Półrocznik Językoznawczy Tertium. Tertium Linguistic Journal 8 (2) (2023)
www.journal.tertium.edu.pl
https://doi.org/10.7592/Tertium.2023.8.2.245

A Morpho-syntactic Analysis of Human-moderated Hate Speech Samples from Wykop.pl Web Service

Inez Okulska
NASK National Research Institute, Warszawa, Poland
inez.okulska@nask.pl

Anna Kołos
NASK National Research Institute, Warszawa, Poland
anna.kolos@nask.pl

Abstract

The dynamic increase in user-generated content on the web presents significant challenges in protecting Internet users from exposure to offensive material, such as cyberbullying and hate speech, while also minimizing the spread of wrongful conduct. However, designing automated detection models for such offensive content remains complex, particularly in languages with limited publicly available data. To address this issue, our research collaborates with the Wykop.pl web service to fine-tune a model using genuine content that has been banned by professional moderators. In this paper, we focus on the Polish language and discuss the notion of datasets and annotation frameworks, presenting our stylometric analysis of Wykop.pl content to identify morpho-syntactic structures that are commonly applied in cyberbullying and hate speech. By doing so, we contribute to the ongoing discussion on offensive language and hate speech in sociolinguistic studies, emphasizing the need to consider user-generated online content.

Keywords: cyberbullying, hate speech, user-generated online content, automated detection, stylometry

Streszczenie

Morfosyntaktyczna analiza przykładów mowy nienawiści zablokowanych przez moderatorów serwisu Wykop.pl

Dynamiczny wzrost treści generowanych przez użytkowników w sieci stanowi poważne wyzwanie w zakresie ochrony użytkowników Internetu przed narażeniem na obraźliwe materiały, takie jak cyberprzemoc i mowa nienawiści, i jednoczesnego ograniczania rozprzestrzeniania nieetycznych zachowań. Jednak projektowanie zautomatyzowanych modeli wykrywania obraźliwych treści pozostaje złożonym zadaniem, szczególnie w językach o ograniczonych publicznie dostępnych danych. W naszych badaniach współpracujemy z serwisem internetowym Wykop.pl w celu uczenia modelu przy użyciu rzeczywistych treści, które

podlegały usunięciu w procesie moderacji. W niniejszym artykule skupiamy się na języku polskim i omawiamy pojęcie zbiorów danych i metod anotacji, a następnie przedstawiamy naszą analizę stylometryczną treści z serwisu Wykop.pl w celu zidentyfikowania struktur morfosyntaktycznych, które są powszechnie aplikowane w języku cyberprzemocy i mowie nienawiści. Dzięki naszym badaniom mamy nadzieję na wniesienie wkładu w toczącą się dyskusję na temat obraźliwego języka i mowy nienawiści w badaniach socjolingwistycznych, podkreślając potrzebę analizy treści generowanych przez użytkowników w sieci.

Słowa kluczowe: cyberagresja, mowa nienawiści, treści internetowe, automatyczne wykrywanie treści, stylometria

1. Introduction

Due to the rapid increase in user-generated online content, there has been in recent years an emerging challenge to tackle the question of harmful speech, covered by such broad umbrella terms as hate speech, offensive language, cyberbullying, etc. Interdisciplinary studies addressing these notorious issues cover a wide range of academic disciplines, to name only law and international law, psychology, social psychology, sociology, media studies, political science, linguistics, and communication studies (Jaszczyk-Grzyb, 2021: 13–16; Guillén-Nieto, 2023: 1–21). Collaterally, recognizing the insufficiency of relying solely on human content moderation, the tech industry and machine learning have made significant efforts to develop automated technologies for detecting harmful content on the web. These advancements are crucial in effectively protecting internet users from exposure to illicit materials and harmful comments.

In our research (funded by The National Centre for Research and Development), we have established a collaboration with Wykop.pl moderators to develop a deep learning technology aimed at supporting the detection of wrongful content in alignment with the website's internal policy. To achieve this, we employ a combination of state-of-the-art transformer-based language models and stylometry, following a human-in-the-loop approach that leverages human expertise to train the algorithms. To facilitate the training and fine-tuning of our language model for automated detection, we utilize a tool called StyloMetrix[1] (Okulska et al., 2023a), operating within a Python environment, which provides an advanced stylometric text analysis in regard to morpho-syntactic structures, which goes beyond more traditional approaches to focus on thesauri of harmful vocabulary.

[1] See StyloMetrix repository on Github: https://github.com/ZILiAT-NASK/StyloMetrix

In this paper, we aim to elaborate on the results of the stylometric analysis of hate speech text samples from human-based content moderation process on a large social news service. To put our linguistic computation into proper context with all its complexities, we will first shed some light on the challenges that one encounters in efforts to investigate the notion of offensive language and hate speech on the web for automated detection.

