ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 81'42:004.8

ЛИНГВИСТИЧЕСКИЕ МАРКЕРЫ ЯЗЫКА НЕНАВИСТИ В ИНТЕРНЕТ-ДИСКУРСЕ И ИХ ИСПОЛЬЗОВАНИЕ В СИСТЕМАХ АВТОМАТИЧЕСКОГО ОБНАРУЖЕНИЯ

Критский Сергей Димитриевич,
Федеральное государственное автономное образовательное учреждение высшего образования Российский университет дружбы народов имени Патриса Лумумбы, г. Москва
adv.sk2521@yandex.ru

Аннотация

В статье рассматриваются лингвистические маркеры языка ненависти (hate speech) в пространстве интернет-дискурса и анализируются подходы к их использованию в системах автоматического обнаружения враждебного контента. Выявлены ключевые лексико-семантические, грамматические и прагматические маркеры, характеризующие враждебные высказывания в сетевом пространстве. Проведён анализ методов машинного обучения и моделей на основе трансформеров (BERT, ELECTRA), применяемых для идентификации языка ненависти в текстах социальных сетей. Установлено, что наибольшую сложность для автоматических классификаторов представляют имплицитные формы враждебности: ирония, кодированная лексика и контекстно-зависимые высказывания. Сделан вывод о необходимости междисциплинарного подхода, объединяющего лингвистическую теорию и компьютерные технологии.

Ключевые слова: язык ненависти, лингвистические маркеры, интернет-дискурс, речевая агрессия, автоматическое обнаружение, машинное обучение, NLP, классификация текстов.

LINGUISTIC MARKERS OF HATE SPEECH IN INTERNET DISCOURSE AND THEIR USE IN AUTOMATIC DETECTION SYSTEMS

Kritskiy Sergey Dimitrievich,
Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow

ABSTRACT

The article examines the linguistic markers of hate speech in the space of Internet discourse and analyzes approaches to their use in systems for automatic detection of hostile content. The key lexico-semantic, grammatical and pragmatic markers characterizing hostile statements in the online space have been identified. The analysis of machine learning methods and transformer-based models (BERT, ELECTRA) used to identify hate speech in social media texts is carried out. It has been found that implicit forms of hostility are the most difficult for automatic classifiers: irony, coded vocabulary, and context-sensitive statements. The conclusion is made about the need for an interdisciplinary approach combining linguistic theory and computer technology.

Keywords: hate speech, linguistic markers, Internet discourse, speech aggression, automatic detection, machine learning, NLP, text classification.

Актуальность

Проблема распространения языка ненависти в интернет-пространстве приобрела глобальный характер. Враждебный контент в социальных сетях представляет серьёзную угрозу для социальной стабильности и прав человека. Язык ненависти — это высказывания, направленные на дискриминацию или разжигание вражды к лицам или группам на основании расовой, этнической, религиозной, гендерной или иной принадлежности. В российском правовом поле данное явление регулируется ст. 282 УК РФ, однако масштаб интернет-коммуникации делает ручной мониторинг невозможным. Лингвистический анализ маркеров враждебности является фундаментальной базой для создания автоматизированных систем, поскольку именно языковые средства выступают формальным носителем враждебной интенции [1, с. 11].

Цель исследования

Систематизировать лингвистические маркеры языка ненависти в интернет-дискурсе и проанализировать подходы к их использованию в современных системах автоматического обнаружения враждебного контента.

Материалы и методы исследования

Исследование выполнено методом аналитического обзора научной литературы по лингвистике, компьютерной обработке естественного языка (NLP) и дискурс-анализу. Рассматриваемый материал включает данные о функционировании враждебных высказываний в русскоязычных и англоязычных интернет-источниках: комментариях, постах в социальных сетях, записях в мессенджерах. Методологическую основу составляют дискурс-анализ, контент-анализ и сравнительный метод.

Результаты и их обсуждение

Лингвистические маркеры языка ненависти представляют собой разноуровневые языковые средства, по которым высказывание идентифицируется как враждебное. На лексико-семантическом уровне ключевым маркером является инвективная лексика — бранные слова, пейоративы, этнофолизмы. О.С. Коробкова установила, что этнофолизмы формируются посредством метафорического переноса, суффиксальной деривации и заимствований из жаргона; экстралингвистические факторы (миграция, политические конфликты) непосредственно влияют на появление таких единиц [2, с. 203]. Н.Е. Петрова и Л.В. Рацибурская описывают средства речевой агрессии в медиатекстах: словообразовательные неологизмы с негативной коннотацией (например, «квазиспециалист», «гайдаровщина»), жаргонную и просторечную лексику, иноязычные заимствования в дискредитирующей функции, агрессивные метафоры и сравнения [3, с. 34].

