Toxicity Detection for Ukrainian-Language Texts in the TextAttributor System

Nataliia Darchuk[1,†], Oксana Zuban[1,*,†], Valentyna Robeiko[1,2,†], Yuliia Tsyhvintseva[1,3,†] and Mykola Sazhok[2,*,†]

[1] Taras Shevchenko National University of Kyiv, 60 Volodymyrska Street, Kyiv, 01033, Ukraine
[2] Institute for Information Technologies and Systems of the National Academy of Sciences of Ukraine, 40 prospekt Akademika Hlushkova, Kyiv, 03187, Ukraine
[3] Institute of the Ukrainian Language of the National Academy of Sciences of Ukraine, 4 Mykhailo Hrushevskyi Street, Kyiv, 01001, Ukraine

Abstract

This paper presents the development and results of two distinct modules designed to detect toxic language in Ukrainian texts, primarily implemented within the "TextAttributor 1.0" expert system. The first module utilizes a rule-based approach, analyzing text through predefined linguistic rules and lexicon-based methods, while the second employs machine learning techniques, specifically leveraging the fastText and LLAMA-3 models, to automatically detect toxic content. The rule-based module outputs a detailed linguistic analysis, mapping toxic vocabulary using a precompiled lexicographic database, while the machine learning module calculates toxicity based on statistical models. The performance of both methods was evaluated by comparing their results on a corpus of Ukrainian texts, with the Pearson correlation coefficient employed to assess their alignment. The results demonstrate the system's capacity to effectively identify toxic content, contributing to ongoing efforts to mitigate the spread of harmful information. This paper contains rude texts that only serve as illustrative examples.

Keywords

toxic text detection, sentiment analysis, Ukrainian language, lexicon-based method, deep learning, text classification

1. Introduction

The study of text toxicity represents a relatively new field of inquiry within the broader domain of sentiment analysis. The tonality of a text is a significant element that influences how it is perceived and understood by the reader. It also enables the author to achieve their communicative objectives. For this reason, the task of determining the tone of a text is of great interest today not only to modern linguists and computer scientists but also to political scientists, managers, marketers, advertisers, image makers, and other professionals working with a particular brand. This task has become particularly crucial with the advent of the Internet, as it has provided a new platform for the analysis of media texts that "transmit, store, and reproduce information that influences public opinion" [1]. The increased attention to the negative, toxic, and emotional component of textual information occurred during the hybrid war period. The capacity of contemporary Internet media and social networks to exert a deleterious influence on vast numbers of individuals through the introduction of destructive and other harmful information into the mind or subconscious, which

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

[*] Corresponding author.
[†] These authors contributed equally.

n.darchuk@knu.ua (N. Darchuk); oxana.zuban@knu.ua (O. Zuban); valentyna.robeiko@knu.ua (V. Robeiko); julivoznuk@gmail.com (Yu. Tsyhvintseva); sazhok@gmail.com (M. Sazhok)
0000-0001-8932-9301 (N. Darchuk); 0000-0002-2644-3892 (O. Zuban); 0000-0003-2266-7650 (V. Robeiko); 0000-0002-9684-3840 (Yu. Tsyhvintseva); 0000-0003-1169-6851 (M. Sazhok)

leads to an inadequate perception of reality, underscores the acute urgency of the problem of "ecology" and the protection of the Internet space.

The issue of addressing the proliferation of destructive online content is a matter of global concern and significance. Since 2019, June 18 has been designated by the United Nations as the International Day against Hate Speech. This year, the Council of Europe held the Week against Hate Speech on June 17-20 in Strasbourg. In 2022, the Committee of Ministers of the Council of Europe developed Recommendations on combating hate speech [3]. It is important to note that there is no clear and universally accepted definition of toxic speech or hate speech. The definitions of these concepts vary across international documents in the legislative, social, and linguistic fields. However, they all have one thing in common: they describe texts that manifest aggression against other people, nations, social groups, etc. and that violate human rights.

The full-scale Russian-Ukrainian war has highlighted the urgency of this issue, as hate speech and other forms of aggressive propaganda represent a crucial element of the Russian propaganda apparatus in its efforts to erode the identity and integrity of the Ukrainian nation. To address this issue, it is essential to develop tools for automated analysis and detection of Ukrainian-language textual content that negatively impacts an individual's psychological state, public consciousness, and infringes upon the rights and legitimate interests of users, society, and the state.

In light of the pressing necessity for such tools, our team has developed IT solutions for the automated identification of toxicity in Ukrainian-language text. These tools are integrated into TextAttributor 1.0 [4], a linguistic parameterization expert system for Ukrainian-language media texts. The research tasks were divided into two successive stages: (a) development of the module for generating linguistic expertise of toxic text, detecting the toxicity index according to dictionaries and rules and (b) forming a dataset of toxic texts using this module and human expertise for machine learning. Additionally, a deep learning model was trained and evaluated by the dataset, forming a machine learning module for toxicity detection. The concept of developing an automated system for assessing the toxicity of text is rooted in the team's extensive research experience in computational linguistics and a particular interest in sentiment analysis [5, 6, 7, 8, 9].

The objective of the present article is twofold: firstly, to analyze the implementation of two methods – the lexicon and rules-based method, and the deep learning-based method – in the automatic analysis of the toxicity of Ukrainian-language texts; secondly, to investigate the effectiveness of the TextAttributor 1.0 system using the two methods on the texts analyzed by the system. The object of this study is Ukrainian-language texts. The subject of this study is the criteria and methods used for the automatic identification of toxic Ukrainian-language texts. The study employs a range of methods, including: a rules and lexicon-based sentiment analysis method, a deep learning-based method, and a combination of linguistic methods, namely component analysis, distributional analysis, and taxonomic analysis. Additionally, statistical methods, such as toxicity indexing and the calculation of the Pearson correlation coefficient for sample data, are utilized. Graphical modeling of statistical data is also employed as a method to visualize 520 Ukrainian-language texts that were submitted by users of the TextAttributor 1.0 web application.

2. Related Works

The latest developments in the field indicate that the task of automatic toxicity detection is currently solved mainly by applying deep learning methods based on various architectures (CNN, LSTM, BERT) [10, 11, 12] and, less frequently, by the use of traditional machine learning methods based on TF-IDF [13] or lexicon-based methods. Lexicon-based methods, despite their limitations as demonstrated by experimental data, offer valuable insights into stylistic and lexical features, making them useful tools for linguistic text analysis [14, 15]. Cross-lingual learning and translation techniques have gained increasing prominence, providing effective solutions to the challenges of text classification across diverse linguistic contexts [16]. Many studies have focused on English and other languages with abundant resources, largely due to the availability of extensive datasets. In particular, researchers have paid close attention to contextual embeddings, such as BERT and fastText [17], which enable efficient handling of misspellings, rare words, and newly introduced terms. Multilingual embeddings, like mBERT, further enhance the ability to process multilingual offensive or toxic content [18].

Recently, there have also been publications devoted to the detection of toxic language in low-resource languages, including Ukrainian. This is primarily due to the release of multilingual large language models (LLMs) and the development of new methods for creating datasets, including the development of translated datasets and the generation of synthetic data. To date, there is no publicly available expert-annotated toxic text dataset. The only existing corpus of this kind is the Ukrainian tweets corpus [9], which can be filtered by toxic keywords (Ukrainian obscene lexicon) provided by the author. Thus, researchers are yet to find a solution to this issue. In their publication [19], the authors describe three approaches to creating a corpus of Ukrainian-language toxic texts with binary markup (toxic and non-toxic). (i) translation from English; (ii) toxic samples filtering by toxic keywords; (iii) crowdsourcing data annotation for phrases containing five to twenty words. Furthermore, researchers delineate and contrast three methodologies for identifying toxicity: Prompting of LLMs, Cross-lingual transfer approach and Fine-tuning of LLMs on different types of data. The results are somewhat controversial, as each approach demonstrates efficacy on a different dataset. In [20], the authors present a bullying detection model for Ukrainian language. In order to construct the model, the researchers created a dataset by means of machine translation from English to Ukrainian. The authors assess the efficacy of the zero-shot technique and evaluate the performance of contemporary multilingual models and embeddings (mBERT, XLM-R, LASER, MUSE). As a result, the final detection model exhibits promising metrics. The authors conclude that, in the context of low-resource languages, the classification accuracy of a given model tends to increase in proportion to the number of samples used, regardless of their origin.

