Anna Bugajska
http://orcid.org/0000-0001-6078-7405
Jesuit University Ignatianum in Kraków
anna.bugajska@ignatianum.edu.pl

Paulina Dziedzic
https://orcid.org/0009-0006-7162-3207
Tischner European University in Kraków
dziedzicp01@gmail.com

DOI: 10.35765/pk.2023.4203.32

A Linguistic Analysis of Sexism-Related Hate Speech in Social Media

ABSTRACT

The aim of this article is to present the functioning of a dual algorithm/human analysis and to investigate the means with which to study hate speech, especially sexism-related hate speech, in the online environment, focusing on social media comments and hashtags. Another aim is to investigate new linguistic trends in contemporary online hate speech that can be revealed via quantitative hate speech analysis. In the first part, the concept of hate speech is briefly introduced in a linguistic context. In the second part, an example of a Twitter hashtag is analyzed. In the third part, an algorithm for the identification of sexism-related hate speech from the corpus available at hatespeechdata.com is discussed. The article demonstrates the methods of evaluating selected types of online content for the presence of hate speech. It is made evident that algorithm-based hate speech qualification is an insufficient tool for identifying hate speech and that qualitative analysis by a trained linguist is necessary.

KEYWORDS: hate speech, algorithm, social media, language, sexism

STRESZCZENIE

Lingwistyczna analiza mowy nienawiści związanej z seksizmem w mediach społecznościowych

Artykuł ma na celu przedstawienie funkcjonowania analizy dualnej algorytm-człowiek oraz sposobów badania w szczególności mowy nienawiści związanej z seksizmem w środowisku internetowym, z naciskiem na komentarze i hashtagi w mediach społecznościowych, oraz zbadanie nowych trendów językowych we współczesnej mowie nienawiści w Internecie, które można

Suggested citation: Bugajska, A. & Dziedzic, P. (2023). A Linguistic Analysis of Sexism-Related Hate Speech in the Social Media. 🅭 🅯 *Perspectives on Culture*, 3(42), pp. 549–560. DOI: 10.35765/pk.2023.4203.32

Submitted: 28.02.2023 Accepted: 31.08.2023

ujawnić za pomocą ilościowej analizy mowy nienawiści. W pierwszej części pokrótce wprowadzono pojęcie mowy nienawiści, odnosząc się do kontekstu językowego. W drugiej części przeanalizowano przykładowy hashtag Twittera. W trzeciej części wykorzystano algorytm identyfikacji mowy nienawiści na tle seksizmu z korpusu dostępnego na stronie hatespeechdata.com. W artykule przedstawiono metody oceny wybranych typów treści internetowych pod kątem obecności mowy nienawiści. Zostaje dowiedzione, że algorytmiczna kwalifikacja mowy nienawiści jest niewystarczającym narzędziem w identyfikacji mowy nienawiści i konieczna jest analiza jakościowa przeszkolonego językoznawcy.

SŁOWA KLUCZE: mowa nienawiści, algorytm, media społecznościowe, język, seksizm

Introduction

Today, a range of phenomena related to linguistic crime and abuse is garnering interest in the studies of digital communication, due to the production of a large number of comments within social networks (Assimakopoulos et al., 2017). Although some policies have been instituted by the administrators of social media outlets, there is an observable lack of consistent legislation and tools to identify and manage undesirable content. For the most part, the online community is believed to be self-regulatory, i.e., ideally the users themselves should be able to create a space free from hate and other abuse. However, to facilitate the process of identifying and eradicating such content, it is common to use algorithms that detect it using predetermined characteristics. The aim of this article is to present the functioning of such a dual algorithm/human analysis using examples and to investigate the means with which to study hate speech, especially sexism-related hate speech, in the online environment, focusing on social media comments and hashtags. Furthermore, the article is intended to investigate new linguistic trends in contemporary online hate speech that can be revealed via quantitative hate speech analysis. In the first part, the concept of hate speech is briefly introduced in a linguistic context. In the second part, an example of a Twitter hashtag is analyzed. In the third part, an algorithm is used for the identification of sexism-related hate speech from the corpus available at hatespeechdata.com. Thus, the article demonstrates the methods of evaluating selected types of online content for the presence of hate speech. It is made evident that algorithms for determining hate speech are an insufficient tool and that qualitative analysis by a trained linguist is necessary.