2. Datasets and annotation framework for automated harmful content detection

Before one is able to train the model and learn from its computation in human-computer interaction, annotated datasets are required, which already poses a fundamental challenge. Aside from the debate on the quality of the data, for the English language, a number of benchmark datasets consisting of posts scraped from popular social network services (SNS), such as Twitter, Reddit, Facebook etc., are available online. As to the lower-resourced languages, public datasets are extremely scant. In 2019, Pol-Eval, a Sem-Eval-inspired campaign to evaluate NLP tools for the Polish language, a task to automate the detection of cyberbullying has been designed and a dataset consisting of 11 041 tweets from Polish Twitter has been made publicly available (Ptaszynski et al., 2019; Rybak et al., 2020) as part of the KLEJ benchmark (*Kompleksowa Lista Ewaluacji Językowych*)[2]. However, even with available datasets, challenges still arise. One is the human annotation of the dataset, which involves labeling the data either for binary classification (e.g., offensive vs. neutral) or for multiclass classification (e.g., offensive vs. hate speech vs. neither). This requires conceptualizing criteria and methods for the annotation framework. It is worth noting that recent studies have raised concerns about low inter-annotator agreement, indicating the presence of annotators' biases and the need for agreed-upon definitions (Ross et al., 2016; Banko et al., 2020). This suggests that the reliability of existing datasets may be questionable due to the subjective nature of the annotations. These challenges surrounding dataset annotation and reliability subsequently impact the generalizability and applicability of models. When a model pre-trained on a particular dataset is utilized for prediction tasks using different data, studies have reported a significant decrease in F1 scores, which is a commonly used metric for evaluating the accuracy of a model's

[2] Only recently, a subsequent dataset of harmful Twitter content has been released by deepsense.ai, the creators of TrelBERT model dedicated to cyberbullying detection (CBD). It is a limited dataset containing 1000 samples, out of which 10.6% was deemed harmful in the annotation process (Szmyd et al., 2023: 19–20).

performance. This highlights the need to carefully consider the impact of dataset biases and annotation methods on the model's ability to generalize to new data.

2.1. Cyberbullying and hate speech criteria

Before commencing any annotation endeavour, it is imperative to establish definitive stipulations and benchmarks for identifying instances of offensiveness, cyberbullying, and hate speech. These terms have become the focal point of machine learning tasks in recent years. Drawing upon insights from sociological and social psychological studies, hate speech typically encompasses expressions targeting distinct group attributes, including but not limited to ethnicity, complexion, religion, gender, and sexual orientation[3] (Warner and Hirschberg, 2012). Conversely, other forms of offensive discourse are perceived to predominantly target individuals. This perspective resonates with numerous inquiries into the automated detection of hate speech. Additionally, hate speech is further defined as „language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Davidson et al., 2017: 512), or as „language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics" (Fortuna et al., 2018: 85:5). Despite the nuances inherent in these definitions, there is a general consensus within the natural language processing (NLP) community that hate speech necessitates a purposeful assault aimed at a collective, predicated upon either stereotypical perceptions or factual traits pertinent to the group's identity (de Gibert et al., 2018: 11). Such an interpretation aligns with the demarcations outlined in the guidelines of prominent online platforms such as Facebook, YouTube, and Twitter (MacAvaney et al., 2019; Fortuna et al., 2018).

Notwithstanding the concise and relatively consistent definition, challenges arise when applying it to assess genuine text samples. Initially, the characteristics of safeguarded groups evolve over time, leading to the emergence of new targets for verbal denigration, which in turn necessitates revisiting the definitions of protected attributes. Furthermore, derogatory terms can be used outside of hate-filled content depending on context, while disparaging messages might be conveyed through ambiguous or metaphorical phrasing. Additionally, humor and sarcasm can often soften the impact of hateful expressions, rendering their evaluation more subjective. Lastly, this also involves an element of negative stereotyping that might be socially accepted, thereby blurring the line between innocuous generalizations and hate speech (Paz et al., 2020).

[3] It is noteworthy to observe that the essence of this definition is closely linked with the legal culture of North America, having initially been articulated by John T. Nockleby in *Encyclopedia of the American Constitution* in 2000 (cf. de Gibert et al., 2018: 11).

Furthermore, definitions referring to diverse target groups „exhibit inaccuracies when juxtaposed with each other, not only in the cross-linguistic but the intra-lingual perspective as well" (Adamczak-Krysztofowicz et al., 2016: 13).

Consequently, this engenders inevitable disparities and imprecisions within the annotation process, whether undertaken through crowdsourcing or expert curation. Recent observations have also frequently underscored that the inherent social biases of human annotators exert an influence on automated classifiers (Mostafazadeh Davani et al., 2023).

2.2. Automated detection experiments and data annotation

There are multiple approaches to the annotation of large datasets, including online crowdsourcing, designing a balanced group of lay annotators (based e.g. on criteria of age or gender) to reduce the possible bias, or expert-curated methods. In some studies, it is claimed that, in general, amateur annotators are more likely to overuse such labels, as hate speech, than professionally trained ones (Waseem, 2016).