3. A lexicon-based method for toxicity detection of Ukrainian-language texts

The lexicon-based method is employed to determine the toxicity of a text. This entails identifying toxic words within the text in accordance with pre-compiled dictionaries of toxic vocabulary. The text is then evaluated on a toxicity scale according to the number of instances of toxic vocabulary identified. This method is employed in conjunction with automatic morphological and syntactic analysis, namely automatic rules-based text analysis. For this reason, the dictionary-based toxicity assessment may also be referred to as the dictionary and lexicon-based method or the rules-based method. The accuracy and completeness of the toxicity determination using this method are contingent upon the scope and quality of the compiled lexicon, as well as the quality of the lemmatization procedure for the text. The scope and semantics of the words in the toxic dictionary are contingent upon the style, genre, and subject matter of the texts for which the analysis system is being constructed.

The efficacy of the dictionary-based approach is contingent upon its capacity to provide a comprehensive linguistic analysis of text toxicity, culminating in the formulation of an expert opinion. This entails the identification of a list of lexical toxic units that characterize the text in question. In contrast, the considered machine learning method is limited to binary classification of text toxicity based on the features and degree of toxicity. It is worth noting that the lexicon-based method can be used to assess the toxicity of texts of varying lengths, even in the absence of toxic datasets. One disadvantage of the method based on dictionaries and rules is the limited nature of the lexicon. It is inherently incomplete because communication generates new means of expressing toxicity and new toxic discourses that transform neutral vocabulary into the vocabulary of destructive influence. This in turn requires the ongoing addition of new terms to the lexicon.

3.1. Lexicographic lists of toxic words

In the present study, our task was to compile lexical lists of toxic lexical means (including idiomatic expressions) regardless of the topic of toxic texts in Ukrainian-language online media discourse.

A toxic text is typically defined as an offensive comment or publication that exhibits one or more of the following characteristics: harassment, threat, obscenity, cyberbullying, trolling, indignation, and identity-based hate text [10, 11, 12]. The listed features are distinctly aggressive pragmatic practices of the sender in the communicative process. However, in our opinion, such a definition of a toxic text does not take into account the fact that a toxic text is the result of toxic communication, which may not contain an aggressive pragmatic stance but may include information about events, facts, phenomena that negatively affect the psychological and emotional state of both the sender and the recipient, or express emotions and emotional evaluations that reflect the psychological instability of the participants of the communication. Such texts have become a feature of wartime media communication, containing words such as: *війна (war), ворог (enemy), московський (muscovite), окупант (occupier), росія (russia), рф (rf), тривога (alarm)* etc., those that serve the function of destructive psychological influence; *страшно (scary), розлючений (angry), жахливий (terrible), наляканий (frightened)* etc., those that express negative evaluations. Therefore, in our study, the concept of "toxic text" is interpreted in a somewhat broader manner: A toxic text is defined as a text that not only contains indications of aggressive communication (harassment, threats, obscenities, cyberbullying, trolling, outrage, and identity-based hate speech) but also verbalizes negative facts, emotions, and assessments. These destructive emotion-generating words cause the recipient to experience anxiety, fear, confusion, shame, guilt, oppression, and control. Within the concept of toxicity, we also differentiate between hate speech as generally aggressive communication, and identity-based hate speech — texts that display aggression through discrimination based on a person's identity, such as nationality, race, skin color, origin, gender, health status, sexual orientation, religiosity, or other features.

Accordingly, the lexical lists of toxic vocabulary were compiled with the notion of this broader interpretation of the concept of a "toxic text" based on the following data: 1) A textual sample of approximately two million words [21], comprising texts from blogs, news sites, online publications, comments to online publications from social networks, and so forth; 2) a database of semantic taxa [5], compiled from Ukrainian journalistic texts totaling 40,000 words; 3) a tonality dictionary of Ukrainian vocabulary compiled by O. Tolochko [8]. The compilation of these lists was carried out automatically using specially developed software for data search and import.

3.1.1. The "Dictionary of Emotionogens"

Lexicographic List No. 1, the "Dictionary of Emotionogens," contains over 5,000 lexemes that verbalize negative facts, emotions, and assessments, causing the recipient to experience anxiety, fear, confusion, shame, guilt, oppression, and control. The dictionary includes words as independent parts of speech with a negative tonality (rated "-2" or "very negative"), namely nouns (e.g., *вада (flaw), вбивця (murderer), заборгованість (debt), ігнор (ignore), обвал (collapse), рабство (slavery), шантаж (blackmail)*), adjectives (e.g., *важкий (difficult), егоїстичний (selfish), облудливий (deceptive), радіаційний (radioactive), убивчий (deadly), ядерний (nuclear)*), adverbs (e.g., *важко (difficultly), задиркувато (provocatively), ризиковано (riskily), убивчо (lethally), шкідливо (harmfully)*), verbs (e.g., *вбивати (to kill), забороняти (to prohibit), завербувати (to subvert), ігнорувати (to ignore), обнулити (to nullify), ризикувати (to risk), шахраювати (to scam)*) and adjectives formed from verbs (e.g., *обдертий (torn), вибитий (knocked out), вибито (knocked), заблокований (blocked), ігнорований (ignored), розбазарений (squandered)*).

3.1.2. The "Hate Speech Dictionary"

The lexical list, designated No. 2, "Hate Speech Dictionary," contains 3,000 lexemes that verbalize aggressive communication, including harassment, threats, obscenities, cyberbullying, trolling, outrage, and identity-based hate text. The list covers the following lexical groups: negative names for people – 1,620 (e.g., *авантюрист (scammer), бабій (womanizer), гей (fagot), гопник (thug), грантожер (grant-eater,), даун (retard), жиробас (fatty), катюга (torturer), лайдак (scoundrel), лесбуха (lesbo), москаль (muscovite), п'явка (leech)*), obscene vocabulary – 613 (e.g., *ахуєнний (fucking awesome), бляхойоб (fucker), доєбатися (to fuck with), єбало (mug face), напиздник (scammer), перехуярити (to fuck up)*) and abusive vocabulary and vulgarisms – 787 (e.g., *афігівати (to freak out), бздун (coward), гадити (to shit), гівноблог (shitpost), дістати (to get fucking upset), довбаний (fucked up), капздець, лайно (bullshit), лохануться (to screw up), медіасрач (media shitstorm), набухатися (to get shitfaced), падло (bastard)*). Each item in the list is marked with semantic characteristics according to the developed classifications. For example, in the dictionary "Negative Names for People," the lexemes are grouped into 18 categories that correspond to key semantic features: negative designations of a person based on age (*бабулєта (old crone)*), sexual orientation (*ґей (gay), голубий (queer), гріховодник (sinner), лесбуха (lesbian), педик (fag)*), gender (*баба (woman)*), appearance (*дилда (beanpole), дистрофік (dystrophic), дрищ (skinny), жиробас (fatty)*), antisocial behavior (*аферист (swindler), аферистка (swindler), бандит (bandit), блазень (clown), гангстер (gangster), пахан (kingpin), перевертень (traitor)*), social activity / passivity (*ватник (pro-kremlin supporter), перебіжчик (defector), перебіжчиця (defector), пофігіст (indifferent)*), nationality (*бандерівець (banderite), укроп (ukrop), москаль (muscovite), кацап (katsap), жид (kike), негритос (negro)*), social status (*байстрюк (bastard), безхатько (homeless), бомжара (bum), бомжиха (bum)*), profession and financial status (*барига (pusher), банкрут (bankrupt), гендляр (trader), голяк (pauper), ділок (fence)*), political affiliation (*нацик (nazi), лівак (leftist)*), place of residence (*даунбас (downbas), педоросія (pedorussia), селюк (hillbilly)*), religion (*безбожник (atheist), сектант (sectarian)*), intellectual abilities (*балванчік (dummy), валянок (idiot), графоман (hack), довбограй (moron)*), diseases (*алкаш (drunkard), аутист (autistic), даун (down)*), and comparison to plants or animals (*амеба (amoeba), баран (sheep), видра (otter), вівця (lamb), жаба (frog), пацюк (rat), свиня (pig)*) etc. Each word of hate speech received has been annotated as belonging to one of the following categories: s – sexism, r – racism, e – ageism, etc.

