Санкт-Петербургский государственный университет

БАРХАТОВА Валерия Андреевна

Выпускная квалификационная работа

Автоматическое выявление агрессии в текстах сетевых сообществ

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805. «Компьютерная и прикладная лингвистика»
Профиль «Компьютерная лингвистика»

Научный руководитель:
доцент, Кафедра математической лингвистики,
Хохлова Мария Владимировна
Рецензент:
доцент, ФГБОУВО «Высшая школа экономики»,
Шульгинов Валерий Александрович

Санкт-Петербург
2021

Оглавление

Введение

В современном обществе во многих сферах социальной жизни наблюдаются сложности в коммуникативном поведении людей, в результате которых речевое общение носит грубый и недоброжелательный характер. Можно заметить, что «высокая степень категоричности авторов публикаций в современной прессе создает впечатление о недопустимом уровне агрессивности речевой коммуникации в современном русском обществе…»[1], поэтому изучение речевой агрессии вызывает в последнее время немалый теоретический и практический интерес: появляется все больше научных работ, освещающих данную проблему. Решение вопросов, связанных с речевой агрессией, в частности, ее устранением во многом связывают с пропагандой толерантности в речевой коммуникации.

В последнее время речевая агрессия перешла из реального общения в виртуальное. Столкнуться с агрессией в интернете сегодня можно намного чаще, чем в офлайн-жизни. Это связано с тем, что интернет умеет хранить, все то, что, когда-то там было оставлено, а многочисленные научные работы, освещающие проблему киберагрессии, утверждают, что проявлять агрессию в интернет-сообществах намного проще, чем в реальной жизни из-за наличия анонимности в интернете. Кроме того, киберагрессия несмотря на то, что появилась она совсем недавно, способна наносить огромный вред психологическому здоровью той части общества, которая является активными пользователями сети интернет, в частности, социальными сетями.

Мы считаем, что существует способ для предотвращения распространения речевой агрессии в социальных сетях путем создания автоматического классификатора, обученного на данных, состоящих из агрессивных и неагрессивных комментариев.

[1] Карпенко Л. А. Психология. Словарь. М., 1990. 494 с.

Таким образом, основной **целью** нашего исследования является создание автоматического классификатора для выявления агрессии в текстах сетевых сообществ.

Поставленная цель потребовала решения следующих **задач**:

- изучение подходов к определению речевой агрессии;
- анализ существующих описаний речевой агрессии в работах других исследователей;
- анализ полученных результатов с точки зрения цели и гипотез исследования;
- представление классификации речевой агрессии;
- определение лингвистических средств выражения речевой агрессии;
- изучение проблемы киберагарессии, ее видов, причин и поводов;
- анализ исследований, посвященных сверточным нейронным сетям и определение их эффективности для классификации текстовых данных;
- сбор текстовых данных для оформления датасета;
- предварительная обработка собранных текстовых данных;
- обучение классификатора с использованием сверточной нейронной сети;
- тестирование классификатора и оценка полученных результатов обучения.

В качестве **объекта** данного исследования выступает речевая агрессия в русском языке в Интернете. **Предметом** исследования являются признаки агрессии.

Основным материалом исследования послужил созданный датасет, состоящий из агрессивных и неагрессивных комментариев социальной сети ВКонтакте.

Гипотеза данного исследования нами сформулирована следующим образом: речевую агрессию в текстах русскоязычных сетевых сообществ можно выявить автоматически с помощью обученного классификатора, созданного на основе сверточных нейронных сетей.

Отсутствие на данный момент датасета для анализа речевой агрессии и высокая опасность киберагрессии и, как следствие, необходимость ее идентификации обусловливают **актуальность** данной работы.

Разработанный классификатор речевой агрессии, направленный на ее выявление в текстах русскоязычных сетевых сообществ, является первым в своем роде для русского языка, в чем и заключается **новизна** проведенного исследования.

Теоретическая значимость нашего исследования связана с тем, что нами подробно изучен феномен речевой агрессии, проанализированы работы разных исследователей, определены лингвистические особенности, а также проблема киберагрессии и причины для борьбы с ней.

Данная работа характеризуется высокой **практической значимостью**. В ходе ее создания был разработан классификатор на основе сверточных нейронных сетей, способный выявить речевую агрессию в текстах сетевых сообществ.