In our research, we conducted a series of machine learning experiments based on two datasets: 1) the afore-mentioned, publicly available KLEJ benchmark dataset for cyberbullying detection (CBD), 2) Wykop.pl dataset of real-life comments banned by the moderators, which has been shared with us within a joint project (funded by NCBiR). We mainly focused on fine-tuning the model to learn how to predict the right label (either harmful or non-harmful) for each text sample. For the Wykop.pl dataset, the best-performing model proved to be large RoBERTa fine-tuned with StyloMetrix embeddings, which achieved 94% of accuracy for the binary classification. However, applying the same methodology to KLEJ CBD dataset did not result in satisfying results, which calls for taking annotation bias into consideration while generalizing to different datasets. In the current paper, however, we do not aim to elaborate on the model performance, but instead we intend to focus solely on stylometric analysis of harmful content to shed some sociolinguistic perspective on social media/social news posts and comments. While these contribute to one of the most fundamental topics for tech industry and machine learning, there are very few studies on their language specificity.

2.2.1. Wykop.pl banned content dataset

As opposed to the KLEJ dataset[4], content banned from Wykop.pl that we used in our experiments is part of authentic data reported by the users and subsequently deemed harmful by the moderators in accordance with the website's policy. Whereas Wykop.pl as a social news service is known for praising freedom of expression and encouraging users to find a safe space to discuss news in an informal environment, which often involves humour, sarcasm and contributes to satire (Sowiński, 2018), users are invited to report any content regarded as non-compliant with the general rules (hate, violence, illegal content, adult content, SPAM, advertisement, etc.). When picking a reason to report certain content in terms of its offensiveness, one can choose between multiple labels as "it attacks me", "it attacks others", "promotion of hatred and violence, drastic content" etc.[5], resulting from the internal moderation policy of Wykop.pl. In our dataset of harmful Wykop.pl content, called BAN-PL (Okulska et al., 2023b)[6], we included samples referring to the labels concerning the notions of cyberbullying and hate speech. For the stylometric analysis purposes, we analyzed a randomly sampled subset of 1585 entries and comments.

3. Stylometric analysis

3.1. StyloMetrix methodology

While large language models do not provide much room for human interpretation of the performance of a particular metric, our tool supporting BERT-based models with stylistic, morpho-syntactic metrics offers a different perspective to elaborate on linguistic information. The basic StyloMetrix model consists of 118 expert-designed metrics included in 9 groups (Descriptive, Grammatical forms, Graphical, Inflection, Lexical, Psycholinguistic, Punctuation,

[4] The KLEJ CBD dataset consists of Polish tweets from popular Twitter accounts (11 041 text samples). Each sample was first annotated by 2-3 lay persons and finally evaluated by an expert curator. The annotation covered 3 classes (0: non-harmful, 1: cyberbullying, 2: hate speech and other harmful contents), while for the binary classification task the latter two were merged into one to discriminate harmful against non-harmful content. The annotators were provided with general guidelines to identify offensive content, including disclosure of personal data, threats, personal attacks, ridiculing, the accumulation of profanity etc. (Ptaszynski et al., 2019: 94). The state-of-the-art (SOTA) model for the automated detection task proved to be TrelBERT, a BERT-based language model trained on Polish Twitter data, released by deepsense.ai (Szmyd et al., 2023).

[5] See the full list of ban reasons within Wykop.pl moderation standards: https://wykop.pl/standardy-moderacji

[6] The initial publicly available iteration of BAN-PL, a Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl Web Service, encompasses 24,000 samples of anonymized content, partitioned into 12,000 pieces for the "harmful" and 12,000 for the "neutral" (non-harmful) class. See the repository on Github: https://github.com/ZILiAT-NASK/BAN-PL. In future, the entire dataset consisting of 691 662 pieces of content will be made available (Okulska et al., 2023b).

Syntactic, Word formation), while additional custom metrics can be added to the existing ones at any time. Regardless of the length of the text sample, the tool computes vectorization for each metric which is normalized in regard to the number of tokens, namely assigned with a value between 0 and 1. Therefore, even samples of varying lengths can be reliably compared with each other. Computational analysis with StyloMetrix provides human linguists with vast statistical information on grammatical patterns according to the conviction that individual style does not rely on lexical choices only, but is also reflected in grammar.

3.2. Harmful speech – linguistic features

An in-depth analysis of both datasets can lead to several conclusions, which are assessed both by expert human knowledge, as well as computed metrics. While KLEJ consists of tweets associated with popular Twitter accounts, the text samples are more likely to comment on current political events and therefore consist of more proper names or indirect allusions to given public figures. Offensive text samples collected from Wykop.pl moderators do not follow current news that closely, they tend to include more general strong opinions on different political and social matters. Therefore, the mean value for L_PERSN (Lexical: Personal noun), as well as for L_PLACEN (Lexical: Place name) is considerably higher for KLEJ dataset (1.94%, 1.10% respectively)[7], as opposed to Wykop.pl content (1.10% and 0.84%). More importantly, the wider range of topics covered by the Wykop.pl dataset, as compared to KLEJ, is reflected in the crucial metric referring to diversity of vocabulary, namely the type-token ratio (TTR), which is calculated by dividing the number of different words (types or types of lemmas) in the sample by the total number of words (tokens). TTR for Wykop.pl banned content is therefore 11.6 percentage points higher (91.67%), as compared to KLEJ. Given the differences in the very nature of content derived from these two datasets, online availability of KLEJ benchmark, as well as the well-understood limitations of the current paper, we will therefore focus on linguistic examples taken from the Wykop.pl dataset. Out of a dataset of 1580 items banned by the moderators due to offensiveness, we manually relabeled 70 samples as examples of generalized hate speech targeted at specific victim groups (Black people, Jews, Muslims, Indians, LGBT, women), which can be deemed to unquestionably fulfill the definition of hate speech. We computed the stylometric analysis for this very limited subset separately and decided to display

