DOI: 10.55648/1998-6920-2025-19-2-81-88 УДК 004.852

# Модель классификации вербальной агрессии в неструктурированных текстах

И. Е. Воронина, М. К. Пастревич

Воронежский государственный университет

*Аннотация*: Рассматривается программный комплекс для автоматической классификации агрессивной лексики в неструктурированных текстах в информационном пространстве с учетом контекстного анализа высказываний на лексико-семантическом уровне. Экспериментальным путем был выбран один из многих векторизаторов. Усовершенствована существующая классификация речевой агрессии, предназначенная для машинного обучения.

*Ключевые слова*: классификация, модель, обучение, вербальная агрессия, нейронная сеть.

*Для цитирования*: Воронина И. Е., Пастревич М. К. Модель классификации вербальной агрессии в неструктурированных текстах // Вестник СибГУТИ. 2025. Т. 19, № 2. С. 81–88. https://doi.org/ 10.55648/1998-6920-2025-19-2-81-88.



Статья поступила в редакцию 24.10.2024;
переработанный вариант – 04.12.2024;
принята к публикации 10.12.2024.

## 1. Введение

В современном информационном мире распространение в социальных сетях таких понятий, как ненависть и агрессия, стало серьезной проблемой, требующей особого внимания исследователей. В глобальном смысле классификация вербальной агрессии представляет собой довольно сложную задачу, которая обуславливается многообразием выражений различных агрессивных намерений и контекстуальными условиями. Следует отметить, что в информационном пространстве в момент коммуникации с оппонентом существует возможность проявления агрессивной лексики, от проявления которой не застрахован ни один пользователь.

Классификацией речевой агрессии занимались лингвисты, например Седов К. Ф. [1], Енина Л. В. [2], Михалская А. К. [3], Паламарчук Н. А. [4], Шейгал Е. И. [5], а также психологи, например Баас А. [6], Жельвис В. И. [7], но ими не рассматривалась автоматическая классификация.

Вопросами классификации неструктурированных текстов в информационном пространстве занимались такие ученые, как Рубцова Ю. В. [8], Капитанов А. И. [9], Котельников Е. В. [10], Кижаева Н. А. [11], однако ими не затрагивалась проблема классификации вербальной агрессии в неструктурированных текстах.

Набирающие популярность нейросети решают многие задачи при работе текстом (ответы на вопросы, написание писем и статей, составление персональных советов, написание приглашения на встречу), но не могут классифицировать вербальную агрессию в текстах.

В связи с этим возникает необходимость создания автоматической классификации речевой агрессии в неструктурированных текстах в информационном пространстве.

Для осуществления поставленной цели необходимо решить такие задачи, как анализ существующих классификаций агрессии, выбор одной из них и дальнейшее ее усовершенствование, а также создание специализированного программного комплекса для автоматической классификации вербальной агрессии в неструктурированных текстах в информационном пространстве.

## 2. Материалы и методы

В дальнейшей работе под речевой агрессией понимается направленное на объект интенциональное авторское действие, отличительным свойством которого является заключение в культурную и национальную специфику коммуникации и выражение с помощью различных специальных языковых средств. Целью такого действия является подавление воли оппонента или же осуществление на него коммуникативного давления [12].

Среди многообразия классификаций вербальной агрессии самой подходящей для классификации неструктурированных текстов в информационном пространстве являются [5]:

1. Эксплицитная. Представлена в виде брани, различных речевых угроз, например: «Вешать таких баранов надо».
2. Манипулятивная. К ней относят запрет на речь, например: «Да заткнись ты уже».
3. Имплицитная. Данный вид словесной агрессии характеризует косвенные речевые акты, иронию или сарказм, например: «Ой, глянь, самый умный тут».

Но необходимо добавить дополнительное условие для дальнейшей классификации – «нейтральная лексика». Следует отметить, что при работе с обучающим набором данных необходимо учитывать контекст, когда одна фраза может иметь разный смысл.

При разработке программного комплекса автоматической классификации вербальной агрессии в неструктурированных текстах в ходе анализа различных алгоритмов классификации был выбран мультиномиальный наивный байесовский классификатор (MNB). Отметим, что данный метод при прогнозировании метки текста каждый раз вычисляет вероятность меток для входного текста. В момент распределения используемые параметры вектора выглядят следующим образом — $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$, где $n$ – количество признаков, а в случае классификации текста считается размером словарного запаса. Параметры $\theta_y$ учитываются с помощью сглаженной версии подсчета относительной частоты:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}, \tag{1}$$

где $N_{yi} = \sum x \in T$, $T$ – тренировочный набор, $N_{yi}$ является общим количеством всех функций класса $y$. Если $\alpha \geq 0$, тогда учитываются функции, которые отсутствуют в обучающих выборках, предотвращая затем нулевую вероятность в последующих вычислениях. Если $\alpha = 1$, то параметр является сглаживанием Лапласа [14].