Работа состоит из введения, основной части, разделенной на две главы, заключения, библиографического списка и приложений. Во **введении** приведены краткий обзор области, которой посвящена работа, и квалификационные параметры исследования. **Глава I** содержит теоретический обзор существующих описаний речевой агрессии, а также ее видов и лингвистических средств выражения. Кроме того, в главе описана сущность проблемы киберагрессии. В **Главе II** представлены особенности сверточных нейронных сетей, их архитектура, этапы и результаты сбора датасета и создания классификатора для выявления речевой агрессии в тестах сетевых сообществ, а также приведен анализ результатов нашего эксперимента. В **заключении** подводятся итоги исследования и приводятся потенциальные направления его продолжения и расширения.

Глава I. Агрессия в интернете и ее лингвистические особенности

1. Речевая агрессия в лингвистике

1.1. Понятие речевой агрессии в лингвистике

Агрессия давно и плодотворно исследуется как в зарубежной, так и в отечественной науке, но чаще всего рассматривается как феномен социальной психологии. Одной из актуальной для лингвистики задач становится поиск и создание таких моделей речевого взаимодействия, использование которых позволяет снизить количество ненависти и враждебности. Для решения этой задачи исследователям необходимо подробно изучать механизмы речевого воздействия, потому что даже любое слово, обращенное к другому человеку, является попыткой воздействия на него, поэтому для улучшения процесса коммуникации необходимо расширить понимание и знание законов воздействия.

По мнению К.Ф. Седова, интерес лингвистики к изучению речевой агрессии «вызван работами по изучению политического дискурса»[2]. Это связано с тем, что «присутствие агрессивных речевых проявлений в публичной официальной речи политиков невольно заставляет задуматься об аналогичных проявлениях в иных коммуникативных сферах, а неолингвистика все больше обращает внимание на агрессию в повседневной коммуникации»[3]. Таким образом, повседневная коммуникация оказывается наиболее ярким пространством для возникновения диссонансного взаимодействия людей.

Обычно речевая агрессия возникает в столкновениях между участниками коммуникативного действия. И.Н. Горелов и К.Ф. Седов называют такую ситуацию «коммуникативным конфликтом, в основе которого находится агрессия, выраженная с помощью языковых средств»[4]. Для коммуникативного конфликта

[2] Седов К.Ф. Агрессия как вид речевого воздействия // Прямая и непрямая коммуникация. Саратов: «Колледж», 2003. С. 112.
[3] Там же.
[4] Горелов И.Н., Седов К.Ф. Основы психолингвистики. М., 2001. 159 с.

характерно наличие хотя бы у одного участника коммуникации стремления снять психологическое напряжение за счет собеседника. Но это стремление не возникает на пустом месте, ему предшествует другой феномен – «чувство фрустрации, т. е. психологический дискомфорт, возникающий при невозможности добиться какой-либо цели»[5]. В межличностном взаимодействии фрустрация может возникнуть только в том случае, если коммуникативный партнер нарушает нормы поведения (например, невымытая посуда, опоздание на работу, экстравагантный внешний вид), т. е. все, что выходит за рамки нормы, может спровоцировать коммуникативный конфликт и, следовательно, речевую агрессию.

Коммуникативный конфликт может происходить в любой сфере нашей жизни. Основным триггером его возникновения, как мы сказали ранее, является нарушение нормы поведения. В своей работе мы собираем и анализируем текстовые данные интернет-пользователей, поэтому нам важно понимать, как коммуникационный конфликт может существовать именно в сетевых сообществах.

А.А. Тиллабаева и В.А. Шульгинов в своей работе подробно рассмотрели речевое поведение интернет-пользователей в ситуации конфронтационного общения[6]. Они выделили несколько ситуаций, которые чаще всего встречаются в описываемой ими ситуации: обсуждение достоверности предъявляемой информации, отсутствие установки пользователя на проверку фактов и нарушение правил орфографии и пунктуации. Мы видим, что в каждом случае происходит нарушение определенной нормы. В первом случае пользователи сталкиваются с такой информацией, которая противоречит всем известным нормам, что вызывает недоверие и провоцирует обсуждение. Во втором случае норме противоречит поведение другого пользователя, который по каким-то причинам отказывается проверять факты. В третьем случае проявляется реакция на нарушение норм

[5] Там же.

[6] Тиллабаева А.А., Шульгинов В.А. Речевое поведение интернет-пользователей в ситуации конфронтационного общения // Слово.ру: балтийский акцент. 2020. Т. 11. №4. С. 45-57.

орфографии и пунктуации. Кроме того, исследователи замечают, что не всегда такие комментарии пользователей могут привести к конфронтационному общению.