[7] These figures refer to text samples labeled as harmful in the KLEJ benchmark dataset. For the neutral class of tweets, the mean percentage of tokens concerned with personal names and places is even higher (2,27%).

these examples in the paper due to the notoriety and detrimental social impact of the online phenomenon, which has still been understudied in the Polish context.

3.2.1. The use of nouns

The frequency of nouns in the Wykop.pl dataset is on average 12 percentage points higher than that of verbs, with a mean value of 32.04%. This indicates that nearly one-third of tokens in the corpus constitute nouns. Notably, the frequency of potentially offensive nouns in the vocative case is conspicuously high, standing at 2.43%, which points to a high frequency of appellative forms of address. However, for the subset of generalized hate speech, the value of nouns in the vocative case is ten times lower which can indicate that the nature of hateful statements does not require the incidence of personal attacks. Also, it is worth emphasizing the distribution of nouns across different cases. Specifically, nouns in the nominative case account for more than a third of all noun occurrences, with the accusative and genitive cases being the second and third most frequent, respectively. For hate speech specifically, the use of the genitive case slightly surpasses that of the accusative case.

Table 1. Examples of nouns in nominative, accusative and genitive cases[8]

Case	Example in Polish	English translation
Nom.	(1) "To te jeb*** **czarnuchy** przesadzają od zawsze **Niewolnictwo** powinno zostać przywrócone i **kara** śmierci dla takich śmieci jak ty psie"	(1) "It's those f***ing **n*** who have been exaggerating forever **Slavery** should be reinstated and the **death penalty** for trash like you dog"
Gen.	(2) "nie ma czegoś takiego jak Palestyna jest plemię **barbarzyńców** modlących się do skały i **proroka pedofila**" (3) "niech USA im wpier*** to może reszta tych inżynierów **piasku** się uspokoi"	(2) "there is no such thing as Palestine there is a tribe **of barbarians** praying to a rock and a **paedophile prophet"**. (3) "let the US f*** them up then maybe the rest of these **sand** engineers will calm down".
Acc.	(4) "już niedługo zamkniemy **polki** w obozach tylko dojdziemy do władzy" (5) "Kiedy znów będzie można linczować **bambusów**"	(4) "we will soon be locking **Polish women** in camps as soon as we come to power" (5) "When will it be possible to lynch **bamboos** [Polish racial slur for Black people] again".

[8] This section of the paper discusses examples of hate speech. The authors do not support the use of harmful language, nor any of the harmful representations quoted.

In hate speech, nouns in the nominative case often tend to support declarative statements based on some sort of false pretences (Adamczak-Krysztofowicz, Szczepaniak-Kozak, 2017: 295-296) appealing to the projection of common sense, as in the first of the above examples ("n*** have been exaggerating forever", "slavery should be reinstated"). In sample 2, one can discern the use of the nominative case in a catchy, rhetorically sharp statement consisting of contrasting assertions "there is no such thing as [noun in Nom.], there is [another noun in Nom.]". First assertion concerned with some neutral content ("Palestine" as a toponym) is being negated by the second one using pejorative religionyms[9] ("praying to a rock and a paedophile prophet" and primitivisms ("a tribe of barbarians") instead. Apart from the question of rection of a verb requiring the use of the genitive case (such as "modlić się do", "to pray to"), the most fundamental syntactic role of nouns in the genitive case is noun adjunct, which is also clearly reflected in the above example ("tribe of barbarians", "to a rock and a paeodophile prophet"). The use of a modifier in the form of a noun adjunct is also clearly visible in example 4, in which people of Arabian origin are called "sand engineers", a patronizing term referring to the Middle Eastern allegedly poor economy. The use of the accusative case is mostly used to address the objects of a certain proclaimed action ("to lock Polish women", "to lynch negroes").

3.2.1. The use of verbs

The frequency of verbs in the Wykop.pl dataset is 20,11% on average, with the present tense being the dominating one (7,62% of all tokens). Past tense and future tense occurrences are significantly lower (2,73% and 1,09% respectively). To a vast extent, the use of present tense in offensive language is driven by harmful predicate nominatives with second person singular verbs being the top category for conjugated verbs (5.74% of all tokens). These values contribute to a wide range of slurs directed at individuals, such as "you are a moron". In examples of generalized hate speech, the use of verbs in tenses seems to be more balanced and fulfils certain rhetorical goals. Second person singular verbs are nearly ten times less frequent, whereas the priority is given to first and third person plural verbs. This may be understood as syntactic polarization (Guillén-Nieto, 2023: 79–82).