Особую роль играет дихотомия «свой — чужой», реализуемая через разделительные местоимения («мы» и «они»), обобщающие негативные характеристики и стереотипные предикаты. А.А. Сычев, Е.А. Коваль и Н.В. Жадунова установили, что основой языка вражды в отношении мигрантов являются негативные стереотипы, конструирующие восприятие мигрантов как источника экономических, медицинских и политических угроз, и предложили содержательные (7 типов угроз безопасности) и формальные (текстуальные, контекстуальные, метаконтекстуальные) критерии классификации [4, с. 802].

На грамматическом уровне маркерами выступают восклицательные и побудительные конструкции, риторические вопросы с имплицитной агрессией, императивы с угрожающей семантикой. Ю.В. Щербинина подчёркивает, что речевая

агрессия реализуется через специфический набор языковых средств — лексику, интонацию и способ организации высказывания, целью которых является «коммуникативное подчинение адресата, осуществление коммуникативного давления» [5, с. 33]. Т.А. Воронцова, анализируя троллинг и флейминг, показала, что в интернет-коммуникации агрессивные речевые стратегии активизируются благодаря анонимности и дистанцированности участников, при этом провокативное поведение нередко принимает системный характер [6, с. 112].

На прагматическом уровне выделяются дегуманизация (сравнение людей с животными), делегитимация (отказ группе в праве на человеческое достоинство) и использование фейковой информации. А.Н. Баранов обращает внимание на скрытые (имплицитные) утверждения, когда враждебный смысл не выражен прямо, а выводится из контекста, пресуппозиций и импликатур — такие случаи представляют наибольшую трудность для лингвистической экспертизы и автоматических систем [7, с. 45].

Переход к автоматическому обнаружению требует формализации выявленных маркеров. В обзоре P. Fortuna и S. Nunes систематизированы подходы к автоматической детекции: от словарных фильтров и классических алгоритмов машинного обучения (SVM, Naive Bayes, Random Forest) до нейросетевых моделей. Ранние системы опирались на лексические признаки — наличие слов из словарей оскорблений, однако давали высокий процент ложноположительных (нейтральные тексты с ненормативной лексикой) и ложноотрицательных результатов [8, с. 5]. T. Davidson и соавторы на выборке из 25000 твитов показали, что расистские и гомофобные высказывания чаще корректно распознаются как язык ненависти, тогда как сексистские нередко классифицируются лишь как «оскорбительные», что указывает на необходимость учёта семантических и контекстуальных признаков [9, с. 512].

Современные трансформерные модели (BERT, RoBERTa, ELECTRA) существенно продвинули качество детекции благодаря способности учитывать контекстные зависимости и улавливать скрытые семантические паттерны. На наборе MetaHate (1,2 млн образцов из 36 датасетов) модель ELECTRA достигла F1-score 0,898, превосходя BERT и RoBERTa. Однако сарказм, кодированный язык (замена оскорбительных слов эвфемизмами или неологизмами) и аннотационный шум в обучающих данных остаются нерешёнными проблемами. А.Н. Величко, исследуя автоматическое определение агрессии на русскоязычных данных методом ансамбля случайных лесов с акустическими признаками, достигла показателя UAR 76,5%, продемонстрировав конкурентоспособность комбинированных подходов в мультиклассовых задачах [10, с. 185].

Выводы

Лингвистические маркеры языка ненависти образуют многоуровневую систему: лексико-семантические (инвективы, этнофолизмы, пейоративы, агрессивные метафоры), грамматические (побудительные конструкции, риторические вопросы, императивы) и прагматические (дегуманизация, делегитимация, имплицитные утверждения) компоненты. Автоматические системы обнаружения эволюционировали от словарных фильтров к нейросетевым моделям на основе трансформеров, способным учитывать контекст и семантику. Ключевой проблемой остаётся детекция имплицитных форм враждебности — сарказма, кодированного языка и контекстно-зависимых высказываний. Перспективным направлением является создание русскоязычных аннотированных корпусов и разработка гибридных моделей, сочетающих лингвистические правила и методы глубокого обучения.