Корпус данных, реализуемый в программном комплексе, имеет следующий вид:

$$(x_i, y_i)_i^L = 0, \tag{2}$$

где $x_i \in \mathrm{R}^n$ – $i$-й комментарий пользователя, $y_i \in \{0, 1, 2, 3\}$ является метками класса. Для каждого комментария устанавливает соответствующую метку функция $F$:

$$F(x_i) = y_i \tag{3}$$

Помимо этого, используется оценка максимального правдоподобия, которая определяет различные неизвестные параметры при распределении вероятностей. Данная оценка вычисляется по следующей формуле:

$$P\left(\frac{x}{y}\right) = \frac{N!}{\prod_{i=1}^{n} x_i!} \cdot \prod_{i=1}^{n} (Pw_i)^{x_i}. \tag{4}$$

Корпус данных состоит из комментариев пользователей таких социальных сетей, как Одноклассники, Вконтакте, Пикабу; мессенджеров Telegram (группы «Мой и твой Воронеж», «MOUNT SHOW»), WhatsApp (группа «Чат дома»). Размер корпуса данных равен 143232 комментариям, собранным и обработанным вручную в соответствии с выбранным определением речевой агрессии, манипулятивной лексики и классификацией вербальной агрессии. Распределение комментариев представлено на рис.1.

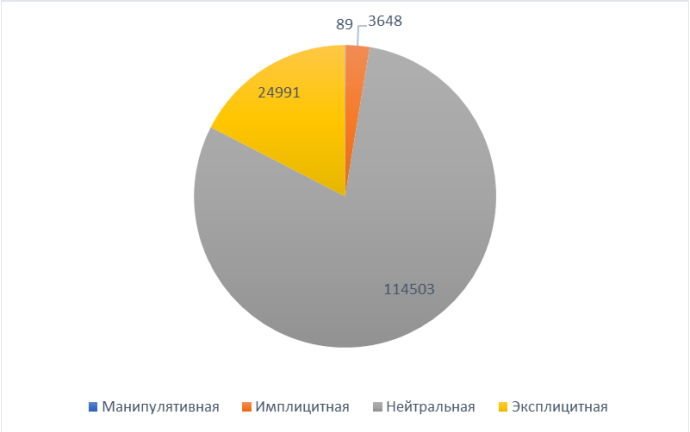


Рис.1. Распределение комментариев

В связи с тем, что при работе с классификацией текста необходимо использовать оценку ее качества, то при разработке программного комплекса для классификации вербальной агрессии в информационном пространстве применяются следующие метрики:

- Accuracy – является долей верных ответов среди всех предсказаний;
- Precision – необходимая часть истинно положительных ответов среди всех возможных положительных ответов;
- Recall – является необходимой долей истинно положительных ответов среди всех возможных верных ответов;
- F-мера – гармоничное среднее между точностью и полнотой.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{5}$$

где $TP$ – истинно положительное решение, когда классификатор правильно отнес объект к рассматриваемому классу; $FP$ – ложно положительное решение, в результате которого классификатор неверно отнес объект к рассматриваемому классу; $FN$ – ложно отрицательное решение, в результате которого классификатор неверно утверждает тот факт, что объект не принадлежит к рассматриваемому классу; $TN$ – истинно отрицательное решение, в результате которого классификатор дает верное утверждение о том, что объект не принадлежит к рассматриваемому классу [13].

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

$$F\text{-}мера = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{8}$$

## 3. Комплекс для классификации вербальной агрессии в информационном пространстве

При разработке программного комплекса для классификации речевой агрессии помимо мультиномиального байесовского классификатора был использован векторизатор HashVectorizer, показывающий более высокую скорость и точность при работе с большими данными, по сравнению с другими векторизаторами. В результате точность модели составляет 0.88.

Программный код разрабатывался в среде Python с использованием таких библиотек, как:

- Pandas (используется для визуализации данных);
- Re (необходима для сопоставления данных с шаблонами);
- Scikit-learn (позволяет работать не только с большим набором данных и в машинном обучении, а также поддерживает возможности классификации или уменьшения размерности);
- Matplotlib (используется для построения графиков);
- Flask (применяется для создания веб-сайтов).

