is needed) 7	('kaja', 'kallase') 2	('ta', 'on') (he/she is) 12
('õpetajate', 'palka') (teachers, salary) 6	('ta', 'on') (he/she is) 7	**('narkarid', 'joodikud') (drug addicts, alcoholics) 2**	('on', 'üks') (is one) 12
('eesti', 'on') (Estonia, is) 6	('väga', 'tubli') (very good) 6	**('vene', 'impeeriumi') (Russian Empire's) 2**	('eesti', 'on') (Estonia is) 11

Table 11. Most popular bigrams for both models and in both categories: toxic and nontoxic results.

Similarly to the word frequencies analysis section, we also analyzed bigrams unique only to toxic class to see if there are any word patterns indicating toxicity.

In Broad Model (Table 12), toxic comments primarily revolve around the same themes of teacher salaries and financial matters, evidenced by bigrams such as "teachers' salary" and "money comes." However, notable changes can be observed, including the direct imperative bigram ('saa', 'aru'),

meaning "understand!" This change suggests a shift towards more assertive or confrontational language within toxic comments. The name of the prime minister at the time is continuing to be popular, however in this case it is presented not as a full name but as a name in genitive case ('kaja', 'kallase').

Strict Model results reveals a more explicit use of offensive language, with bigrams denoting to drug addicts and alcoholics and phrases like ‘shit in the wheel’ as well as mentions of money ('raha', 'meie') and descriptive words like crisis ('kriisis', 'elada') or poor conditions ('kehva', 'seisu') around the same topics (teachers, education system, government budget). These findings suggest that toxic comments tend to use more colorful and emotional language to express complaints or criticisms, which can make the conversation more negative and toxic.

Broad Model		Strict Model	
Bigrams in unique to toxic dataset	Translations	Bigrams in unique to toxic dataset	Translations
('kaja', 'kallase') 8	(Kaja Kallas's) 8	('kehva', 'seisu') 2	(poor condition) 2
('õpetajate', 'palka') 6	(teachers' salary) 6	**('narkarid', 'joodikud') 2**	(drug addicts, alcoholics) 2
('saa', 'aru') 5	(understand!) 5	('raha', 'meie') 2	(money, our) 2
('inimesed', 'on') 5	(people are) 5	('sita', 'rataskes') 1	(shit in the wheel) 1
('on', 'meie') 5	(is our) 5	('maha', 'müüs') 1	(sold away) 1
('aasta', 'pärast') 4	(after a year) 4	('tahtis', 'kriisis') 1	(important, in crisis) 1
('on', 'üks') 4	(is one) 4	('kriisis', 'elada') 1	(live, in crisis) 1
('üle', 'võimete') 4	(beyond the capabilities) 4	('elada', 'pankurid') 1	(live, bankers) 1
('õpetajad', 'on') 4	(teachers are) 4	('ametnike', 'näol') 1	(officers' faces) 1
('raha', 'tuleb') 4	(money comes) 4	('üks', 'rehknut') 1	(one, swindlers) 1

Table 12. Bigrams unique to only toxic comments for both models.

Following the same structure, we split the results between comments made by people under men and female politicians’ posts to see if there were any differences and what they were.

Firstly, we will look at the result of Broad Model (Table 13). Comments from both male and female politicians’ post still have Kaja Kallas and Kaja Kallase as the most frequently used bigram. The topics discussed remain consistent, and there are no noticeable toxic phrases that stand out.