[9] In the sociolinguistic analysis of hate speech, the term "religionym" is employed to denote the concept of "identification by means of the assumed religious denomination the foreigner belongs to" (Adamczak-Krysztofowicz, Szczepaniak-Kozak 2017: 298).

Table 2. Examples of verbs in the present, past and future tenses

Tense	Example in Polish	English translation
Present	(1) "Zabawne **jest** twoje lewactwo i w sumie głupota bo w imię poprawności politycznej **nazywasz** mnie tłukiem chociaż w głębi bardzo dobrze **zdajesz** sobie sprawę że większość egzotycznych przybyszy **to** nie są kardiochirurdzy czy choćby wykwalifikowani pracownicy (…)" (2) "Kobiet **się nie kocha**. Kobiety **się bije**"	(1) "What's funny **is** your extreme leftist politics and altogether stupidity because in the name of political correctness you **call** me a bum even though deep down you **are** very well aware that most of the exotic newcomers **are** not cardiac surgeons or even skilled workers" (2) "Women **are** not to be loved. Women **are** to be beaten"
Past	(3) "Egipcjanie wbrew wizerunkowi stworzonemu przez filym z Hollywood **byli** czarni Piramidy się same **nie wybudowały** Sfinks pismo obrazkowe i tak dalej to robota czarnych To że całościowo czarni ssą po całości bo z tej części która **była** wysokorozwinięta **zostały** ruiny nie znaczy że czarni nigdy niczego nie **osiągnęli** To tak jak brudasy kiedyś ostoja nowoczesnej nauki i medycyny obecnie ruchacze owiec i terroryści"	(3) "The Egyptians, contrary to the image created by the Hollywood filmmakers, **were** black The pyramids **did not build** themselves The Sphinx the pictorial writing and so on are the work of blacks Just because overall blacks suck because the part that **was** highly developed is left in ruins does not mean that blacks never **achieved** anything It's like the dirty people once the stronghold of modern science and medicine now sheep f***ers and terrorists"
Future	(4) "No ładnie poleciały bany za obrażanie ciemnoskórych protestujących Co **będzie** następne karanie za obrażanie białych? Czy może rasizm działa tylko w jedną stronę (…)" (5) "Mam nadzieję że są w USA jakieś białe organizacje które **uratują** honor i **każą** czarnym np pełzać po ziemi xD"	(4) "Well nice bans flew for insulting dark-skinned protesters What **will be** next punishment for insulting whites?" (5) "I hope there are some white organizations in the U.S. that **will save** the honor and **make** blacks, for example, crawl on the ground xD"

It has already been argued that in terms of both generalized and directed hate speech there is more focus on the present, as opposed to general social media comments (ElSherief et al., 2018). The use of present tense, apart from phatic expressions, often combined with offensiveness, emphasizes not only individual current feelings, but most certainly shapes the persuasive grammar of alleged common sense assertions that decrease the perceived level of subjectivity. Studies have also shown that using present (vs. past) tense contributes to a higher degree of persuasion in terms of consumer behavior (Packard et al., 2023). With respect to hateful opinions on social and political subjects, the present tense deprived of any subjective modifiers, modal phrases or conditional structures can also be seen as more persuasive. Along with the aforementioned significant degree of nouns in the nominative case, it tends to exhibit the highest

clout by stating the allegedly obvious, as in the case of such sentences as "there is no such thing as Palestine, there is a tribe of barbarians", "It's those f***ing n*** who have been exaggerating forever", or "most of the exotic newcomers are not cardiac surgeons or even skilled workers".

In alternative instances, users often employ the past tense to underpin their persuasive discourse by juxtaposing divergent statements, employing a pseudo-historical rationale. In example 3, the past tense is leveraged to portray Egyptian culture as an integral facet of Black people's heritage, reflecting a favorable sentiment only to establish a contrasting statement on the present ("overall blacks suck"). While the past tense is instrumental in introducing a more elaborate depiction that conveys a certain scholarly depth underpinning the user's argumentation, the final sentence demonstrates that Polish users can achieve heightened persuasiveness in concisely crafting the contrast between the "now" and the "then" by employing ellipsis to omit verbs entirely ("once the stronghold of modern science and medicine[,] now sheep f***ers and terrorists"). This approach is reinforced by employing nouns in the nominative case, thereby enhancing the overall impact of the persuasive message.

As statistics show, future tense frequency is over two times lower than that of past tense. Examples 4 and 5 display the tendency to use past tense as a consequence of critical attitude towards the present, which goes in two separate directions. On the one hand, it can build assumptions as to what will happen in the future, on the other – it is used to express hopeful postulates as to what should happen in the future to restore order.