3.1.3. Dictionary of Toxic Compounds

Lexicographic List No. 3, "Toxic Compounds," encompasses 1,500 stable phrases that idiomatically reflect toxic sentiment. The items in the list are classified by 26 semantic features, with each combination assigned a semantic label of toxicity (T), identity-based hate speech (IBYS), or expressiveness (E), e.g., IBYS 4. Claims of inferiority, moral flaws (*брудний циган (dirty gypsy), жиди скупі (stingy Jews)*) – IBYS 7. Mention in a derogatory or insulting context, obscene comments (*америкос тупий (dumb yank), баба з бородою (bearded woman), вшивий кацап (lousy katsap), жидівська пика (jewish mug), мусорам слово не давали (pigs have no word around here), гидувати содомітами і мужеложцями (disgusted by sodomites and homosexuals)*) – IBYS 10. Threats of physical destruction or any violence (*вбивати бандерівців (kill the banderites), смерть жидам (death to Jews)*) – T 1. Calls for violence (*голову розбити! (smash his head!), Гройсманяку на гілляку (hang Groysman), достатньо кількох повісити (a few hangings will suffice)*) – T 10. Bullying (*жирний осел (fat ass), кончений малолєтка (degenerate brat)*) – T 11. Caustic remarks and malicious jokes aimed at eliciting an emotional response from readers, or trolling (*безславна каденція (disgraceful tenure), Верховні зрадники (supreme traitors), вічний корупціонер (eternal corruptionist), панове тітушководи (hooligan puppeteers), печерний мракобіс (backward cavemen)*) – E 1. Emotional irritation of the message author, outrage about politics / life / friends, etc.(*атрофований мозок (atrophied brain), виключити хавальник (shut your trap), гнила сутність (rotten essence), грьобаний стид (bloody shame)*) – EPh. Phraseologisms with a toxic tonality (*базарна баба (a chatterbox of a woman), берега не бачити (to lose one's bearings), витирати ноги (to wipe one's feet on), відкривати вогонь (to open fire), вовк в овечій шкурі (a wolf in sheep's clothing), цілувати дупу (to kiss ass)*), etc.

3.2. Implementation within the TextAttributor 1.0 system

The described lexicon-based method for toxicity detection in Ukrainian-language texts was implemented within the TextAttributor 1.0 system as a rule-based module. The general characteristics of the rule-based toxicity detection module are as follows: the input is user-provided text, while the output is: 1) a numerical value of the statistical index of text toxicity; 2) the absolute frequencies of toxic words and phrases classified by semantic classes; 3) the text with toxic words and phrases highlighted.

3.2.1. Automatic Analysis Algorithm

The linguistic statistical analysis of toxicity is performed for each text using the program code for calling the analyzers in the following sequence of actions:

1. The text is tokenized into separate sentences and words;
2. Morphological annotation of words;
3. Followed by a contextual analysis that refines the morphological annotation codes;
4. Lemmatization of words;
5. Identification of toxic vocabulary in the analyzed text: comparison of words in the text with the registry of the toxic lexicon database and calculation of the absolute frequencies of toxic words and phrases;
6. Calculation of the text toxicity index;
7. Generation of a linguistic expert report: a statistical map of toxic vocabulary in addition to the text where toxic words and phrases are highlighted.

3.2.2. Toxic Vocabulary Database

The rule-based module operates on the basis of a lexicographic database of toxic tabled vocabulary, compiled from three lexicographic lists. The respective table (developed based on MS SQL Server) contains 9,500 rows.

Description of the data structure of the table:

[wid] – record identifier;

[did] – dictionary identifier (used for grouping phenomena by dictionary);

[wrd] – word form or lemma; if the column "sltype = 0", then it is a word form; if "sltype = 1", then this column would be a lemma;

[mitka] – sub-category label;

[updateDate] – record update date;

[updatedBy] – the user who updated the record;

[sdcatitemID] – category identifier;

[wrd2] – next words of the phrase after the first (if "sltype = 0") or the second lemma of the phrase, if `sltype=1`;

[wrd3] – the third lemma of the phrase, if "sltype = 1".

[sltype] – an indicator that this line contains a phrase stored as text or as lemmas.

3.2.3. Text toxicity index

The text toxicity index (I_{tox}) is calculated using the following formula:

$$I_{tox} = 10\left(e + |K|(m+t)\right)/n\,, \qquad (1)$$

where: e – the number of emotionogen words (Lexicographic List No. 1), m – the number of hate speech words (Lexicographic List No. 2), t – the number of toxic phrases (Lexicographic List No. 3) in the text analyzed by the system; K – an intensifier of aggressive toxic speech units (K = 2); n – number of words in the analyzed text; 10 – normalization factor for the statistical parameter.

The formula for the statistical parameter – the text toxicity index – takes into account the usage frequency for various classes of vocabulary in the text, differentiated by the semantic features of three lexicons: emotionogenic words (*e*), hate speech lexemes (*m*), toxic phraseological compounds (*t*). To enhance the weight of toxic means of aggressive communication, a coefficient *K* (on a five-point scale [-2, -1, 0, +1, +2] it corresponds to -2) has been introduced into the formula. This coefficient intensifies the weight of hate speech words (Lexicographic List 2) and toxic phraseological compounds (Lexicographic List 3) in determining the level of toxicity. The procedure of multiplying by 10 has been included into the formula to increase the empirical value of the toxicity index for normalizing the linguistic statistical parameters of the stylometric analysis in the TextAttributor 1.0 system.