Men politician posts		Female politician posts	
Bigrams in toxic dataset	Bigrams in non-toxic dataset	Bigrams in toxic dataset	Bigrams in non-toxic dataset
('kaja', 'kallase') 6	('ma', 'olen') (I am) 9	('kaja', 'kallas') 10	('kaja', 'kallas') 11
('ma', 'olen') (I am) 6	('teie', 'perele') (to your family) 9	('on', 'väga') (is very) 7	('on', 'väga') (is very) 8
('meil', 'on') (we have) 6	('väga', 'tubli') (very good) 5	('on', 'meie') (is our) 5	('ma', 'olen') (I am) 7
('on', 'väga') (is very) 6	('on', 'selles') (is in this) 4	('kõik', 'on') (everything is) 4	('meil', 'on') (we have) 7
('on', 'üks') (is one) 6	('on', 'mul') (is for me) 4	('samal', 'ajal') (at the same time) 4	('on', 'kõik') (is all) 6
('on', 'meie') (is our) 5	('ukraina', 'sõda') (Ukraine war) 4	('eesti', 'on') (Estonia is) 4	('saatke', 'mulle') (send to me) 5
('saa', 'aru') (understand) 4	('ma', 'olin') (I was) 4	('meie', 'riigi') (our country's) 4	('on', 'tulnud') (has come) 5
('ta', 'on') (he/she is) 4	('edu', 'sulle') (success to you) 4	('on', 'raha') (is money) 4	('on', 'ta') (is he/she) 5
('aru', 'saada') (to understand) 4	('on', 'nad') (is them) 3	('on', 'kõik') (is all) 4	('kõik', 'on') (all is) 5
('aasta', 'pärast') (after a year) 4	('ma', 'tea') (I don't know) 3	('raha', 'on') (money is) 4	('on', 'vaja') (is needed) 5

Table 13. Gender separated popular bigrams for Broad Model.

The second model (Strict Model) results as shown in Table 14, do have a notable dominance of offensive and toxic language within comments classified as toxic.

Comments under posts by male politicians feature bigrams in the genitive form ('eesti', 'rahva') (Estonian people's), ('kehva', 'seisu') (poor condition), ('on', 'meie') (is our), ('kaja', 'kallase') and ('vene', 'impeeriumi') (Russian Empire's). Genitive form also shows possessions or is considered possessive language (Estonian Cases: Introduction to Basic Estonian Grammar, n.d.). From the offensive language side ('narkarid', 'joodikud') (drug addicts, alcoholics), ('sita', 'rataskes') (shit in the wheel) and (shit-wheel Estonia) all are highlighting the offensive tone.

Female politicians' posts comments bigrams also include swore words like ('käi', 'vittu') (go to hell), ('vittu', 'värdjas') (fuck bastard), ('värdjas', 'vaata') (bastard, look), ('paks', 'siga') (fat pig), ('kuradima', 'nõid') (damn witch), ('nõid', 'astu') (witch, walk), ('loll', 'luud') (stupid bitch). Bigrams referring to money are also presented. Compared to male posts comments, female post comments according to this are more toxic This observation aligns with the findings of the earlier analysis of frequent words, further supporting the conclusion.

Non-toxic bigrams remain consistent with the previously identified patterns, including common phrases used in everyday discussions.

Men politician posts		Female politician posts	
Bigrams in toxic dataset	Bigrams in non-toxic dataset	Bigrams in toxic dataset	Bigrams in non-toxic dataset
('ma', 'olen') (I am) 3	('ma', 'olen') (I am) 12	('rahva', 'raha') (people's money) 3	('kaja', 'kallas') 20
('eesti', 'rahva') (Estonian people's) 2	('teie', 'perele') (to your family) 9	('on', 'suurepärane') (is excellent) 2	('on', 'väga') (is very) 14
('ühte', 'paati') (in the same boat) 2	('meil', 'on') (we have) 7	('raha', 'meie') (money is our) 2	('on', 'kõik') (is everything) 10
('kehva', 'seisu') (poor condition) 2	('ta', 'on') (he/she is) 7	**('käi', 'vittu') (go to hell) 1**	('ma', 'olen') (I am) 9
('on', 'meie') (is our) 2	('on', 'väga') (is very) 7	**('vittu', 'värdjas') (fuck bastard) 1**	('meil', 'on') (we have) 9
('kaja', 'kallase') 2	('on', 'üks') (is one) 7	**('värdjas', 'vaata') (bastard, look) 1**	('kõik', 'on') (everything is) 8
('narkarid', 'joodikud') (drug addicts, alcoholics) 2	('mul', 'on') (I have) 6	**('paks', 'siga') (fat pig) 1**	('on', 'meie') (is our) 8
('vene', 'impeeriumi') (Russian Empire's) 2	('saa', 'aru') (understand) 5	**('kuradima', 'nõid') (damn witch) 1**	('eesti', 'on') (Estonia is) 7
('sita', 'rataskes') (shit in the wheel) 1	('on', 'meie') (is our) 5	**('nõid', 'astu') (witch, walk) 1**	('meie', 'riigi') (our country's) 7
('rataskes', 'eesti') (shit-wheel Estonia) 1	('raha', 'on') (money is) 5	**('loll', 'luud') (stupid bitch) 1**	('mul', 'on') (I have) 7