Наличие или отсутствие речевой агрессии в данном случае напрямую зависит от дискурсивных особенностей конкретного интернет-сообщества или определяется конфронтационный репликой, которая содержит в себе прямое оскорбление оппонентов, выраженное с помощью языковых средств. Можно сказать, что для возникновения коммуникативного конфликта и проявления речевой агрессии недостаточно только одного нарушения поведения. Необходимо также наличие других факторов.

Дальнейший анализ исследований и попытка определить речевую агрессию как феномена показывают отсутствие единого подхода к изучению речевой агрессии. Мы выяснили ранее, что она неразрывно связана с коммуникативным конфликтом и даже является его основой, поэтому сопоставим взгляды разных исследователей на речевую агрессию.

В общем виде «речевая (вербальная) агрессия – обидное общение; словесное выражение негативных эмоций, чувств или намерений в оскорбительной, грубой, неприемлемой в данной речевой ситуации форме»[7]. Она характерна как для устной, так и письменной речи, и может выражаться эксплицитно или имплицитно. Для выражения разного типа агрессии используются разные языковые средства.

В стилистическом словаре представлено следующее определение: «речевая агрессия – использование языковых средств для выражения неприязни, враждебности, манеры речи, оскорбляющая чье-либо самолюбие, достоинство»[8].

Другой исследователь, К.Ф. Седов, рассматривает речевую агрессию с лингвопсихологической точки и определяет ее как «целенаправленное коммуникативное действие ориентирование на то, чтобы вызывать негативное

[7] Щербинина Ю.В. Русский язык: Речевая агрессия и пути ее преодоления. М., 2012. 224 с.
[8] Стилистический энциклопедический словарь русского языка / под ред. М. Н. Кожиной. М., 2006. 696 с.

эмоционально-психологическое состояние у объекта речевого воздействия»[9]. Отечественный лингвист В.И. Жельвис в своих работах изучает речевую агрессию в социопсихологическом ракурсе как психологически оправданное действие[10]. Другие исследователи, например, А.К. Михальская[11] и Л.В. Енина[12] отмечают, что речевая агрессия может служить средством выплескивания эмоции и снятия эмоциональной напряженности.

В своих исследованиях Л.В. Енина определяет речевую агрессию как «сферу речевого поведения, которая мотивирована агрессивным состоянием говорящего»[13]. Однако «агрессивное состояние» как некий психологический феномен находит свое отражение в выборе языковых и речевых средств в том случае, если речевая агрессия выступает в эксплицитных формах. Имплицитные формы речевой агрессии, как правило не экспонируют психофизическое состояние адресанта[14].

С прагмалингвистической точки зрения речевая агрессия является способом речевого воздействия, представленное как однонаправленное эмоциональное негативное воздействие на участника общения.

Социолингвистический подход определяет речевую агрессию как широкое лингвистическое явление[15] и рассматривает ее в связи с современной ситуацией в обществе.

---

[9] Седов К.Ф. Агрессия как вид речевого воздействия // Прямая и непрямая коммуникация. Саратов: «Колледж», 2003. С. 112

[10] Жельвис В.И. Поле брани. Сквернословие как социальная проблема. М., 1997. 330 с.

[11] Михальская А. К. Русский Сократ: Лекции по сравнительно-исторической риторике. − М.: Изд. центр. «Academia», 1996. − 192с.

[12] Енина Л. В. «Речевая агрессия и речевая толерантность в средствах массовой информации // Российская пресса в поликультурном обществе% толерантность и мультикультурализм как ориентиры провеччионального поведения. М., 2002. С. 104-110.

[13] Енина Л.В. Катартический характер речевой агрессии в сверхтексте лозунгов и источники ее смягчения // Вопросы стилистики: Антропоцентрические исследования. Саратов, 1999. Вып.28. С. 105.

[14] Закоян Л.М. Речевая агрессия как предмет лингвистических научных исследований // Полилингвальность и транскультурные практики. 2008. №2. С. 46-52.

[15] Там же.

Юрислингвистика рассматривает понятие «оскорбление» как проявление речевой агрессии[16], потому что именно понятие «оскорбление» закреплено на законодательном уровне и включает в себя все виды речевой агрессии. Далее в работе мы подробно рассмотрим статус «оскорблений» в нашей стране и понаблюдаем, как он соотносится с речевой агрессией и чем отличается от киберагрессии.