3.2.4. Results of the Rule-Based Toxicity Detection Module

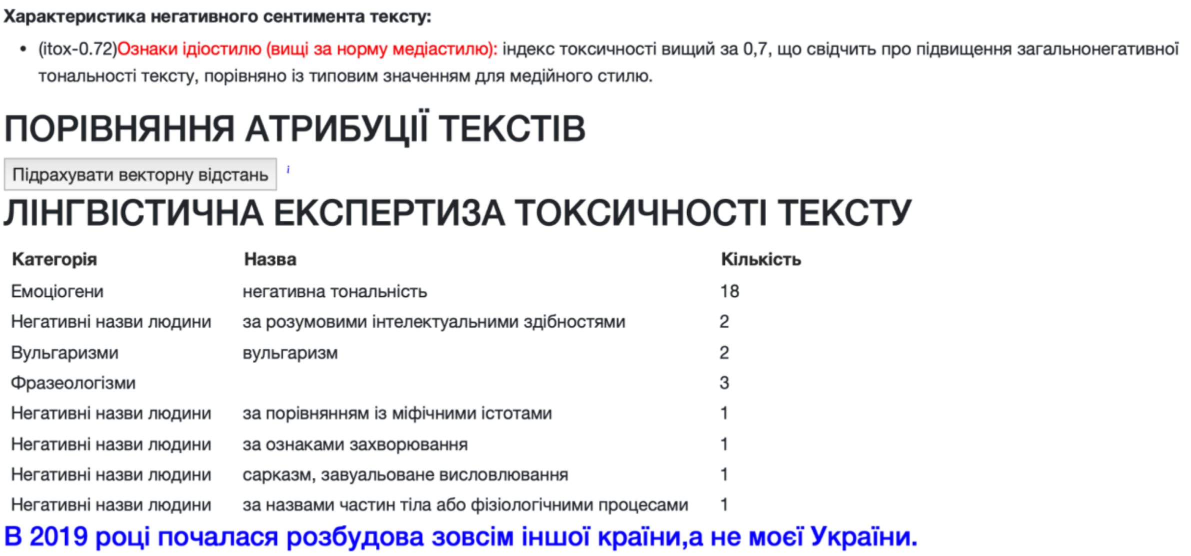
Let us consider the results of toxic speech analysis using a sample text: *«В 2019 році почалася розбудова зовсім іншої країни, а не моєї України» ("In 2019, a completely different country began to form, rather than my Ukraine"),* which consists of 52 sentences and 500 words. The toxicity index of the analyzed text (Itox) has been assessed at 0.7, indicating a slightly higher level of toxicity compared to the upper limit of the average toxicity value for the media style in the Ukrainian language (Fig. 1).

The system generates a statistical map of the linguistic expert report on text toxicity (Fig. 1) which presents a list of word semantic classes identified by the system in the analyzed text, according to the classification markers of the toxic vocabulary database: Semantic category – column 1; name of semantic feature – column 2; number of words/phrases in the analyzed text according to semantic features – column 3. In particular, the statistical map of the text *"In 2019, a completely different country began to form, rather than my Ukraine"* systematizes the following linguistic data: emotionogens – 18; negative names for a person based on intellectual ability – 2; vulgarisms – 2; negative names for a person by comparison with mythical creatures – 1; negative names for a person based on health characteristics – 1; negative names for a person (sarcasm, idiomatic expression) – 1; negative names for a person based on body parts or physiological processes – 1.

The verbalization of statistics for identified toxic vocabulary in the text, presented in the statistical map according to semantic categories, is visualized in a separate window using bold black font marking specific lexical means (Fig. 2).

By comparing the statistical map data with the text, one can systematize toxic means, for example: emotionogens – 18 (nouns: *війна (war), жах (horror), жертва (victim), жорстокість (cruelty), загибель (death), корупція (corruption), пропаганда (propaganda), психлікарня (psychiatric treatment), терор (terror)*; adjectives: *болючий (painful), некерований (uncontrollable), приречений (doomed), страшний (terrible)*; adverb: *на жаль (unfortunately)*; verbs: *воювати (to fight), ненавидіти (to hate), ховати (to bury)*); vulgarisms – 2 (*бидло (scum), кончений (finished)*); toxic phrase *лизати дупу (to kiss ass)*; negative person descriptors based on various characteristics (*журнашлюха (journawhore), ідіот (idiot), сатана (satan), совок (sovok)*).

The provided text also includes indication of other features: 1) phraseoligisms are highlighted in blue font (*правити бал (to rule the roost), мати на увазі (to have in mind)* etc.); some words and symbols not recognized by the system's automatic morphological analysis are highlighted in red font (*блть, гауляйтер, Лисі, кварталити, Монатики, оркостан, рсня, Усики, хіхіх, Чонгар*). The system does not consider these units when calculating the text toxicity index. Testing the system's performance demonstrates the need to refine the lexicographic database, especially regarding neologisms, author's individual word usage, proper names, and special abbreviations, as well as to enhance the text preprocessing and lemmatization procedure. We consider this one of the primary prospective tasks.



Характеристика негативного сентимента тексту:

- (itox-0.72)Ознаки ідіостилю (вищі за норму медіастилю): індекс токсичності вищий за 0,7, що свідчить про підвищення загальнонегативної тональності тексту, порівняно із типовим значенням для медійного стилю.

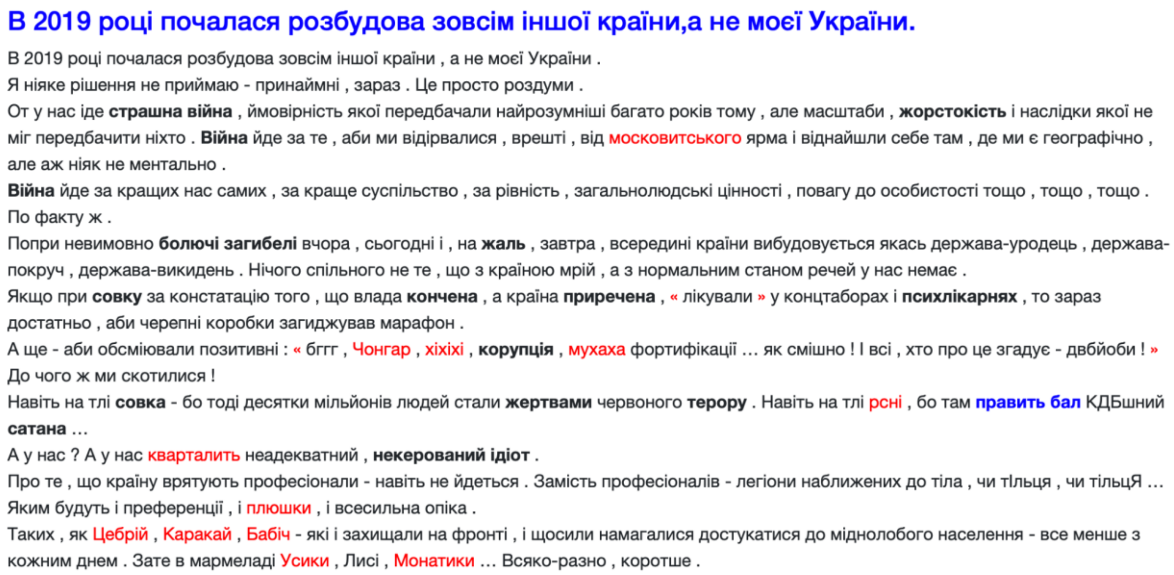
ПОРІВНЯННЯ АТРИБУЦІЇ ТЕКСТІВ

ЛІНГВІСТИЧНА ЕКСПЕРТИЗА ТОКСИЧНОСТІ ТЕКСТУ

Категорія	Назва	Кількість
Емоціогени	негативна тональність	18
Негативні назви людини	за розумовими інтелектуальними здібностями	2
Вульгаризми	вульгаризм	2
Фразеологізми		3
Негативні назви людини	за порівнянням із міфічними істотами	1
Негативні назви людини	за ознаками захворювання	1
Негативні назви людини	сарказм, завуальоване висловлювання	1
Негативні назви людини	за назвами частин тіла або фізіологічними процесами	1

В 2019 році почалася розбудова зовсім іншої країни,а не моєї України.

Figure 1: The text toxicity index and the statistical map of the linguistic expert report on text toxicity. All categories under the **Категорія** column are explained below.



В 2019 році почалася розбудова зовсім іншої країни,а не моєї України.

В 2019 році почалася розбудова зовсім іншої країни , а не моєї України .

Я ніяке рішення не приймаю - принаймні , зараз . Це просто роздуми .