Hate speech and its linguistic analysis

The concept of online hate speech is relatively new, having appeared in the virtual realm a few years ago. In the international terminology, there is no single legally binding definition of the term *hate speech*. What stands out, however, is its basic feature, which may be described as the use of pejorative, offensive, and/or vulgar terms in relation to a factor that identifies a person or group of people. Such factors may be related to religion, ethnicity, race, nationality, gender, or many other distinguishing features of an individual/group. Also, there are a number of documents which attempt to provide a legal framework within which one can consider the notion of hate speech (for an exhaustive discussion, see e.g. Assimakopoulos et al., 2017; Carlson, 2021; Pejchal, 2020). For instance, in the United States, it has been identified by the Supreme Court as contradicting the basic constitutional right to freedom of speech (UNSPAHS, 2019), while in the European Union we can invoke the *Framework decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*. According to the latter, the following behaviors are punishable:

- public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, color, descent, religion, or belief or national or ethnic origin;
- the above-mentioned offense when carried out by public dissemination or the distribution of tracts, pictures, or other material; and
- publicly condoning, denying, or grossly trivializing crimes of genocide, crimes against humanity, and war crimes as defined in the Statute of the International Criminal Court (Articles 6, 7, and 8) and crimes defined in Article 6 of the Charter of the International Military Tribunal, when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group.

Instigating, aiding, or abetting in the commission of the above offenses is also punishable (*Framework decision*…, 2008).

It is worth noting that some authors distinguish between hate speech, offensive speech, and hate crime (Carlson, 2021). Hate speech can be understood as a broad phenomenon involving multimodal means of expression, not only lexical (Carlson, 2021). However, for the purposes of this article, the study will be limited to language samples. As Victoria Guillén Nieto (2022) claims in her book, *Hate Speech: Linguistic Approaches*, there has been no significant study of hate speech in the field of linguistics and it is treated as a peripheral phenomenon despite, in fact, being rather common. In this context, it is worth mentioning the paper "Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media"

(ElSherief et al., 2018), which provides a useful categorization of the phenomenon and proposes a nuanced view of hate speech. The authors divide hate speech into directed and generalized hate, as presented in Table 1.

Table 1. Directed and Generalized Hate

Directed hate	Generalized hate
"A shit sucking Muslim bigot like you wouldn't recognize history if I crawled up your c*bt. You think photoshop is truth machine" [sic]	"Why do so many filthy wetback helf--bread sp*k savages in #LosAngeles? None of them have any right at all to be here." [sic]
"shut the fuck up you stupid nigger I honestly hope you get brain cancer" [sic]	"Ready to make headlines. The #LGBT community is full of whores spreading AIDS like the Black Plague. Goodnight. Other people exist, too." [sic]
"bitch you breathe too much, shut the fuck up" [sic]	"Bringing weird niggers into my mentions!"
	"I'm partial, faggot..homo...both fairly describes the same sick twisted Assholes out there." [sic]

Note. Based on ElSherief et al. (2018).

As can be seen, importantly for this article, hate speech can be defined broadly and various factors need to be taken into account. Stereotypes, which are often present in online hate speech, can be both deprecatory and positive. Therefore, taking the context into account is the most difficult part of detecting hate speech online. It requires extra effort that cannot be digitized or automated, such as manual correction by annotators and qualifiers or using the "flagging" function on social media platforms. The latter mechanism is important because, for example, on YouTube comments are not initially screened, even automatically, and only when a user flags the content as inappropriate or offensive do human analysts investigate.

Research description and results

The research was carried out with the use of three different methods: analysis of a Twitter hashtag and of online databases with the help of an algorithm, with variables and interpretation based on standard linguistic tools and features described by Caleffi (2015), Scott (2017), and Zappavigna (2015), among others. The research was on the topic of sexism and hate speech directed towards feminists. Each of these will be subsequently discussed.

#HowToSpotAFeminist: An analysis of a Twitter hashtag

A hashtag is a tool that allows specific content to be tracked and followed (Scott, 2017, p. 2). It is defined as metadata (containing information about other data) that groups posts on the same topic. The metadata could be, for example, the date of publication and the medium or the location of the computer network on which the post was created. It is also a source of data about users of the same range of interests (Zappavigna, 2015, pp. 2–4). A hashtag is depicted by a "hash symbol" (#) preceding a text that omits spaces in a form of notation, as in the following example (hashtags emphasized by italics): @George: Tesla Long-Awaited Electric Pickup Will Be With No Handles *#Tesla #ElectricVehicles #PickUps #Automotive*.