Table 14. Gender separated popular bigrams for Strict Model.

The analysis of bigrams from comments under posts by Estonian politicians reveals different patterns of Estonian expressions and toxicity. In Broad Model, popular bigrams, such as references to the Prime Minister and themes related to financial matters, show consistency across comments under posts by both male and female politicians. However, Strict Model exposes a stark contrast, with comments under posts by female politicians showing a notably higher dominance of offensive language and offensive references compared to those under male politicians' posts. Despite variations in toxicity levels, non-toxic bigrams remain consistent with everyday conversational patterns across both models.

4.3 Sentiment analysis on toxic comments

A small sentiment analysis was carried out on comments categorized as toxic from both models. In the results from Broad Model, it was found that 92% of the comments exhibited a negative sentiment, with a breakdown of 95% for male politicians' comments and 89% for female politicians. Additionally, 4% of the comments were neutral, while another 4% were positive, with a similar distribution between male and female comments. In contrast, the findings from Strict Model revealed that a staggering 97% of the comments had a negative sentiment, leaving only 2 out of 75 comments with a positive sentiment.

4.4 Adjectives, verbs, and nouns in toxic classes.

In this section most popular adjectives, verbs and nouns are compared between Broad Model and Strict Model with the distinction between comments from male and female politician’s posts.

From Broad Model (Table 15) it is notable that nouns most frequently used in comments from male politicians align perfectly with those observed in the general statistics overview of the data. On the other hand, comments from female politicians' posts mostly feature adjectives such as "average," "last," and "sick." The use of "average" likely pertains to discussions regarding average salary, with "salary" also emerging as one of the most popular nouns.

Men			**Women**		
Noun	**Verb**	**Adjective**	**Noun**	**Verb**	**Adjective**
Inimene (person)	Teadma (to know)	Hea (good)	Inimene (person)	Tahtma (to want)	Keskmine (average)
Õpetaja (teacher)	Võima (can)	Noor (young)	Riik (country)	Võtma (to take)	Hea (good)
Palk (salary)	Tahtma(to want)	Kõrge (high)	Raha (money)	Andma (to give)	Viimane (last)
Riik (country)	Võtma (to take)	Suur (big)	Õpetaja (teacher)	Ütlema (to say)	Haige (sick)
Raha (money)	Maksma (to pay)	Raske (hard)	Palk (salary)	Võima (to allow)	Suur (big)

Table 15. Most popular nouns, verbs and adjectives in men and women post comments in toxic class. For Broad Model

In Strict Model results (Table 16), adjectives with negative implications such as "stupid," "sick," and "sneaky" are prevalent, along with verbs like "to screw," "to lie," and "to leave," which naturally carry negative meanings and tone. Nouns continue to revolve around themes related to teachers, salaries, and money. Positive adjectives like wonderful and right can in this context be used in a sarcastic way for them to be in toxic category.