От у нас іде **страшна війна** , ймовірність якої передбачали найрозумніші багато років тому , але масштаби , **жорстокість** і наслідки якої не міг передбачити ніхто . **Війна** йде за те , аби ми відірвалися , врешті , від московитського ярма і віднайшли себе там , де ми є географічно , але аж ніяк не ментально .

Війна йде за кращих нас самих , за краще суспільство , за рівність , загальнолюдські цінності , повагу до особистості тощо , тощо , тощо .

По факту ж .

Попри невимовно **болючі загибелі** вчора , сьогодні і , на **жаль** , завтра , всередині країни вибудовується якась держава-уродець , держава-покруч , держава-викидень . Нічого спільного не те , що з країною мрій , а з нормальним станом речей у нас немає .

Якщо при **совку** за констатацію того , що влада **кончена** , а країна **приречена** , « лікували » у концтаборах і **психлікарнях** , то зараз достатньо , аби черепні коробки загиджував марафон .

А ще - аби обсміювали позитивні : « бггг , Чонгар , хіхіхі , **корупція** , мухаха фортифікації … як смішно ! І всі , хто про це згадує - двбйоби ! »

До чого ж ми скотилися !

Навіть на тлі **совка** - бо тоді десятки мільйонів людей стали **жертвами** червоного **терору** . Навіть на тлі рсні , бо там **править бал** КДБшний **сатана** …

А у нас ? А у нас кварталить неадекватний , **некерований ідіот** .

Про те , що країну врятують професіонали - навіть не йдеться . Замість професіоналів - легіони наближених до тіла , чи тІльця , чи тільцЯ …

Яким будуть і преференції , і плюшки , і всесильна опіка .

Таких , як Цебрій , Каракай , Бабіч - які і захищали на фронті , і щосили намагалися достукатися доміднолобого населення - все менше з кожним днем . Зате в мармеладі Усики , Лисі , Монатики … Всяко-разно , коротше .

Figure 2: Text fragment analyzed in the rule-based toxicity detection module. All spotted words are explained below.

4. Machine learning methods for toxicity detection of Ukrainian-language texts

4.1. Data

For our experimental study of toxicity in media texts using machine learning methods, two datasets containing short Internet texts from Ukrainian-language blogs, comments, articles, etc., have been prepared.

Dataset 1. For the initial analysis of the issue, a trial corpus of media texts was formed, consisting of 668 documents featuring hate speech and expert annotations (each text sample was classified by an expert into a specific category — neutrality or hate speech). 226 texts were identified as toxic. The obtained dataset was randomly split into a training set (600 texts, 192 of which are toxic, 94,905 word samples, 11,852 stems) and a test set (68 texts, 34 of which are toxic, 10,800 word samples).

Dataset 2. The training set was augmented by incorporating annotated 11,387 text documents, among which 2,155 are toxic. The test set remained unchanged.

4.2. Linguistic Features

The typical pipeline of Machine Learning methods involves preprocessing the text data, extracting features, and then using a classifier to determine if the text is toxic. Feature extraction techniques include:

- Bag-of-Words (BoW) is a simple approach where text is represented as an unordered collection of words, ignoring grammar and word order but preserving frequency.
- Term Frequency-Inverse Document Frequency (TF-IDF) enhances the BoW model by weighing word frequency relative to its importance across documents, thus reducing the weight of commonly used words.
- Pre-trained word embeddings like Word2Vec or GloVe can be used to represent words as dense vectors, capturing semantic similarities.
- N-grams capture short phrases or word sequences to better model context compared to BoW, useful for identifying common toxic word patterns.
- Other linguistic features may include sentiment scores, lexical resources like swear word dictionaries, Part-of-Speech (POS) tags, and syntactic features.
- In experimental research described in this paper we rely on BoW and word embeddings.

In experimental research described in this paper we rely on BoW and word embeddings.

4.3. Classical Machine Learning Algorithms

Classical, or primitive, Machine Learning Approaches use linguistic features and classical algorithms like SVM, logistic regression and Naive Bayes. They require significant domain knowledge and are limited in handling complex language patterns.

Naive Bayes: Suitable for text classification due to the assumption of word independence, which often works well despite its simplicity.

Logistic Regression: A widely used linear model for binary classification.

Support Vector Machines (SVM): Works well with high-dimensional data like text, especially when combined with kernel methods to separate non-linear toxic and non-toxic classes.

Random Forest and Decision Trees: Tree-based models are sometimes used to capture non-linear relationships, but they tend to struggle with sparse data in large text corpora.

In this paper we describe experiments exploiting the Naive Bayes approach.

4.4. Deep Learning Methods

Deep Learning Approaches leverage neural networks, RNNs, CNNs, LSTMs, and especially transformers like BERT for sophisticated contextual understanding of text. These models can automatically learn features but require substantial computational resources and large datasets.

Hybrid Approaches combine the strengths of both classical and deep learning methods for better performance and robustness.

Challenges across both methods include handling nuanced language, interpretability, data imbalance, and avoiding bias.

In this paper our choice is fastText [22], an efficient and widely-used method for word representation and text classification, which can be used as a feature extraction technique in text toxicity detection.

FastText's main advantages are its relative simplicity, speed, and ability to handle large vocabularies and datasets. It is particularly efficient because it:

- uses rather shallow neural network, which is much faster than deep models,
- embeds subword (character n-grams) information, allowing better generalization,
- utilizes hierarchical softmax for computational efficiency during classification, especially for large-scale problems.

The subword presentation is important since the Ukrainian language is highly inflective, and plenty of unseen words as well as words with spelling errors must be covered. The effectiveness of the selected approach is also determined by the technical conditions of system operation and available textual resources for model training.

4.5. LLM Prompting

Text toxicity detection using Large Language Models (LLMs), such as GPT-3 or GPT-4, is an emerging area in Natural Language Processing (NLP) that leverages advanced capabilities of LLMs for identifying harmful, abusive, or toxic language in text. Here are the methods and strategies for using LLM prompting to detect toxicity:

- Direct Toxicity Classification Prompting
- Chain-of-Thought Prompting (Step-by-Step Reasoning)
- Few-Shot Learning (Providing Examples)
- Zero-Shot Prompting with Contextual Framing
- Prompt Tuning for Toxicity Detection
- Debiasing and Calibration Prompting
- Toxicity Detection as Part of a Larger System (Hybrid Models)

In this work we consider one of the most straightforward methods involves directly prompting the LLM to classify text as "toxic" or "non-toxic" exploiting Llama v. 3 [23]. This was tested by providing the LLM with a clear instruction such as:

> Classify the following sentence as 'toxic' or 'non-toxic':
> "You are an idiot and nobody likes you."

The LLM is provided with text snippets, and based on its training and understanding of language, it identifies the toxic content. Implementation of this approach is quick and simple, no fine-tuning required. Another advantage is that LLMs can identify nuanced and context-dependent toxicity that may escape simpler classifiers. On the other hand, (a) LLMs may not always be consistent or reliable without additional constraints or examples, (b) false positives or negatives could occur due to ambiguity or complexity in language, (c) huge amounts of computation resources are required, including a powerful GPU with 24GB of RAM.