The phrase preceded by the hashtag symbol automatically becomes a clickable link that redirects the user to a page summarizing all publications where the same hashtagged phrase appears. Since the hashtag phenomenon is relatively new, research on its linguistic aspect has only recently begun to emerge, so it is worth paying attention to its morphosyntactic features. Morphology determines how a word is conceived and comprehended, while syntax defines a sentence's structure (Kibort, 2010). The use of hashtags is contributing to a new mechanism of word formation (Caleffi, 2015). The idea of adding a hashtag to a string of words generates a new linguistic entity that may constitute different parts of speech or may perform different functions in a sentence. The number of hashtagged words and the type of characters – whether numbers, letters, acronyms, abbreviations, or entire words – should mainly be considered when evaluating them.

Table 2. Classification of Hashtags According to Syntactical and Linguistic features

Type of hashtag	Example
Abbreviation/acronym	#FBF (Flashback Friday)
Entire single word	#mountains
Sentence	#HowToSpotAFeminist
Phrase	#ThursdayThoughts, #quoteoftheday
Letters and digits	#RAF100

Note. Based on Caleffi, 2015, p. 53.

Following Caleffi's work, groups of acronyms, single words, sentences, and phrases composed of numbers and letters or solely letters were distinguished. The most characteristic feature of acronyms, whether they consist of only letters or are alphanumeric, is that they do not always refer only to commonly known abbreviations. On the contrary, they appear at a dynamic pace and can refer to current trends. As proof, in Table 2 we demonstrate an example of an acronym that stands for "Flashback Friday" and refers to uploading photos or posts referring to past events – on Fridays. It therefore embodies terminology used only in specific communities, in this case, social media. The most common type of hashtag is the one followed by solely one word. As a rule, these are nouns (including first names or proper names) or adjectives that constitute metadata in relation to the overall publication. Consequently, we distinguished other categories, that is, sentences and phrases. Since the classification of hashtags refers to constructions containing a sentence and a subject, while the phrases most often consist of a noun and an adjective or noun phrases, they were considered separately. Sentences in a hashtag form do not contain spaces. For this reason, it is common to capitalize every word. This practice is also used in reference to phrasal constructions. Aphorisms, which tend to be pragmatically loaded, appear in the sentence category, which is possible thanks to the longer, hash-tagged construction (Caleffi, 2015).

According to the systemic functional linguistics (SFL) theory, which examines the relationship between language and the role it plays in society, language has three functions: experiential, interpersonal, and textual. If the use of hashtags is treated as a linguistic phenomenon with a certain linguistic meaning, one could attempt to analyze it in terms of fulfilling these three functions of language (Zappavigna, 2015, p. 6).

Table 3. Hashtag Classification Based on Language's Functions

Function	Example
Experiential/empirical	Nose ring, tattoos, partly shaved head #howtospotafeminist
Interpersonal	A lot of people travel to the mountain to see the autumn leaves at the weekend. #sobeautiful
Textual	The #Picture is #SoBeautiful & #Inspirational.

Note. Based on Zappavigna (2015, p. 6).

In the use of hashtags, metacommentary resonating throughout the post is a representative attribute of the interpersonal function. The hash-tagged items most commonly found for this feature are emotionally charged epithets. In the example specified in Table 3, the noun "picture" and the adjectives "beautiful" and "inspirational" act as tags in addition to forming a coherent linguistic structure. Finally, the textual function mainly refers to the construction and organization of sentences in a post. Hash-tagged words, regardless of what part of speech they constitute, may be embedded within a sentence, forming its component element (Zappavigna, 2015). The experiential function in a text is fulfilled by a hashtag when its application emphasizes the topic mentioned in the text. It assumes a lemmatized feature, indicating what the post is about. The content of the post is not a sentence itself, but only an enumeration of features of a person or a group of people unknown to the reader. In this context, the hashtag is the main point of the message; it indicates the topic of the post. However, without the use of the hashtag, the post itself would be incomprehensible for the recipient. This is because it does not contain a subject, nor does it indicate the purpose of the statement. The act of tagging the question "how to spot a feminist" leads us to conclude that the characteristics mentioned within the message are meant to refer to the entire feminist community.

The hashtag "#HowToSpotAFeminist" originated in a Twitter post published in 2015 by a radio host promoting his show. The publication rapidly gained great popularity among opponents of the feminist movement. Considering the algorithm-based analysis that computers perform, it is likely that the sentence would not be categorized as offensive. This stems from the fact that the comment does not contain keywords: commonly known vulgarisms or slurs against a particular person or a group. Automated methods for detecting hate speech are prone to error due to the missing emotional and contextual factors. However, if one were to analyze the hashtag more deeply in terms of its syntactic and semantic construction, the very act of juxtaposing the predicate "spot" with the object "feminist" in the same sentence indicates its stereotypical nature and spawns the basis for discriminatory stimuli. The author of the hashtag and his followers seem to assume that the personal or physical features they name are typical of all feminists. Thus, the hashtag comes down to creating a set of assumptions about feminists in the minds of the audience, or to reinforcing an existing stereotype. In consequence, the reproduction of negative stereotypes may lead to discriminatory behavior, which in turn may entail physical or verbal violence in the form of hate speech. The stereotypical judgments generated in relation to the hashtag are presented in Table 4. Each of the statements below pigeonholes people who support the feminist movement. It is worth noting that the posts imply that being a woman is synonymous with being a feminist.