Men			**Women**		
Noun	**Verb**	**Adjective**	**Noun**	**Verb**	**Adjective**
Aasta (year)	Teadma (to know)	Õige (right)	Inimene (person)	Vaatama (to see)	**Loll (stupid)**
Aeg (time)	Elama (live)	Noor (young)	Riik (country)	**Keerama (to screw)**	Õige (right)
Palk (salary)	Tahtma(to want)	Vana (old)	Raha (money)	**Valetama (to lie)**	Suurepärane (wonderful)
Riik (country)	Võtma (to take)	Alatu (sneaky)	Pea (head)	Jagama (to share)	**Haige (sick)**
Rahvas (nation)	Jääma (to stay)	Raske (hard)	Rahvas (nation)	Lahkuma (to leave)	**Vale (wrong)**

Table 16. Most popular nouns, verbs and adjectives in men and women post comments in toxic class. For Strict Model

During the data collection period, prominent topics emerged in the comments, notably centering around teacher salaries and governmental responses to the topic. The prevalence of related vocabulary suggests an active involvment from the people with these problems, given that teacher salaries are linked to government budget allocations. Interestingly, across different subgroup

analyses, including distinctions between models (toxicity levels), and gender, the thematic consistency is the same.

A linguistic trend that was observed was the frequent use of genitive cases, reflecting possessive language, particularly shown in toxic comments. However, at broader subgroup levels, such as toxicity classifications and gender differentiations, the overall discussions and comments appear to be relatively civil, as indicated by the low percentage of comments labeled as toxic by both Broad Model (33%) and Strict Model (5%).

Also, within the subset of toxic comments, there are instances of pronounced toxicity characterized by using swear words and adjectives with negative or disturbing implications. Additionally, manual elimination of words for the Estonian language toxicity database revealed evampels of racial generalizations in toxic comments, including terms like "slaavlased" (Slavs), "mustlased" (Gypsies), "venelased" (Russians), and "moslemid" (Muslims).

Additionally, gender-separated analyses uncovered differences, with comments on posts by female politicians attracting more spam and exhibiting higher toxicity levels, a trend supported by bigram analysis. These findings show the complex dynamics of online discussions in Estonia, highlighting both constructive dialogue and instances of toxicity, particularly in conversations surrounding sensitive topics such as government policies and societal issues.

Limitations

There are some limitations that I would like to point out.

As this paper uses public data at the time of data collection, we do not know and cannot include comments that were previously removed by Facebook due to toxicity or comments that were reported as bad and then removed.

Conclusion

This thesis aimed to analyze the extent of toxicity that is in the comment sections of Estonian politician Facebook posts with a focus on understanding the dynamics between gender, language usage, and online discussions. 10 female and 10 male politicians from XV parliament composition that include all political parties represented in the parliament were selected. The Facebook posts and comments were collected in December 2023 (scraped starting from 03.03.2023 to 9.12.2023, however the majority of posts were made in November and December 2023).

As toxicity is hard to define, we primarily used Kim et al. (2021) definition, where **toxicity is defined as text containing disrespect, name-calling, or the use of terms that are racist, sexist, or xenophobic.** However, given the focus on the Estonian language, we have also included Estonian-specific swear words and other hurtful terms and expressions.

To research toxicity in the comments, we proposed the following research questions:

RQ1: What differences exist in the word usage in Facebook comments on posts made by male and female politicians?

RQ2: How does the frequency of toxic language differ between comments on posts made by male and female politicians?

RQ3: To what extent do commenters exhibit gender bias, discriminatory language, or name-calling (labeling) in the male and female politician posts comments?

Addressing these questions required training of a toxicity classifier to identify and categorize toxic language within the dataset. The following toxicity analysis of word usage patterns revealed interesting insights. While no huge differences were observed in the surface-level word usage between comments on posts by male and female politicians, a deeper examination showed slight differences, particularly in the usage of toxic language. Comments on posts by female politicians showed a higher density of toxic language, displaying potential gender dynamics within online political discussions (RQ1 & RQ2).