4.6. Results

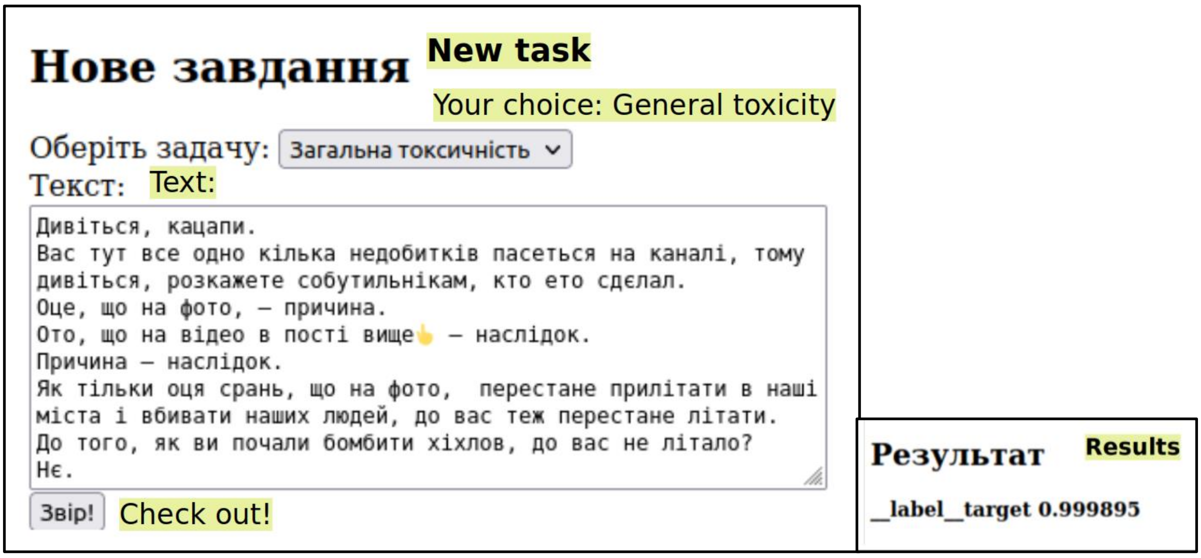
Evaluating models for toxicity detection requires a careful choice of metrics because of the class imbalance and Precision, Recall, and F1-Score meets these requirements: precision helps in minimizing false positives (important to avoid over-censoring), while recall ensures detection of most toxic content and F1 generalizes these two metrics.

- Classical techniques: The following results were obtained from the testing of Dataset 1 and 2: F1 – 71 %; precision – 60 %; recall – 85 %.
- Deep Learning: The best result according to the generalized metric (F1 = 79.4%) was obtained for the following hyperparameter values: the dimension of the space of the vector representation of words – 56, the initial learning rate – 0.15, the number of epochs – 500, the lexical context – bigrams, the length of subwords – from 2 to 5, the decision-making threshold – 0.4. At the same time, for a sensitivity of 85%, an accuracy of about 70% is achieved.

- LLM prompting: The following results were obtained on the test sample: F1 – 75.3 %; precision – 74.3 %; recall – 76.5 %.

4.7. Integration with the TextAttributor 1.0 system

The experimental system for determining the toxicity of media texts based on ML is implemented in a "client-server" architecture using a REST interface. A basic web user interface has been developed, allowing users to input text and receive a response from the system on whether the text is toxic (__label__target) with a confidence score ranging from 0 to 1. Figure 3 shows an example of using the basic web interface to assess the toxicity of a given media text. In turn, using the REST API, TextAttributor acts as a client, sending the user-entered text to the server, receiving a response, and visualizing the result along with other parameters calculated for the given text.



Оберіть задачу: Загальна токсичність

Дивіться, кацапи.
Вас тут все одно кілька недобитків пасеться на каналі, тому дивіться, розкажете собутильнікам, кто ето сделал.
Оце, що на фото, – причина.
Ото, що на відео в пості вище👆 – наслідок.
Причина – наслідок.
Як тільки оця срань, що на фото, перестане прилітати в наші міста і вбивати наших людей, до вас теж перестане літати.
До того, як ви почали бомбити хіхлов, до вас не літало?
Нє.

__label__target 0.999895

Figure 3: Displaying the results of toxicity analysis in a basic web interface.

5. Experiments and discussions

Let us consider the experimental comparison of the toxicity indices of a single text, determined by the two methods considered.

5.1. Systematization of experimental data and the gradation of the toxicity index

The experimental study aims to compare the results of the system's performance using two methods – rules-based and machine learning – to determine the effectiveness of these methods in identifying toxic texts. To achieve this goal, a sample of 520 texts was compiled and analyzed using the TextAttributor 1.0 system. The sample included texts of varying lengths (from 33 words to 46,000 words), encompassing different themes, styles, and genres. Each text was evaluated for toxicity using two indices: ToxR is the toxicity index based on the lexicon and rule-based method, while ToxML is the toxicity index as determined by the machine learning method.

Using the two methods in the TextAttributor 1.0 system, toxicity grading occurs on two scales: ToxR is graded on a scale from 0 to 5.78, while ToxML, based on machine learning, ranges from 0 to 1.00. ToxML is limited to a range of 0–1.00, with a decision threshold for text toxicity set at 0.4 (0.35). (neutral text) 0 < low-toxicity text < 0.4 ≤ toxic text ≤ 1.00. The decision threshold was determined experimentally based on the best F1 score for the trained model. In contrast, ToxR has no upper limit or decision-making threshold value for the system. Establishing toxicity using this method involves the identification of lexical markers of text toxicity, their statistical evaluation—the text toxicity index—while the decision on toxicity is left to the linguist.

During the linguistic analysis of the toxicity of texts examined by the system, the average ToxR value for the media style of the Ukrainian language was empirically measured at 0.4 (0.35). This value will be considered as the decision threshold on a text's toxicity, despite the fact that this numerical value results from statistical analysis of entirely different features than those used in the calculation of ToxML. Empirical numerical values of toxicity indices, determined by two methods for each text in the sample, were ranked in descending order of numerical values: from highest to lowest. Ranking the toxicity indices, calculated by the two methods, allows for the distribution of texts according to toxicity levels:

- ToxML: toxic texts with an index of 0.4+ constitute 52.7% of the sample (276 texts).
- ToxR: toxic texts with an index of 0.4+ constitute 58.8% of the sample (306 texts).

The remaining texts in the sample exhibit either a low degree of toxicity (0 < 0.4) or are devoid of toxic characteristics (neutral text = 0):

- ToxML: low-toxicity texts with an index below 0.4 constitute 37.7% of the sample (196 texts); neutral texts with a toxicity level of 0 constitute 9.6% (50 texts).
- ToxR: low-toxicity texts with an index below 0.4 constitute 37.5 % of the sample (195 texts); neutral texts with a toxicity level of 0 constitute 3.7 % (19 texts).

The percentage ratio of toxic and low-toxicity texts, as determined by the two methods, demonstrates close results; however, the numerical toxicity indicators determined by the two methods for a single text can be drastically different. Ranking texts by descending numerical values of ToxML while preserving the numerical values of ToxR for these texts indicates that:

1. 276 texts are toxic by ToxML; however, 84 of these texts exhibit low toxicity (ToxR < 0.4) or are non-toxic (ToxR = 0) according to the ToxR method, meaning only 192 texts are assessed as toxic by both methods.
2. 244 texts are low-toxicity or neutral (ToxML < 0.4), but 107 of these texts exhibit high toxicity according to the ToxR method (0.4 ≤ ToxR), meaning that only 137 texts are assessed by the system as low-toxicity/non-toxic by both methods.

Thus, it can be stated that 329 out of 520 texts were classified as toxic/low-toxicity/emotionally neutral by both methods. Thus, we are presented with a question: Is there a statistical relationship between the two variables—empirical numerical values of the toxicity indices determined by the two methods?

5.2. Pearson correlation between toxicity indices determined by two methods

Comparing the two sets of numerical data – ToxML and ToxR toxicity indices – we have formulated a task to determine whether there is a statistical dependency in the changes of ToxML and ToxR toxicity levels on a sample of 520 texts, i.e., to ascertain the degree of correlation between the two data sets: the ranked list of ToxML and corresponding numerical values of ToxR. The correlation between the two data sets was measured using three methods: linear Pearson correlation (0.2068); Spearman rank correlation (0.2513); Kendall rank correlation (0.1753).