Программная реализация задачи автоматической классификации вербальной агрессии состоит из следующих частей:

- модуль нормализации текстовых сообщений (приведение набора данных к первой нормальной форме);
- модуль получения векторного представления текстовых данных (подключение HashVectorizer);
- модуль для классификации текстовых сообщений.

На рис. 2 представлен общий вид классификатора.

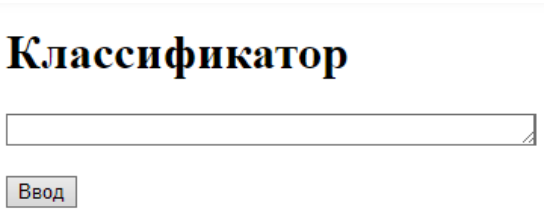


Рис.2. Классификатор вербальной агрессии

На рис.3. представлена классификация эксплицитной лексики.

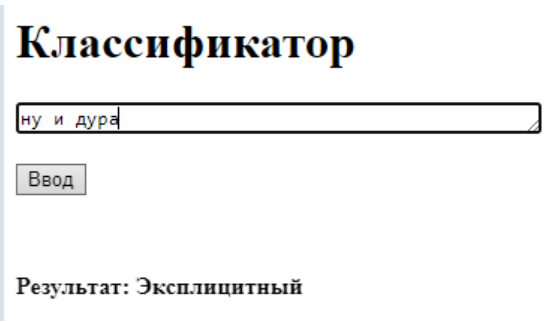


Рис.3. Эксплицитная лексика

На рис. 4 представлен пример классификации эксплицитной лексики.

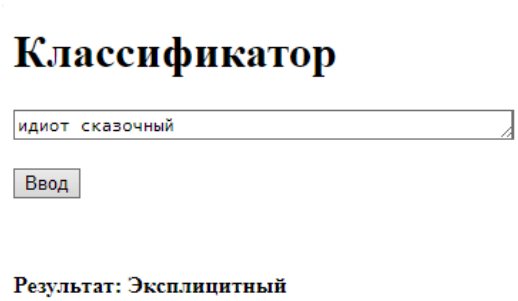


Рис.4. Пример эксплицитной лексики

На рис. 5 представлена классификация нейтральной лексики.

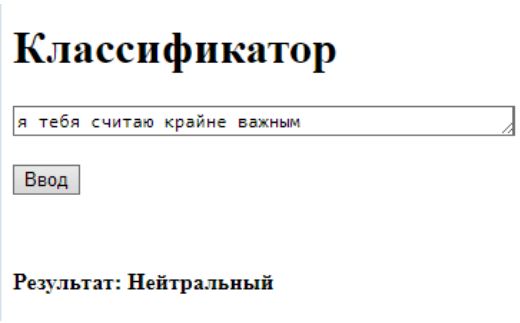


Рис.5. Нейтральная лексика

На рис. 6 представлен пример нейтральной лексики.

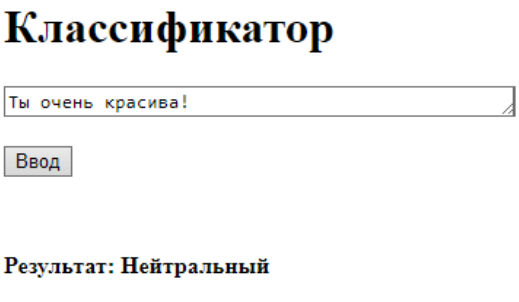


Рис.6. Пример нейтральной лексики

На рис. 7 представлена классификация имплицитной лексики.

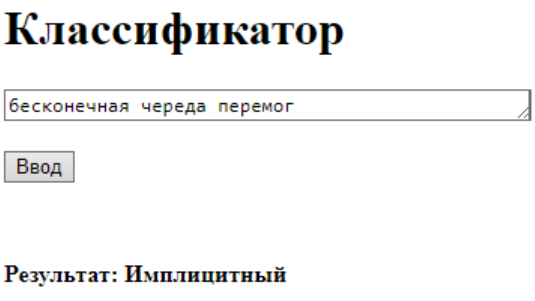


Рис.7. Имплицитная лексика

На рис. 8 представлен пример классификации имплицитной лексики.

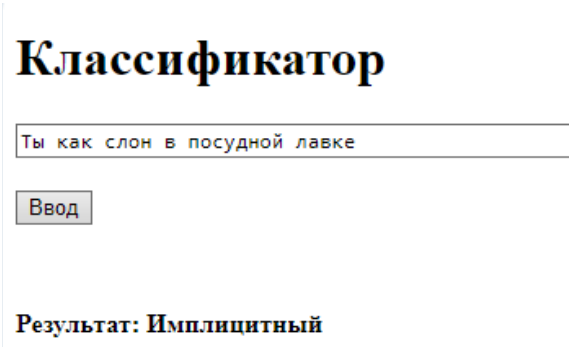


Рис.8. Пример имплицитной лексики.