Рассмотрев точки зрения разных исследователей, занимающихся изучением речевой агрессии, мы выяснили, что данный феномен имеет сложную и многогранную структуру, который позволяет рассматривать речевую агрессию не только как психологическое, но и как лингвистическое явление и как сложное коммуникативное действие, целью которого является создание негативного эмоционально-психологического состояния у объекта речевого воздействия.

В своей работе для исследования речевой агрессии мы остановимся только на лингвистическом подходе, т. е. будем подробно рассматривать языковые средства выражения речевой агрессии, чтобы в дальнейшем использовать полученную информацию для сбора и анализа текстовых материалов, которые будут использованы для обучения классификатора.

1.2. Виды речевой агрессии

В предыдущем параграфе мы выяснили, что существует несколько подходов к определению понятия речевой агрессии, поэтому опираясь на изученный ранее материал и работы К.Ф. Седова, который внес значительный вклад в изучении речевой агрессии как лингвистического феномена, определим классификацию видов речевой агрессии.

1. Вербальная – невербальная

[16] Кусов Г.В. Оскорбление как иллокутивный лингвокультурный концепт: Автореф. дис… канд. филол. наук. Волгоград, 2004. 27 с.

Основным отличием вербальной от невербальной агрессии является средство ее выражения. Так, для проявления вербальной характерно использование всех возможных языковых средств, а для невербальной – жестов, эмоций, мимики и др. Исследователи отмечают, что каждая культура имеет свой набор вербальных и невербальных форм проявления агрессии[17].

2. Прямая – косвенная

Прямая агрессия, в отличии от косвенной, содержит открытую, явную враждебность, которая может проявляться в таких формах, как угроза, оскорбление и другие:

- *фу нахуй мерзотварь гавно уебищно пиздофемка;*
- *Мда. Ты типичная срайлиш.*
- *сам иди туда, тупорылый.*

Одним из самых распространенных способов косвенной агрессии является ирония, в основе которой находится насмешка:

- *Вообще ничего не успеваю делать!*
- *Конечно, ты у нас вся такая занятая и деловая.*

3. Инструментальная – неинструментальная

Для неинструментальной агрессии характерно использование агрессии ради агрессии, где один участник коммуникации пытается за счет другого осуществить катартическую разрядку:

- *Иди ты нахер чмошник обосанный!*
- *Сам ты пиндос ебанный…*

Для инструментальной агрессии характерно стремление к достижению какой-либо цели:

- *чудище, не заткнешься сейчас, рот гавном вымажу.*

[17] Стернин И.А. Введение в речевое воздействие. Воронеж, 2001. 227 с.

В данном случае мы видим угрозу, которая содержит результат невыполнения действия коммуниканта.

4. Инициативная – реактивная

Реактивная агрессия чаще всего проявляется в качестве защиты от агрессора, когда в ответ на агрессию адресант сам проявляет агрессию:

– *ну ты и тряпка!*

– *сам ты тряпка!*

Кроме того, выделяют также разновидность реактивной агрессии – коммуникативный саботаж, который обычно проявляется в блокировании вопросно-ответной коммуникации методом ответа вопросом на вопрос[18]:

– *Ты не хочешь как-то извиниться перед Кириллом?*

– *А почему я должен перед кем-то извиняться?*

Важно понимать различие между пассивной агрессии и реактивной. Ключевым моментом является то, что реактивная – это результат влияния агрессии на агрессию, а пассивная – это использование метода прекращения контакта.

5. Спонтанная – подготовленная

Важным отличием в этой бинарной классификации является особенности порождения высказывания. Спонтанная агрессия проявляется как мгновенная реакция на какое-либо действие:

Ребенок уронил вазу. Его мама говорит:

– *Что ты сделал, идиотина!*

Подготовленная агрессия характеризуется «спланированностью»: человек заранее обдумывает какую-то ситуацию (особенно это часто бывает во время бессонной ночи перед коммуникативным актом), а некоторые даже могут иметь

[18] Николаева Т.М. О принципе «некооперации» и/или о категории социолингвистического воздействия // Логический анализ языка: Противоречивость и аномальность текста. М., 1990.

«заготовки», т. е. запланированные фразы, если, например, плотно работают или общаются со своим недоброжелателем и активно думает о нем каждый день:

Мужчина в разговоре со своим коллегой негативно отзывается о начальнике, сыне главного директора:

– *Я всегда знал, что он ни на что не способен, мамкино отродье.*

6. Непосредственная – опосредованная

Эта оппозиция выделяется на основе характера коммуникативного контакта. Непосредственная речевая агрессия возникает в одном пространстве и времени. К ним можно отнести любое коммуникативное столкновение, которое происходит в реальном времени и закреплено за одним местом (например, конфликт учителя и ученика на уроке географии 25 марта). Опосредованная же может осуществляться в разных хронотопах. Этот вид агрессии характерен для сплетен, обсуждений и осуждений, когда агрессия направлена на человека, который в момент проявления агрессии находится в совершенно другом месте.