Let us consider the value of Pearson's linear correlation coefficient, which is most frequently used in humanities and social studies to analyze correlation rather than causal dependency. The correlation coefficient r = 0.2068 is low and indicates a weak positive correlation. However, with an error p = 0.001, considering the degrees of freedom f = 600, the critical correlation coefficient value is r = 0.134. The empirical value of Pearson's coefficient, r = 0.2068, is higher than the critical value and demonstrates high significance for a sample size of 520 texts. This implies that, although the degree of correlation is weak, the indicator has high reliability.

Let us consider a scatter plot for Pearson's coefficient of variation (0.2068), which indicates a weak measure of joint variability between two random variables – two toxicity indices, ToxML and ToxR, determined by two methods for a single text (Fig. 4). For the convenience of graphical modeling of correlation, the empirical values of ToxR were converted to a scale from 0 to 1 using the min-max normalization technique: $X_{norm} = (X_i - X_{min}) / (X_{max} – X_{min})$.

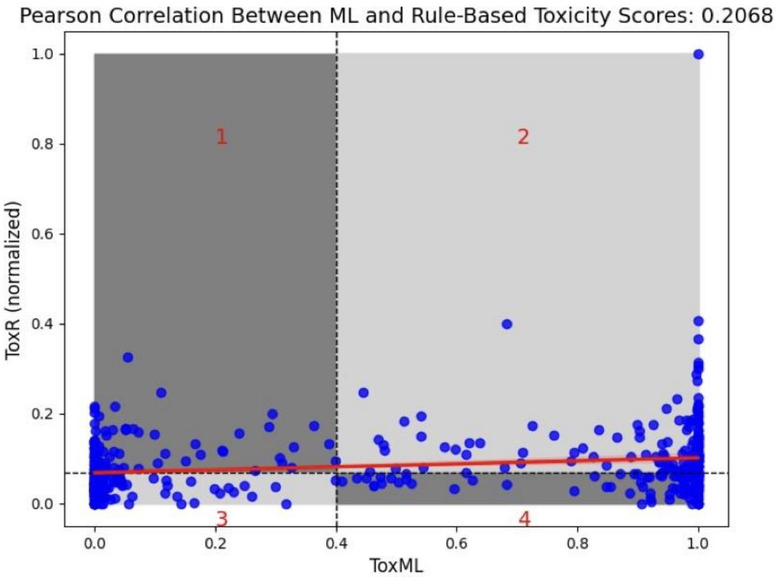


Figure 4: Scatter plot with Pearson's coefficient of variation for all 520 collected customer's texts.

X-axis: numerical values of toxicity levels determined by the machine learning method (ToxML – independent variables); Y-axis, the numerical values of the toxicity index determined by the rule-based method (ToxR – dependent variables). Each point on the diagram represents a text, with its coordinates being two toxicity indices. The red line is a regression line modeling the relationship between the two variables. Given a weak positive Pearson coefficient, the regression line has a low slope. This indicates a vague trend of increasing Y values with the increase in X values. Considering the decision threshold of 0.4 for determining text toxicity/low toxicity for ToxML and 0.07 (after min-max normalization) for ToxR, four zones were identified on the graph: 1) Texts toxic by ToxR, but low-toxicity or neutral according to ToxML – dark gray zone; 2) Texts toxic according to both methods – light gray zone; 3) Texts considered low-toxicity or neutral according to both methods – light gray zone; 4) Texts toxic by ToxML, but low-toxicity or neutral by ToxR – dark gray zone. The correlation of variables (ToxR and ToxML) for texts in zone 1 and zone 4 consequently affects the coefficient of variation, as an inversely proportional correlation is observed in these zones. Low X values correspond to high Y values (Zone 1), whereas high X values correspond to low Y values (Zone 4). We consider the tests of these areas to be contentious with respect to toxicity indicators.

The results of automatic toxicity analysis also demonstrate similar statistical data by the two methods in determining the toxicity/low-toxicity/neutrality of texts, which, according to the scatter plot, are located in zones 2 and 3. To determine the degree of correlation between the statistical data obtained using the two methods for these texts, Pearson's correlation coefficient was calculated for a sample of 329 texts, of which 192 are toxic, 133 are low-toxicity, and 4 are neutral texts. The correlation coefficient r = 0.5344 is relatively high and indicates a moderate positive correlation. Moreover, the average degree of correlation has high reliability, as with p = 0.001 and considering the degrees of freedom f = 350, the critical correlation coefficient value is r = 0.175. The empirical value of Pearson's coefficient, r = 0.5344, is higher than the critical value and demonstrates high significance for a sample size of 329 texts.

A scatter plot was also constructed for Pearson's coefficient of variation (0.5344) (Fig. 5).

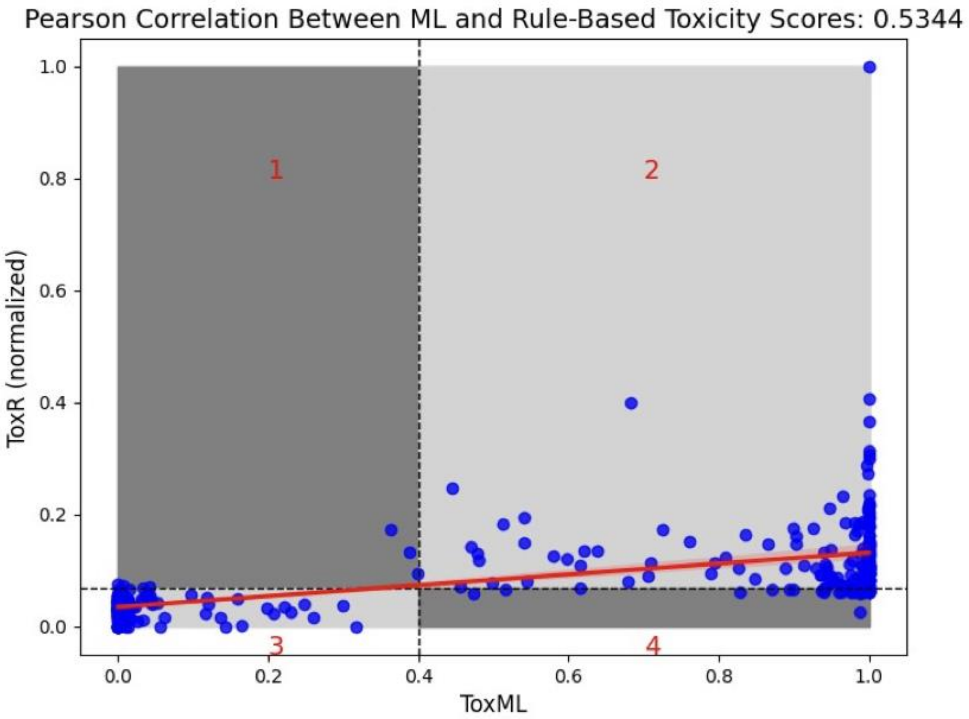


Figure 5: Scatter plot with Pearson's coefficient of variation for 329 customer's texts with similar statistical data by the two methods in determining the toxicity/low-toxicity/neutrality.

The regression line, which models the average positive correlation coefficient on the diagram, has a significantly greater slope compared to the diagram in Fig. 4, which indicates a moderate measure of joint variability between the two random variables ToxR and ToxML. This indicates a clear trend of increasing Y values with the increase in X values.