Table 4. Stereotypes Attributed to Feminists Using the Tag #HowToSpotAFeminist

Example	Category
#HowToSpotAFeminist - Usually fat & ugly, always inherently unlikeable, supremely hypocritical, snarky, annoying, deluded, intransigent.	appe-arance, attitude
#HowToSpotAFeminist is someone who studied Social Sciences, but wants to earn like an Engineer. And if she doesn't, she calls it "inequality"	education
They're the only one's without a date #HowToSpotAFeminist	love life
#Howtospotafeminist They are wearing pantsuits!!!!	dress code

Note. Based on comments found on Twitter.

Table 4 categorizes the examples given according to the type of stereotypes attributed to feminists. Thus, the first and last ones emphasize features of the appearance and character of female feminists. While the irony is palpable, the utterances are free of vulgarisms or offensive words. The second and third examples refer to personal features, education, and love life. According to these comments, feminists claim equal pay under conditions of unequal positions on the labor market, with a stereotypical reference to supposedly less demanding studies and jobs that women tend to perform (e.g., sociology vs. engineering).

As can be seen from the above analysis, if the samples were subjected to an automated hate speech detection mechanism, some of them could be missed, whereas in fact they would qualify as hate speech. This is due to the fact that neither the hashtag itself nor the comments contain keywords, which are central to the methodology of algorithms. Thus, qualitative analysis is necessary in such cases.

Analysis of a sexism-related online hate speech database

The analysis of the online hate speech database was performed with an algorithm written in the programming language Python. The features of Natural Language Tool Kit – a set of Python libraries for statistical language processing for English – and of the Tweepy library – a library for connecting with the Twitter platform – were used to create the code. By means of a special questionnaire, permission was requested from Twitter for an individual key, which was thereafter incorporated into the code to conduct the mass research. In parallel, other libraries containing various statistical functions for linguistic features were imported. With them, the necessary data pieces could be extracted from tweets and subjected to statistical analysis. The data set was compared with a control group, data with no hate speech detected. The aim of this research was to analyze the linguistic properties of words/phrases, such as the number of words in a sentence, the number of compound sentences in one comment, the lexical diversity (a measurement of the number of different words present in a text), the number of lexemes, the variety in parts of speech, and gender statistics.

The database concerns the subject of sexism. It was created using 10,460 comments which, as indicated on the website hatespeechdata.com, were classified by annotators into Group 1 (offensive comments) or Group 2 (free of hate speech utterances).

Table 5. Results of Quantitative Comparative Analysis for Most Frequent Variables in the Sexism-Related Database

	Hate speech (2,740)	**Non-hate speech (7,720)**
Noun	4.56	3.24
Verb	190	1.35
Personal pronoun	0.99	0.78
Adjective	1.24	1.06
Adverb	1.06	0.77
Preposition/ Subordinating conjunction	1.25	1.24
Proper name	2.54	1.68
Determiner	1.11	0.97
Mean number of words per message	21.0	11.0
Mean typos corrected	1.0	1.0
Lexical diversity	0.209009	0.144926

Note. Based on the database available at hatespeechdata.com.

As reflected in Table 5, all the variables appeared statistically more frequently in the Group 1 comments than in the Group 2 comments. However, this difference never reached 2.0. Statistically, 1.90 verbs appear in each hate speech comment, whereas there are 1.35 in every utterance of Group 2. The largest variation was identified in the mean length of comments. While in Group 2 the utterances averaged 11 words, in Group 1 the average was 21. The fact that hate speech utterances appear to be nearly twice as long as those free of offensive language may be the main contributing factor to the higher occurrence of other variables seen in Group 1, such as the individual parts of speech listed in Table 5, the lexical diversity, or the use of the autocorrection tool.

Table 6. Gender Statistics

	Group 1: Hate speech	Group 2: Non-hate speech
Female	2.45	2.10
Male	3.68	2.83

Note. Based on the database provided at hatespeechdata.com.

According to figures presented in Table 6, in both groups, female names appeared about 1.5 times less frequently than male names (1.5 and 1.34, respectively); the latter were more prevalent in the sexist statements than in others.