## 4. Заключение

Данный комплекс может применяться как администратором групп в социальных сетях для возможной блокировки контента, так и пользователем для определения проявления вербальной агрессии в комментариях или постах.

Как показали исследования, манипулятивная лексика в информационном пространстве используется редко, в связи с чем некорректно отображается при работе комплекса. Для улучшения точности программного комплекса необходимо расширение корпуса данных именно в этой части.

На данный момент программный комплекс представлен в web-версии, в дальнейшем планируется расширение до плагина. Помимо этого, планируется расширение корпуса данных для последующего обучения модели.

## Литература

1. *Седов К. Ф.* Агрессия и манипуляция в повседневной коммуникации / К. Ф. Седов // Юрислингвистика-6: инвективное и манипулятивное функционирование языка. – 2005. – № -. – С. 89-105.
2. *Енина Л. В.* Современные российские лозунги как сверхтекст: автореф. дис. … канд. филол. наук : 10.02.01 / Енина Лидия Владимировна. – Екатеринбург, 1999. – 18 с.
3. *Михальская А. К.* Основы риторики: Мысль и слово: учебное пособие для учащихся 10-11 классов / А. К. Михальская. – Москва : Просвещение, 1996. – 416 с.
4. *Паламарчук Н. А.* Способы выражения агрессии в текстах Интернет-комментариев / Н. А. Паламарчук // Актуальные проблемы теоретической и прикладной лингвистики. – 2011. – С. 19-28.
5. *Шейгал Е. И.* Семиотика политического дискурса: монография / Е.И.Шейгал. – Волгоград: Перемена, 2000. – 367 с.
6. *Buss A. H.* The Psychology of Aggression / A.H. Buss. – Michigan: Wiley, 1961. – 307 с.
7. *Жельвис В. И*, Поле брани. Сквернословие как социальная проблема в языках и культурах мира / В. И, Жельвис. – Москва : Ладомир, 1997. – 453 с.
8. *Рубцова Ю. В.* Методы и алгоритмы построения информационных систем для классификации текстов социальных сетей по тональности: дис. … канд. техн. наук : 05.13.17 / Рубцова Юлия Владимировна. – Новосибирск, 2019. – 141 с.
9. *Капитанов А. И.* Разработка методики и алгоритмов линейной классификации неструктурированной текстовой информации в технических системах: дис. … канд. техн. наук : 05.13.01 / Капитанов Андрей Иванович. – Москва, 2022. – 122 с.
10. *Котельников Е. В.* Методология интеллектуального анализа мнений при обработке текстовой информации на основе правдоподобного вывода: автореф. дис. … канд.

техн. наук : 05.13.17 / Котельников Евгений Вячеславович. – Нижний Новгород, 2019. – 39 с.
11. *Кижаева Н. А.* Исследование паттернов в текстах на основе динамических моделей: дис. … канд. физ.-мат. наук : 01.01.09 / Кижаева Наталья Александровна. – Санкт-Петербург, 2018. – 86 с.
12. *Черкасова М. Н.* Речевые формы агрессии в тестах СМИ: моногр./ М. Н.Черкасова. – Ростов-на-Дону / Ростовский государственный университет путей сообщения, 2011. – 123 с.
13. Scikit-learn [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes (дата обращения: 30.09.2024).

**Воронина Ирина Евгеньевна**
д.т.н., профессор кафедры программного обеспечения и администрирования информационных систем, Воронежский государственный университет (ФГБОУ ВО ВГУ, 394018, Россия, г. Воронеж, Университетская площадь, 1), +7 (473) 220-82-66, e-mail: irina.voronina@gmail.com.

**Пастревич Марина Константиновна**
преподаватель кафедры программного обеспечения и администрирования информационных систем, Воронежский государственный университет (ФГБОУ ВО ВГУ, 394018, Россия, г. Воронеж, Университетская площадь, 1), +7 (473) 220-82-66, e-mail: mirstat@mail.ru.