Furthermore, the examination of gender bias and discriminatory language exposed noteworthy trends, including the frequent use of Prime Minister Kaja Kallas's given name and the utilization of specific offensive terms targeted at women. These findings show the versatile nature of online interactions (RQ3).

In addition to improving our understanding of online communication dynamics, this study contributes significantly to the development of Estonian language processing tools. By establishing a novel Estonian language toxicity database including 167 offensive words sourced from social media platforms like Facebook, alongside the introduction of a state-of-the-art toxicity detection classifier (one trained on very toxic comments and other on more broader spectrum of comments) and a systematic analysis pipeline, the research provides a valuable framework for future investigations in this domain.

Reference list

Alsharef, A., Aggarwal, K., Sonia, Koundal, D., Alyami, H., & Ameyed, D. (2022). An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding. Computational Intelligence and Neuroscience, 2022, e8467349. https://doi.org/10.1155/2022/8467349

Antti Virtanen, Kanerva, J., Rami Ilo, Luoma, J., Juhani Luotolahti, Tapio Salakoski, Ginter, F., & Sampo Pyysalo. (2019). Multilingual is not enough: BERT for Finnish. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1912.07076

Azri, H., Belbachir, H., & Meddah, F. G. (2022). Identifying spam activity on public Facebook pages. CIT. Journal of Computing And Information Technology/Journal of Computing And Information Technology, 29(3), 133–149. https://doi.org/10.20532/cit.2021.1005221

Bojanowski, P., Grave, É., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Chen, Z., Shen, S., Hu, Z., Lu, X., Mei, Q., & Liu, X. (2019). Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification. The World Wide Web Conference on - WWW '19. https://doi.org/10.1145/3308558.3313600

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. Journal of Communication, 64(4), 658–679.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.org. https://arxiv.org/abs/1810.04805

Diego Reforgiato Recupero, Consoli, S., Gangemi, A., Andrea Giovanni Nuzzolese, & Spampinato, D. (2014). A Semantic Web Based Core Engine to Efficiently Perform Sentiment Analysis. Lecture Notes in Computer Science, 245–248. https://doi.org/10.1007/978-3-319-11955-7_28

Dragoni, M., & Diego Reforgiato Recupero. (2016). Challenge on Fine-Grained Sentiment Analysis Within ESWC2016. Communications in Computer and Information Science, 79–94. https://doi.org/10.1007/978-3-319-46565-4_6

Elnaggar, A., Bernhard Waltl, Glaser, I., Jörg Landthaler, Scepankova, E., & Matthes, F. (2018). Stop Illegal Comments. https://doi.org/10.1145/3299819.3299845

Enaitz Ezpeleta, Iturbe, M., Iñaki Garitano, Velez, I., & Urko Zurutuza. (2018). A Mood Analysis on Youtube Comments and a Method for Improved Social Spam Detection. Lecture Notes in Computer Science, 514–525. https://doi.org/10.1007/978-3-319-92639-1_43

Enaitz Ezpeleta, Iñaki Garitano, Arenaza-Nuño, I., Hidalgo, G., & Urko Zurutuza. (2018). Novel Comment Spam Filtering Method on Youtube: Sentiment Analysis and Personality

Erakondade reitingud - iganädalane erakondliku eelistuse küsitlus. (n.d.). Erakondade Reitingud. https://reitingud.ee/

Eskelinen, A., Silvala, L., Ginter, F., Pyysalo, S., & Laippala, V. (2023, May 1). Toxicity Detection in Finnish Using Machine Translation (T. Alumäe & M. Fishel, Eds.). ACLWeb; University of Tartu Library. https://aclanthology.org/2023.nodalida-1.68/

Estnltk. (n.d.). Estnltk.github.io. Retrieved May 9, 2024, from https://estnltk.github.io/

Estonian Cases: Introduction to Basic Estonian Grammar. (n.d.). Lingvist. Retrieved May 9, 2024, from https://lingvist.com/course/learn-estonian-online/resources/introduction-to-basic-estonian-grammar/