Список литературы:

1. Комалова Л.Р. Язык и речевая агрессия: аналит. обзор / Отв. ред. Яковлева Э.Б. М.: ИНИОН РАН, 2015. 75 с.
2. Коробкова О.С. Маркеры языка вражды в номинациях этнической принадлежности: социолингвистический аспект // Известия РГПУ им. А.И. Герцена. 2009. № 111. С. 200–205.
3. Петрова Н.Е., Рацибурская Л.В. Язык современных СМИ: средства речевой агрессии: учеб. пособие. М.: Флинта: Наука, 2011. 160 с.
4. Сычев А.А., Коваль Е.А., Жадунова Н.В. Проблема классификации языка вражды в отношении мигрантов (на примере Республики Мордовия) // Регионология. 2018. Т. 26, № 4. С. 798–815.
5. Щербинина Ю.В. Русский язык: Речевая агрессия и пути ее преодоления. М.: Флинта: Наука, 2004. 224 с.
6. Воронцова Т.А. Троллинг и флейминг: речевая агрессия в интернет-коммуникации // Вестник Удмуртского университета. Сер.: История и филология. 2016. Т. 26, № 2. С. 109–116.
7. Баранов А.Н. Лингвистическая экспертиза текста: теория и практика: учеб. пособие. 6-е изд. М.: Флинта, 2018. 592 с.
8. Fortuna P., Nunes S. A Survey on Automatic Detection of Hate Speech in Text // ACM Computing Surveys. 2018. Vol. 51, No. 4. Article 85. P. 1–30.
9. Davidson T., Warmsley D., Macy M., Weber I. Automated Hate Speech Detection and the Problem of Offensive Language // Proceedings of the International AAAI Conference on Web and social media. 2017. Vol. 11, No. 1. P. 512–515.
10. Величко А.Н. Метод анализа речевого сигнала для автоматического определения агрессии в разговорной речи // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2022. № 4. С. 180–188.

References:

1. Komalova L.R. Yazyk i rechevaya agressiya: analit. obzor [Language and Speech Aggression: Analytical Review]. Moscow, INION RAN, 2015. 75 p. (In Russ.)
2. Korobkova O.S. [Hate speech markers in ethnic membership nominations: sociolinguistic aspect]. Izvestiya RGPU im. A.I. Gertsena, 2009, no. 111, pp. 200–205. (In Russ.)
3. Petrova N.E., Ratsiburskaya L.V. Yazyk sovremennykh SMI: sredstva rechevoj agressii [Language of Modern Media: Means of Speech Aggression]. Moscow, Flinta, Nauka, 2011. 160 p. (In Russ.)
4. Sychev A.A., Koval E.A., Zhadunova N.V. [The issue of classification of hate speech against migrants (the case of the Republic of Mordovia)]. Regionologiya, 2018, vol. 26, no. 4, pp. 798–815. (In Russ.)
5. Shcherbinina Yu.V. Russkij yazyk: Rechevaya agressiya i puti ee preodoleniya [Russian Language: Speech Aggression and Ways to Overcome It]. Moscow, Flinta, Nauka, 2004. 224 p. (In Russ.)

6. Vorontsova T.A. [Trolling and flaming: speech aggression in internet communication]. Vestnik Udmurtskogo universiteta. Ser.: Istoriya i filologiya, 2016, vol. 26, no. 2, pp. 109–116. (In Russ.)
7. Baranov A.N. Lingvisticheskaya ekspertiza teksta: teoriya i praktika [Linguistic Expert Examination of Text: Theory and Practice]. 6th ed. Moscow, Flinta, 2018. 592 p. (In Russ.)
8. Fortuna P., Nunes S. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 2018, vol. 51, no. 4, article 85, pp. 1–30.
9. Davidson T., Warmsley D., Macy M., Weber I. Automated Hate Speech Detection, and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and social media, 2017, vol. 11, no. 1, pp. 512–515.
10. Velichko A.N. [Method of speech signal analysis for automatic aggression detection in conversational speech]. Vestnik VGU. Seriya: Sistemnyj analiz i informacionnye tekhnologii, 2022, no. 4, pp. 180–188. (In Russ.)