*Авторы прочитали и одобрили окончательный вариант рукописи.*
*Авторы заявляют об отсутствии конфликта интересов.*
Вклад соавторов: *Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.*

**Model for Classifying Verbal Aggression in Unstructured Texts**

**Irina E. Voronina, Marina K. Pastrevich**

Voronezh State University (VSU)

*Abstract*: a software package for automatic classification of aggressive vocabulary in unstructured texts in the information space is considered, taking into account the contextual analysis of statements at the lexical-semantic level and justification for the choice of a vectorizer and the most optimal classification for machine learning.

*Keywords*: classification, model, training, verbal aggression, neural network.

*For citation*: Voronina I. E., Pastrevich M. K. Sistema i algoritm v upravlenii kachestvom pravotvorchestva RF [System and algorithm in managing the quality of lawmaking in the Russian Federation] // *Vestnik SIBGUTI,* 2025, vol. 19, no. 2, pp. –. https://doi.org/10.55648/1998-6920-2025-19-2-3-9.



The article has been received by the editors 24.10.2024;
revised version – 04.12.2024;
accepted for publication 10.12.2024.

## References


1. *Sedov K. F. Agressiya i manipulyaciya v povsednevnoj kommunikacii* [Aggression and manipulation in everyday communication]. Jurislinguistics-6: invective and manipulative functioning of language, 2005. P. 89-105.
2. *Enina L. V. Sovremennye rossijskie lozungi kak sverhtekst* [Modern Russian slogans as a supertext]. Abstract of a dissertation for a candidate of philological sciences / Abstract of Ph. D. thesis. Ekaterinburg, 1999, 18 p.
3. *Mikhalskaya A. K. Osnovy ritoriki: Mysl' i slovo: uchebnoe posobie dlya uchashchihsya 10-11 klassov* [Fundamentals of Rhetoric: Thought and Word: a teaching aid for students in grades 10-11]. Moscow, Prosveshchenie, 1996. 416 p.
4. *Palamarchuk N. A. Sposoby vyrazheniya agressii v tekstah Internet-kommentariev* [Methods of expressing aggression in the texts of Internet comments]. Actual problems of theoretical and applied linguistics. 2011. P. 19-28.
5. *Sheigal E. I. Semiotika politicheskogo diskursa* [Semiotics of political discourse]. Volgograd, Peremena, 2000. 367 p.
6. *Buss A. H. The Psychology of Aggression.* Michigan: Wiley, 1961. 307 p.
7. *Zhelvis V. I. Pole brani. Skvernoslovie kak social'naya problema v yazykah i kul'turah mira* [Field of battle. Profanity as a social problem in the languages and cultures of the world]. Moscow, Ladomir, 1997. 453 p.
8. *Rubtsova Yu. V. Metody i algoritmy postroeniya informacionnyh sistem dlya klassifikacii tekstov social'nyh setej po tonal'nosti* [Methods and algorithms for constructing information systems for classifying social network texts by tonality]. Abstract of Ph. D. thesis. Novosibirsk, 2019. 141 p.
9. *Kapitanov A. I. Razrabotka metodiki i algoritmov linejnoj klassifikacii nestrukturirovannoj tekstovoj informacii v tekhnicheskih sistemah* [Development of methods and algorithms for linear classification of unstructured text information in technical systems]. Abstract of Ph. D. thesis. Moscow, 2022. 122 p.
10. *Kotelnikov E. V. Metodologiya intellektual'nogo analiza mnenij pri obrabotke tekstovoj informacii na osnove pravdopodobnogo vyvoda* [Methodology of intelligent analysis of opinions in processing text information based on plausible inference]. Abstract of Ph. D. thesis. Nizhny Novgorod, 2019. 39 p.
11. *Kizhaeva N. A. Issledovanie patternov v tekstah na osnove dinamicheskih modelej* [Study of patterns in texts based on dynamic models]. Abstract of Ph. D. thesis. St. Petersburg, 2018. 86 p.
12. *Cherkasova M. N. Rechevye formy agressii v testah SMI* [Speech forms of aggression in media tests]. Rostov-on-Don, Rostov State University of Railway Engineering, 2011. 123 p.
13. Scikit-learn [Electronic resource]. avaliable at: `https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes` (date of access: 09/30/2024).



**Voronina Irina Evgenievna**

Doctor of Engineering Sciences, Professor of the Department of Software and Information Systems Administration, Voronezh State University (VSU, 394018, Russia, Voronezh, University Square, 1), +7 (473) 220-82-66, e-mail: irina.voronina@gmail.com.

**Pastrevich Marina Konstantinovna**

lecturer, Department of Software and Information Systems Administration, Voronezh State University (VSU, 394018, Russia, Voronezh, University Square, 1), +7 (473) 220-82-66, e-mail: mirstat@mail.ru