As much as agitating for future actions plays a crucial role in hate speech, the relatively low frequency of future tense results from prioritizing different grammar forms to convey a similar message. Imperative forms with infinite usage is one way to achieve this goal, as following examples show: 1) "Strzelać do tego bydła bez ostrzeżenia" ("Shoot these cattle without warning"), 2) "profilaktycznie rozstrzelać te manifestantki co by się poczuły jak w ameryce" ("preventively execute these demonstrating women so that they feel like in america"). Overall, the value for infinite usage is relatively high (3,30%), comprising future forms, modal phrases, and quasi-verb complements (defective words that function partially as verbs with non-verbal morphology). For the hate speech subset, it is even more significant with the value of 5,21%. As compared with the infinite usage, imperative form frequency in cyberbullying language is at a similar level of 3,40%, covering mostly offensive and toxic forms in the second person singular, however, in case of hate speech, as expected, this proves to be nearly ten times lower. In generalized hateful utterances there is some room for permissive imperative of a composite form, as in one of the formerly mentioned examples ("niech USA im wpier***", "let the US f*** them up").

To convey hateful messages in a more subjective form, conditional mood is used, however, statistically speaking, it does not play a vital role in the analyzed offensive dataset (only 0.30% of all tokens for cyberbullying, and 0,55% for hate speech). In explicit hate speech, it can serve the purpose of mitigating the incitement for violence, as in the following examples: (1) "przydałby się wujek Adolf aby posprzątać żydostwo i czarnuchów" ("uncle Adolf would come in handy to clean up the jewry and n***"), (2) "ja bym otworzył Auschwitz" ("I would open Auschwitz"). Conditional mood also fits the rhetoric device to imaginatively describe what "they", meaning "others", would allegedly be capable of: "oni by bez żadnych skrupułów zgwałcili ci kobietę na twoich oczach zabili ją a później zabili ciebie" ("they would unscrupulously rape your woman in front of your eyes kill her and then kill you").

3.2.2. The use of adjectives, pronouns, modifiers, and nominal phrases

According to the data analysis, pronouns were found to be the third most frequently utilized part of speech in offensive language, exhibiting a mean frequency of 15.97% (with only 1.5 pp lower frequency for hate speech). This exceeds the incidence of adjectives by more than two times, which were used with a mean frequency of 6.52%. However, within the subset of hate speech, adjectives were relatively more prevalent, accounting for 9.57% of the overall usage. Furthermore, the data highlights the substantial usage of demonstrative pronouns, both in cyberbullying and hate speech instances. These pronouns primarily serve as adjective noun modifiers, such as in phrases like "these cattle" or "those n***." Remarkably, the employment of demonstrative pronouns in this context constituted 5.21% of the overall language usage, which seems high, as compared to all adjectives.

A distinguishing linguistic feature of hate speech is a significant disparity between the usage of first person plural and third person plural pronouns, with the latter being employed more frequently (ElSherief et al., 2016). This phenomenon is consistent with the logic of creating an "us vs. them" dichotomy, commonly found in hate speech and supporting "othering" understood as a discursive practice[10]. The Wykop.pl dataset reveals that pronouns referring to "them" occur almost five times more often than those referring to "us" (0.44% of all tokens vs. 0.09%). Additionally, this discrepancy is similarly apparent in verb conjugation, where the frequency of third person plural verbs is nearly ten times greater than that of first person plural verbs (1.30%

[10] The notion of "othering" has been investigated primarily in the field of critical discourse analysis. Within this area of research, "othering is a technical term used here to describe the manner in which social group dichotomies are represented via language" (Pandey, 2004: 155) with the most fundamental concepts having been elaborated on by Theun van Dijk in the 80's and early 90's.

vs. 0.14% of all tokens). Within the hate speech subset, "us" pronouns vs. "them" pronouns occur with a mean frequency of 0,18%, as opposed to 1,1%.

Table 3. Examples of 3rd person plural pronouns

Pronoun	Example in Polish	English translation
3PL	(1) "Jestem ateuszem i antyklerykałem ale za takie postępki sam własnoręcznie pałowałbym tę tęczową zarazę. Czy **oni** naprawdę muszą wywracać do góry nogami naturalny porządek"	(1) "I'm an atheist and anti-cleric, but I would bludgeon this rainbow plague myself for such advances. Do they really have to turn the natural order upside down"
	(2) "Nie może być tak że biedne Murzyniątka wypiją po browarze i świat **im** się kończy"	(2) "It can't be that poor Negroes drink a brew and the world ends for **them**"
	(3) "Już dawno powinno się zatapiać te **ich** tratwy zamiast **ich** wyławiać"	(3) "It's long overdue to sink those rafts of **theirs** instead of fishing **them** out"

As previously mentioned, the dataset contains a significant proportion of nouns, with over one-third of all tokens classified as such (32.04%). What is more, a substantial amount of the text samples consist of nominal phrases, which are formed by the combination of adjectives, pronouns, adverbs and past participles used as noun modifiers. In fact, these nominal phrases account for more than half of all text samples, comprising 54.83% of the Wykop.pl dataset and 57.35% for hate speech specifically. Below some examples of wordy nominal phrases are presented.