Fahim, A., & Naseem, M. N. (2015). Facebook spam and spam filter using artificial neural networks. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.1098954

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., Elsheikh, A. H., Abualigah, L., & Al-qaness, M. a. A. (2021). Social Media toxicity classification using Deep Learning: Real-World Application UK Brexit. *Electronics*, *10*(11), 1332. https://doi.org/10.3390/electronics10111332

Field, A., & Tsvetkov, Y. (2020). Unsupervised Discovery of Implicit Gender Bias. Arxiv.org. https://arxiv.org/abs/2004.08361

Fichman, P., & Rathi, M. (2022). The Impact of Culture on Online Toxic Disinhibition: Trolling in India and the USA. Hawaii International Conference on System Sciences 2022 (HICSS-55). https://aisel.aisnet.org/hicss-55/dsm/culture/5/

Goodman, J., Cormack, G. V., & Heckerman, D. (2007). Spam and the ongoing battle for the inbox. Communications of the ACM, 50(2), 24–33. https://doi.org/10.1145/1216016.1216017

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. Expert Systems with Applications, 36(7), 10206–10222. https://doi.org/10.1016/j.eswa.2009.02.037

Kim, J. W., Guess, A. M., Nyhan, B., & Reifler, J. (2021). The Distorting prism of Social Media: How Self-Selection and Exposure to Incivility fuel Online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Kobellarz, J. K., & Silva, T. H. (2022). Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English. https://doi.org/10.1145/3539637.3556892

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. Proceedings of the AAAI Conference on Artificial Intelligence, 27(1). https://ojs.aaai.org/index.php/AAAI/article/view/8539

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv.org. https://arxiv.org/abs/1907.11692

Lees, A., Vinh Cao Trần, Tay, Y., Sorensen, J., Gupta, J. P., Metzler, D., & Vasserman, L. (2022). A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3534678.3539147

Marjanović, S. V., Stańczak, K., & Augenstein, I. (2022). Quantifying gender biases towards politicians on Reddit. PloS One, 17(10), e0274317. https://doi.org/10.1371/journal.pone.0274317

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Mertens, A., Pradel, F., Ayjeren Rozyjumayeva, & Jens Wäckerle. (2019). As the Tweet, so the Reply? https://doi.org/10.1145/3292522.3326013

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. Political Communication, 36(2), 214–226

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. Proceedings of the 25th International Conference on World Wide Web - WWW '16. https://doi.org/10.1145/2872427.2883062

O'Callaghan, D., Harrigan, M., Carthy, J., & Cunningham, P. (2021). Network Analysis of Recurring YouTube Spam Campaigns. Proceedings of the International AAAI Conference on Web and Social Media, 6(1), 531–534. https://doi.org/10.1609/icwsm.v6i1.14288

Pavlopoulos, J., Prodromos Malakasiotis, & Ion Androutsopoulos. (2017). Deeper Attention to Abusive User Content Moderation. https://doi.org/10.18653/v1/d17-1117

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020, June 1). Toxicity Detection: Does Context Really Matter? ArXiv.org. https://doi.org/10.48550/arXiv.2006.00998

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation (pp. 1532–1543). Association for Computational Linguistics. https://aclanthology.org/D14-1162.pdf

Recognition. Lecture Notes in Computer Science, 228–240. https://doi.org/10.1007/978-3-319-74433-9_21

Riigikogu. (2023, April 3). XV riigikogu - Riigikogu. https://www.riigikogu.ee/en/parliament-of-estonia/composition/

Riigi Teataja (n.d.). *XV Riigikogu liikmete registreerimine–Riigi Teataja*. https://www.riigiteataja.ee/akt/331032023001

Rohemäe, M. (2023, November 15). Kallas õpetajate palgatõusust: ei saa lubada midagi, milleks raha ei ole. ERR. https://www.err.ee/1609165345/kallas-opetajate-palgatousust-ei-saa-lubada-midagi-milleks-raha-ei-ole

Rudinger, R., May, C., & Van Durme, B. (2017). Social Bias in Elicited Natural Language Inferences. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. https://doi.org/10.18653/v1/w17-1609

Sahami, M., Dumais, S.T., Heckerman, D.E., & Horvitz, E. (1998). A Bayesian Approach to Filtering Junk E-Mail. AAAI Conference on Artificial Intelligence.