In Figure 5, four zones were identified using the same principle: 1) Texts toxic by ToxR, but low-toxicity or neutral according to ToxML - dark gray zone; 2) Texts toxic according to both methods - light gray zone; 3) Texts considered low-toxicity or neutral according to both methods - light gray zone; 4) Texts toxic by ToxML, but low-toxicity or neutral by ToxR - dark gray zone. The correlation of variables (ToxR and ToxML) shows that texts rarely fall into zones 1 and 4; instead, the points in zone 2 clearly represent low-toxicity and neutral texts, while zone 4 contains points representing texts with high toxicity. With an average positive correlation coefficient (Fig. 5), the dispersion of points is significantly lower than with a weak positive coefficient (Fig. 4), yet in both cases, "outliers" from the model's range of inference are observed. These are points of texts that deviate from the general trend of variable correlation and are typically characterized by high levels of toxicity according to one or both methods.

The diagram also clearly shows the formation of three clusters: 1) Cluster of points in the range of 0–0.3 (on the X-axis) represents neutral texts and texts with low toxicity; 2) Cluster of points in the range of 0.8–1.0 (on the X-axis) represents texts with high toxicity; 3) Dispersion of points in the range of 0.4–0.8 represents texts of medium toxicity. In the range of 0.3–0.4, there are 5 points—texts whose degree of toxicity is difficult to interpret because they fall within the decision threshold range. However, 2 points are closer to cluster 1 (neutral texts and low-toxicity texts), while 3 points are closer to cluster 3 (medium toxicity texts).

6. Conclusions and future work

The development and testing of the "TextAttributor 1.0" system's toxicity detection modules offer promising outcomes, particularly in the automatic identification of toxic content in Ukrainian texts. The system combines a rule-based and a machine learning approach, both of which provide valuable, yet distinct, insights into text toxicity. The rule-based module excels in providing a detailed linguistic analysis of toxic vocabulary based on a lexicographic database, making it suitable for in-depth expert analysis. On the other hand, the machine learning module provides a scalable solution for handling large volumes of text, offering an efficient and automated method of classifying toxic content.

However, a moderate correlation between the two methods, as demonstrated by the Pearson correlation coefficient, reveals some discrepancies in toxicity assessment. These differences underline the need for further refinement in both modules to improve the overall system accuracy and reliability.

Key tasks for further improvement include enhancing the machine learning model to analyze full-length texts rather than truncated segments and expanding its training dataset to increase accuracy for longer texts. Additionally, testing more advanced models such as those from the BERT family could further improve classification performance. On the rule-based side, recalibration of the toxicity index formula is required to address limitations in how text size and single occurrences of aggressive language impact the final score.

One of the most promising directions for future work is the integration of rule-based and machine learning methods into a hybrid model. Such a model would leverage the strengths of both approaches, applying rule-based analysis for in-depth, context-sensitive interpretation and machine learning for efficient large-scale classification. This hybrid approach could also enhance the system's ability to detect more subtle forms of toxicity, such as covert hate speech or context-dependent insults.

Further development of the user interface to display toxicity results more intuitively would enhance the system’s usability for non-technical users. Visual tools such as heatmaps of toxic word distributions, graphs showing the progression of toxicity throughout a text, and interactive dashboards could make the system more accessible for use in various domains.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Dictionary of media literacy, 2022. URL: https://filter.mkip.gov.ua/wp-content/uploads/2022/10/slovnyk.pdf [in Ukrainian].

[2] V.M. Petryk, M.M. Prysiazhniuk, L.F. Kompantseva, Ye.D. Skulysh, O.D. Boiko, V.V. Ostroukhov. Suggestive technologies of manipulative influence: study guide. Kyiv, 2011 [in Ukrainian].

[3] Recommendation CM / Rec (2022) 16 of the Committee of Ministers to member States on combating hate speech. URL: https://www.coe.int/en/web/combating- hate-speech/recommendation-on-combating-hate-speech.

[4] TextAttributor 1.0, 2024. URL: http://ta.mova.info

[5] N. Darchuk, O. Zuban, M. Lanhenbakh, Ya. Khodakivska. AGAT-semantics: semantic markup of the Ukrainian language corpus. Ukrainian linguistics, 1(46), 2016, 92–102. https://doi.org/10.17721/um/46 [in Ukrainian].

[6] N. Darchuk. Linguistic approach for development of computer-based sentiment analysis in the Ukrainian language. Science and education a new dimension, 189, (2019), 10–13 [in Ukrainian].

[7] O. Zuban, V. Bilous. Automatic determination of lexical tone of Ukrainian-language text (based on TSN news). In V International Scientific and Practical Conference “Information Technology and Interaction (IT&I’2018)”, 2018, pp. 182–183. http://iti.fit.univ.kiev.ua/wp-content/uploads/ITI-2018.pdf [in Ukrainian].

[8] O. Tolochko. Tonal dictionary of the Ukrainian language, 2023. URL: https://github.com/Oksana504/sentimentdictionary-uk.

[9] K. Bobrovnyk. Automated building and analysis of ukrainian twitter corpus for toxic text detection. In COLINS 2019. Volume II: Workshop, 2019, pp. 55-56.

[10] F. Alkomah, X. Ma. A Literature Review of Textual Hate Speech Detection Methods and Datasets, Information 13.6 (2022). doi: 10.3390/info13060273.

[11] Jahan, M. S., & Oussalah, M. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546 (2023), 126232. doi: 10.1016/j.neucom.2023.126232.

[12] A. Bonetti, M. Martínez-Sober, J. C. Torres, J. M. Vega, S. Pellerin, and J. Vila-Francés, Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks, Applied Sciences 13.10 (2023). doi: 10.3390/app13106038.

[13] F. Rangel, G. Sarracén, B. Chulvi, E. Fersini, P. Rosso. Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. In Proceedings of the CLEF 2021–Conference and Labs of the Evaluation Forum, 2021.

[14] S. Frenda, B. Ghanem, M. Montes-Y-Gómez, P. Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. J. Intell. Fuzzy Syst. 36 (2019), 4743–4752. doi: 10.3233/JIFS-179023.

[15] K. Machová, M. Mach, K. Adamišín. Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions. Sensors, 22(17) (2022), 6468. doi: 10.3390/s22176468.

[16] F.-Z. El-Alami, S. Ouatik El Alaoui, and N. En-Nahnahi. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. J. King Saud Univ. Comput. Inf. Sci., 34(8 Part B) (2022), 6048– 6056. doi: 10.1016/j.jksuci.2021.07.013.

[17] A. G. D'sa,, I. Illina, D. Fohr. Bert and fasttext embeddings for automatic detection of toxic speech. In: 2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies” (OCTA). IEEE, 2020. p. 1-5. doi: 10.1109/OCTA49274.2020.9151853

[18] M. A. H. Wadud, M. F. Mridha, J. Shin, K. Nur., A. K. Saha. Deep-BERT: Transfer Learning for Classifying Multilingual Offensive Texts on Social Media. Computer Systems Science & Engineering, 44(2) (2022). doi: 10.32604/csse.2023.027841.

[19] D. Dementieva, V. Khylenko, N. Babakov, G. Groh. Toxicity Classification in Ukrainian. In Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), 2024, pages 244–255. doi: 10.18653/v1/2024.woah-1.19.

[20] V. Oliinyk and I. Matviichuk. Low-resource text classification using cross-lingual models for bullying detection in the Ukrainian language. Adaptive systems of automatic control: interdepartmental scientific and technical collection, 1 (42) (2023). doi: 10.20535/1560-8956.42.2023.279093

[21] Corpus of the Ukrainian language, 2003–2024. Mova.info: linguistic portal. URL: http://www.mova.info/corpus.aspx [in Ukrainian].

[22] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, FastText.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651, 2016. doi: 10.48550/arXiv.1612.03651.

[23] Meta-Llama. Inference code for Llama models, 2024. URL: https://github.com/meta-llama/llama/.