Table 4. Examples of complex nominal phrases

Example in Polish	English translation
(1) "Są to **negatywnie nastawieni do białego człowieka młodzi mężczyźni** którzy z przyjemnością szerzą tutaj swoją ideologię"	(1) "These are **young men, having a negative attitude towards white men**, who take pleasure in spreading their ideology here"
(2) "Zaczynam widzieć sens w tym ze murzyn powinien być **niewolnikiem skutym łańcuchami i zaciągnięty do najcięższych najbrudniejszych prac"**	(2) "I'm beginning to see the sense in the idea that the black man should be **a slave shackled with chains and enlisted in the hardest dirtiest jobs**"

Conclusion

Both studies on automated hate speech or harmful offensive content detection and sociolinguistic examinations of the devastating phenomenon have been of utmost importance for the reduction of violence and aggression in online societies. The fundamental challenge, however, is the data scarcity. Large, rich and well-balanced corpora from different sources provided with a reliable annotation framework are rather scant, not only for lower-resourced languages. While there are some highly valuable Polish studies investigating the problem of hate speech, they rely on available sources, such as press (e.g. Adamczak-Krysztofowicz et al., 2017), while the constant increase of user-generated online content requires further examinations despite the challenges of data collection. A deep insight into actual banned text samples from a popular web service, which due to the moderation process have been removed and, thus, are no longer available online, can shed some new perspectives on the sociolinguistic studies of hate speech on the web. However, it also imperative to stress the crucial need to develop further rich corpora from different domains.

In our stylometric analysis, our central focus has been a comprehensive exploration of the morpho-syntactic attributes inherent to offensive language, with particular attention directed towards hate speech. The underlying goal of this endeavor has been to illuminate salient grammatical patterns essential for the advancement of automated detection models and the facilitation of linguistic inquiries. Our findings underscore the potential for achieving persuasive impact through the strategic omission of subjectivity, the deliberate avoidance of 1st person singular and plural forms, the adoption of the present tense in lieu of conditional or imperative forms, and the amplification of discourse via the juxtaposition of past and present tenses. Furthermore, our observations pertaining to the utilization of nouns, verbs, and pronouns offer valuable insights, forming a foundational basis for subsequent analyses. It is worth noting that these preliminary insights were derived from a relatively modest corpus and should be validated and extended using more expansive datasets in the future. Moving forward, a rigorous examination of the diversity inherent to modifiers and nominal phrases warrants consideration, as these elements contribute to expressions that can effectively sidestep overt ethnonyms or racial slurs.

While lexical approaches are effective in examining notable shifts in the prevalent vocabulary of derogatory language over time, particularly in relation to the public sphere, it is essential to consider that as automated content detection on the web increases, malicious users may employ strategies to convey hateful messages without explicit profanities or widely

recognized slurs, while adhering to well-established grammatical structures of hate. Within the subset of hate speech, our analysis revealed that 21% of the samples contained explicit profanities, while 70% consisted of various derogatory terms. This indicates that 30% of the samples could potentially be overlooked if only vocabulary-based approaches were employed. Consequently, it can be argued that both lexical and morpho-syntactic approaches are highly valuable in studying the phenomenon of hate speech and curbing its dissemination online, albeit they can be applied to slightly different research inquiries.

Acknowledgements

The project KOMTUR ('System for categorizing, evaluating, and moderating online content for new services and advertising selection') was funded by the National Centre for Research and Development in Poland within the 3/1.1.1/2020 Operational Programme Smart Growth 2014-2020, co-financed with the European Regional Development Fund. We wish to very much thank Wykop.pl web service for sharing with us the data and giving us some insight into the real-life moderation process.

References

Adamczak-Krysztofowicz, Sylwia, Anna Szczepaniak-Kozak (2017) "A Disturbing View of Intercultural Communication: Findings of a Study into Hate Speech in Polish." *Linguistica Silesiana*, 38; 285-310. https://doi.org/10.24425/linsi.2017.117055.

Adamczak-Krysztofowicz, Sylwia, Anna Szczepaniak-Kozak, Magdalena Jaszczyk (2016) "Hate Speech: an Attempt to Disperse Terminological Ambiguities." *Voci* 13; 13–28.

Banko, Michele, Brendon MacKeen, Laurie Ray (2020) "A Unified Taxonomy of Harmful Content." [In:] Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, Zeerak Waseem (eds.) *Proceedings of the Fourth Workshop on Online Abuse and Harms*; 125–137. Retrieved from https://aclanthology.org/volumes/2020.alw-1/. Date: 11.02.2024.

Davidson, Thomas, Dana Warmsley, Michael Macy, Ingmar Weber (2017) "Automated Hate Speech Detection and the Problem of Offensive Language". [In:] Yu-Ru Lin, Meeyoung Cha, Daniele Quercia (eds.) *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017).* Palo Alto, California: AAAI Press; 512–515.