Sheth, A. P., Shalin, V. L., & Sheth, A. P. (2021). Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key. *arXiv (Cornell University)*. https://doi.org/10.1016/j.neucom.2021.11.095

Sydnor, E. (2019). Disrespectful democracy: The psychology of political incivility. Columbia University Press

Tay, Y., Vinh Cao Tran, Ruder, S., Gupta, J. P., Hyung Won Chung, Bahri, D., Qin, Z., Baumgartner, S., Yu, C., & Metzler, D. (2021). Charformer: Fast Character Transformers via Gradient-based Subword Tokenization. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.12672

Tanvir, H., Kittask, C., Eiche, S., & Sirts, K. (2021, May 31). EstBERT: A Pretrained Language-Specific BERT for Estonian (S. Dobnik & L. Øvrelid, Eds.). ACLWeb; Linköping University Electronic Press, Sweden. https://aclanthology.org/2021.nodalida-main.2/

TartuNLP/EstBERT · Hugging face. (n.d.). https://huggingface.co/tartuNLP/EstBERT?fbclid=IwAR38WilbRF8BK44oz5Ie0lIbsSFrRn1F6J398jOnw2x50hB17_zZelqAZqQ

tartuNLP/EstBERT128_sentiment · Hugging Face. (n.d.). Huggingface.co. Retrieved May 9, 2024, from https://huggingface.co/tartuNLP/EstBERT128_sentiment

T. C. Alberto, J. V. Lochter and T. A. Almeida, "TubeSpam: Comment Spam Filtering on YouTube," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2015, pp. 138-143, doi: 10.1109/ICMLA.2015.37.

toxicity noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. (n.d.). https://www.oxfordlearnersdictionaries.com/definition/english/toxicity?q=toxicity

Uiboaed, K. (2018). Eesti keele stoppsõnad / Estonian stop words. Datadoi.ee. https://doi.org/10.15155/re-48

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for Toxic Comment Classification: An In-Depth Error Analysis. ArXiv:1809.07572 [Cs]. https://arxiv.org/abs/1809.07572

Wang, K., Yang, J., & Wu, H. (2021, December 12). A Survey of Toxic Comment Classification Methods. ArXiv.org. https://doi.org/10.48550/arXiv.2112.06412

Xia, Y., Zhu, H., Lu, T., Zhang, P., & Gu, N. (2020). Exploring Antecedents and Consequences of Toxicity in Online Discussions. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–23. https://doi.org/10.1145/3415179

Zaheri, S., Leath, J., & Stroud, D. (2020). Toxic Comment Classification. SMU Data Science Review, 3(1). https://scholar.smu.edu/datasciencereview/vol3/iss1/13/

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semantic Web, 10(5), 925–945. https://doi.org/10.3233/sw-180338