The study of the predominantly used lexemes returned the following results:

Group 1: sexist, call, like, female, men, think, woman, get

Group 2: like, people, get, think, go, really

It is noteworthy that apart from most of the lexemes, which were rather semantically neutral, the offensive comments contained words that classify and specify a particular gender ("woman," "men," or "female"), while the statements that are free of hate speech are also free of gender-specific terms. Instead, the authors of comments that do not categorize any gender as superior to another statistically more often applied the lexeme "people," whose semantic meaning does not indicate any specific sex.

Table 7. Statistics on Personal Pronouns from Sexism-Related Database

	Hate speech (2740)	Non-hate speech (7720)
I	0.33	0.24
you	0.19	0.15
he	0.02	0.02
she	0.05	0.01
we	0.03	0.04
they	0.10	0.07

Note. Based on the database provided at hatespeechdata.com.

In the analysis of the next variable, personal pronouns, the pronouns "I" and "you" were most commonly detected in both groups. Both pronouns were slightly more often applied in the hate speech group than in the group free of vulgarity or disparaging speech. One may deduce that such a trend indicates a higher proportion of direct hate speech, confronting two actors of a message, rather than commenting on a third party.

Conclusions

The research has shown that hate speech needs further definition in both international law and linguistics. Furthermore, it has demonstrated the need for interdisciplinary analysis of content by human linguists with special training and familiarity with the legislative frameworks of a given linguistic area, as the possibilities offered by linguistic algorithms are still limited. The quantitative analysis revealed various patterns and features that may be observed in hate speech, while the qualitative study provided a more profound understanding of those language items in context. The corpus chosen for the study was screened mainly for sexist content; as the literature on the subject suggests, such research can be conducted for any group prone to discrimination.

References

Assimakopoulos, S., Baider, F. H., & Millar, S. (eds.). (2017). *Online hate speech in the European Union: A discourse-analytic perspective*. Springer.

Caleffi, P.-M. (2015). The "hashtag": A new word or a new rule? *SKASE Journal of Theoretical Linguistics*, 12(2), 46–69.

Carlson, C.R. (2021). *Hate speech*. MIT Press.

ElSherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. University of California. Retrieved from: https://arxiv.org/pdf/1804.04257.pdf (access: 25.02.2023).

Framework decision on combating certain forms and expressions of racism and xenophobia by means of criminal law. 2008/913/JHA of 28 November 2008. Eur-lex. (2014). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al33178. (access: 21.02.2023).

Hatespeeechdata.com (2021). Retrieved February 21, 2023.

Kibort, A. (2010). *A typology of grammatical features*. Grammaticalfeatures.net. Retrieved from: http://www.grammaticalfeatures.net/inventory.html (access: 25.02.2023).

Nieto, V.G. (2022). *Hate speech: Linguistic approaches*. De Gruyter.

Pejchal, V. (2020). *Hate speech and human rights in Eastern Europe: Legislating for divergent values*. Routledge.

Rasinger, S. (2018). *Quantitative research in linguistics: An introduction* (2nd Ed.). Bloomsbury.

Romano, A. (2015, May 5). Sexist #HowToSpotAFeminist hashtag is reclaimed by feminists on Twitter. *Mashable*. Retrieved from: https://mashable.com/2015/05/05/how-to-spot-a-feminist (access: 25.02.2023).

Scott, K. (2018). “Hashtags work everywhere”: The pragmatic functions of spoken hashtags. *Discourse, Context & Media*, Vol. 22, 57–64.

Statista.com (2021).

UNSPAHS. (2019). *United Nations strategy and plan of action on hate speech*. Retrieved from: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf (access: 10.01.2023).

Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics*, 25(3), 2–4.

Anna Bugajska – an Associate Professor at the Jesuit University Ignatianum in Krakow, where she heads the Language and Culture Studies Department. Her research interests and publications include the intersections of philosophy, sociology, language, and culture – especially in relation to emergent technologies. She is also involved in various programs and projects concerning contemporary ethical problems. She is the author of *Engineering Youth: The Evantropian Project in Young Adult Dystopias* (2019).

Paulina Dziedzic – a business analyst. She earned her degree from the Tischner European University in Krakow in the field of linguistics. Her interests are focused on exploring the emotional intelligence that influences social media interactions and the social phenomena that shape them. During her work as a content analyst at one of the most famous global online video sharing platforms, she developed a keen interest in the issue of hate speech in social media. Her experience provided her with a firsthand understanding of the extent to which this phenomenon exists online. This experience, combined with her training as an English philologist (specializing in intercultural communication in business), inspired her to pursue research into the linguistic and psychological factors that contribute to hate speech as a conditioned phenomenon in English-language culture.