De Gibert, Ona, Naiara Perez, Aitor García-Pablos, Montse Cuadros (2018) "Hate Speech Dataset from a White Supremacy Forum". [In:] Darja Fišer, Ruihong Huang, Vinodkumar

Prabhakaran, Rob Voigt, Zeerak Waseem, Jacqueline Wernimont (eds.) *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*; 11–20. Retrieved from https://aclanthology.org/volumes/W18-51/ Date: 11.02.2024.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang, Elizabeth Belding (2018) "Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media". [In:] *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12; 42-51. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15041/14891 Date: 11.02.2024. https://doi.org/10.1609/icwsm.v12i1.15041

Fortuna, Paula, Sérgio Nunes (2018) "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys*, 51(4); 85:1 – 85:30. https://doi.org/10.1145/3232676.

Guillén-Nieto, Victoria (2023) *Hate Speech: Linguistic Perspectives*. Berlin, Boston: De Gruyter Mouton.

Jaszczyk-Grzyb, Magdalena (2021) *Mowa nienawiści ze względu na przynależność etniczną i narodową w komunikacji internetowej. Analiza porównawcza języka polskiego i niemieckiego*. Poznań: Wydawnictwo Naukowe UAM.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder (2019) "Hate Speech Detection: Challenges and Solutions." *PloS one,* 14(8); e0221152. https://doi.org/10.1371/journal.pone.0221152.

Mostafazadeh Davani, Aida, Mohammad Atari, Brendan Kennedy, Morteza Dehghani (2023) "Hate Speech Classifiers Learn Normative Social Stereotypes." *Transactions of the Association for Computational Linguistics,* 11; 300–319. https://doi.org/10.1162/tacl_a_00550

Okulska, Inez, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska & Adam Nowakowski (2023a) "StyloMetrix: An Open-Source Multilingual Tool for Representing Stylometric Vectors". arXiv preprint arXiv:2309.12810.

Okulska, Inez, Kinga Głąbińska, Anna Kołos, Agnieszka Karlińska, Emilia Wiśnios, Adam Nowakowski, Paweł Ellerik, Andrzej Prałat (2023b) "BAN-PL: a Novel Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl web service", arXiv:2308.10592v2 [cs.CL]. https://doi.org/10.48550/arXiv.2308.10592

Packard, Grant, Jonah Berger, Reihane Boghrati (2023) "How Verb Tense Shapes Persuasion." *Journal of Consumer Research,* 50(3); 645-660. https://doi.org/10.1093/jcr/ucad006

Pandey, Anjali (2004) "Constructing Otherness: A Linguistic Analysis of the Politics of Representation and Exclusion in Freshman Writing." *Issues in Applied Linguistics,* 14(2); 153–184. https://doi.org/10.5070/l4142005075

Paz, Maria Antonia, Julio Montero-Díaz, Alicia Moreno-Delgado (2020) "Hate Speech: A Systematized Review." *Sage Open,* 10(4). https://doi.org/10.1177/2158244020973022.

Ptaszynski, Michal, Agata Pieciukiewicz, Paweł Dybała (2019) "Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter." [In:] Maciej Ogrodniczuk, Łukasz Kobyliński (eds.) *Proceedings of the PolEval 2019 Workshop*. Warszawa: IPI PAN; 90–110.

Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki (2016) "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis". *Bochumer Linguistische Arbeitsberichte* (*NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*), 17: 6–9.

Rybak, Piotr, Robert Mroczkowski, Janusz Tracz, Ireneusz Gawlik (2020) „KLEJ: Comprehensive Benchmark for Polish Language Understanding." [In:] Dan Jurafsky, Joyce Chai, Natalie Schluter, Joel Tetreault (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 1191–1201. Retrieved from https://aclanthology.org/2020.acl-main.111/ Date: 11.02.2024 https://doi.org/10.18653/v1/2020.acl-main.111.

Sowiński, Rafał (2018) „Rola systemu tagów w serwisie Wykop.pl. Folksonomia czy memy?" *Zeszyty Naukowe Państwowej Wyższej Szkoły Zawodowej im. Witelona w Legnicy*, 3(28); 201–212.

Szmyd, Wojciech, Alicja Kotyla, Michał Zobniów Piotr Falkiewicz, Jakub Bartczuk, Artur Zygadło (2023) "TrelBERT: A Pre-trained Encoder for Polish Twitter." [In:] Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, Roman Yangarber (eds.) *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*; 17–24. Retrieved from https://aclanthology.org/2023.bsnlp-1.3/ Date: 11.02.2024. https://doi.org/10.18653/v1/2023.bsnlp-1.3

Waseem Zeerak (2016) "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter" [In:] David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, Svitlana Volkova (eds.) *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics; 138–142.

Warner, William, Julia Hirschberg (2012) "Detecting Hate Speech on the World Wide Web" [In:] Sara Owsley Sood, Meenakshi Nagarajan, Michael Gamon (eds.) *Proceedings of the*

2012 Workshop on Language in Social Media (LSM 2012); 19–26. Retrieved from https://aclanthology.org/W12-2103/ Date: 11.02.2024.