Appendices

Appendix 1

-1 - Positiivne Kommentaar: "Sa oled suurepärane! Sinusuguseid poliitikuid Eesti vajabki!" Selgitus: Kommentaar on väga positiivses toonis kirjutatud.
0 - Neutraalne Kommentaar: "..." Selgitus: Kommentaar on neutraalne, ei sisalda negatiivsust, positiivsust ega solvanguid.
1 - Kergelt Toksiline Kommentaar: "Ma saan aru, mida sa üritad öelda, aga see kõlab natuke naiivselt." Selgitus: Kommentaar sisaldab kerget kriitikat või irooniat, mis võib tekitada ebamugavust või olla veidi solvav.
2 - Mõõdukalt Toksiline Kommentaar: "Ainult täielik idioot võiks sellise asjaga nõustuda." Selgitus: Kommentaar sisaldab otsest kriitikat, mis võib olla solvav ja sisaldada negatiivset hinnangut isiku suhtes.
3 - Täielikult Toksiline Kommentaar: "Sina litakas persevest ei peaks üldse arvamust avaldama" Selgitus: Kommentaar on selgelt ründav, sisaldab tugevaid solvanguid ja on mõeldud isiku või tema vaadete alavääristamiseks.
X - Seosetu Kommentaar: "Mait Metsast" Kommentaar 2: "Art" Selgitus: Kommentaari sisust on raske aru saada. Kommentaar koosneb ainult nimest (st. kommentaari kirjutamist on alustatud, aga mitte lõpetatud, või on kasutaja lihtsalt ära tagitud).
S - Spämm/Skämm Kommentaar: "Täname, et meeldisite minu kinnitatud lehele ja jätsite sellele kommentaare. Klõpsake lingil, et saata otsesõnum minu fännilehele....facebook.com/" Selgitus: Kommentaar on selgelt kellegi andmete õngitsemiseks mõeldud või on korduv mitmetes kommentaarides.

Appendix 2
Spam examples (in Estonian)

Indrek Nurmits Tere hommikust. Minu nimi on Maria Julianna DABASINE VALKAI. Kasvatan hobuseid. Praegu on mul tohutu varandus. Aga kahjuks olen raskelt haige ja suren. Pärast mitmeid teste diagnoosis arst mul ajuvähi ja pani mind mõistma, et mul pole enam palju aega elada. Tegelikult ma enam seda üle ei ela. Ma olen tõesti meeleheitel ja mu süda veritseb. Saadan teile selle sõnumi, et saaksite minu tervislikust seisundist aimu ja loodan vähemalt teie tähelepanu köita. Tänan teid, sest me ei tunne üksteist tegelikult, kuid ärge kõhelge minuga ühendust võtmast.

Põhjus, mis mind teie juurde ajendab, on järgmine: kirjutan teile selle kanali kaudu heategevuseks teie heaks. Olen siin, et annetada teile mitu eurot. Pea seda kingituseks. Minu perekonnaseis on selline, et mul pole meest ja veel vähem lapsi.

Mul pole muud võimalust teiega suhelda kui siin või kaudu

Tänan, et vastasite pärast minu sõnumi lugemist.

E-post: maria.dabasine02@gmail.com

Argo Hallop Vabandan, et sellisel viisil ühendust võtsin, minu nimi on Sandra KPNEL, olen Prantsuse Polüneesias elav Eesti kodanik. Mul on raske haigus, mis mõistab mind surma, see on emakakaelavähk ja mul on euro suurune summa, mille tahan anda kellelegi, kes on kasvanud Jumalas nagu sina. Nädal tagasi kaotasin oma mehe kohutavas autoõnnetuses, mis mind väga raputas ja ma ei saanud ikka veel abielluda; lapsi meil ei olnud, selle summa tahan enne surma ära anda, sest arstide sõnul on mu päevad loetud. ma ravisin valu;

Ma ei taha teada, kas ravi õnnestub, kui soovite seda kingitust kasutada.

Siin on minu post. e-maili aadress:

miliauskienekpnel@gmail.com

võtke minuga ühendust, rahu olgu teiega

miliauskienekpnel@gmail.com

Appendix 3

Data and models can be found on following link:

https://drive.google.com/drive/folders/1TBiecI-pED-FyIsNQXiRly9L3hpPFiZB?usp=sharing

Appendix 4

Mina, Liina Vesilind,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Understanding toxicity of Estonian politicians Facebook posts comments

mille juhendajad on Uku Kangur, Rajesh Sharma

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Liina Vesilind

15.05.2024