7. Эмоциональная – рациональная

Основным отличием этих оппозиций является наличие/отсутствие в речевом действии рационального начала. Эмоциональная агрессия характеризуется спонтанностью, в отличие от рациональной, которая чаще всего имеет спланированное выступление, например, в шутках. Основным же маркером рациональной агрессии является стремление говорящего на осознанном уровне учитывать при достижении перлокутивного эффекта особенностей коммуникативной ситуации и личностных свойств адресата речи[19].

8. Сильная – слабая

Сильная агрессия характерна для речевого акта, который способен повлиять на изменение эмоционального состояния коммуникативного партнера, например, вызвать у него эмоции страха, гнева, унижения и др. Особенно интересно это можно

[19] Седов К.Ф. Агрессия как вид речевого воздействия // Прямая и непрямая коммуникация. Саратов: «Колледж», 2003. С. 110-113.

заметить при пассивной агрессии, когда целью агрессора является вызвать у другого эмоциональное чувство, реакцию, например, стыд, ущербность, никчемность и т.д., при этом не всегда используется средства прямой агрессии, т.е. агрессор может иметь нейтральную интонацию, эмоционально никак не окрашенную, но как раз сама специфика коммуникативной ситуации позволит второму участнику беседы испытывать неприятные эмоции.

Разделение речевой агрессии на сильную и слабую, по мнению К.Ф. Седова, приводит к выводу, что необходимо также при изучении агрессии осознавать неизбежность существования этого речевого феномена в нашей жизни, а где-то, по мнению исследователя, она даже необходима[20].

Это связано с тем, что все люди совершенно разные, отличаются характером, мировоззрением, жизненными принципами. Все это в последствии становится причиной возникновения конфликтов. Исследователи отмечают, что не все конфликты являются проявлением враждебности. Столкновение мнений наоборот, может порождать современно новые явления, решения проблем. К.Ф. Седов убежден, что в небольших дозах агрессия не ухудшает, а улучшает коммуникативный климат.

Если речевую агрессию использовать рационально, например, в виде шутки, розыгрыша, то она даже способна, например, укрепить отношения в семье, коллективе, сплотить друзей, и стать противоядием для лицемерия, ханжества, и приторности. Таким образом, К.Ф. Седов выделяет последнюю бинарную оппозицию речевой агрессии.

9. Враждебная – невраждебная

Именно враждебная речевая агрессия является ядерной формой изучаемого нами феномена. Ее примерами могут служить все ранее представленные агрессивные высказывания.

[20] Там же.

Невраждебная агрессия встречается чаще всего в неофициальном дружеском общении, в частности, в мужской компании. Ее основными признаками являются, например, шутливые толчки, удары, которыми пользуются здоровые и жизнерадостные молодые люди. Автор считает, что именно невраждебная речевая агрессия является признаком психологического здоровья и нормальных дружеских отношений.

М.М. Бахтин своем исследовании выделяет карнавальный тип взаимоотношения, который может привести к «освобождению от господствующей правды существующего строя, временную отмену всех иерархических отношений, привилегий, норм и запретов», и реализуется в «особой форме вольного фамильярного контакта между людьми»[21], где «создается особый идеально-реальный тип общения, невозможный в обычной жизни. Это вольный фамильярно-площадной контакт между людьми, не знающий никаких дистанций между ними»[22].

Именно для такого типа дружеского взаимоотношения характерно использование невраждебной агрессии. В таком общении можно найти как бы агрессивные речевые выступления, конечная цель которых не сопоставима с целью враждебной речевой агрессии (намерение обидеть или оскорбить собеседника).

Разговор двух подруг:

– *Ну ты балда! Может хватит уже за ним бегать? Это ни к чему хорошему точно не приведет.*

Таким образом, выделенные нами бинарные оппозиции можно представить в виде поля с ядром, в основе которого будут располагаться наиболее очевидные формы речевой агрессии. К таким можно отнести прямую, неинструментальную, инициативную, активную, спонтанную, непосредственную, эмоциональную, сильную и враждебную речевую агрессию. На периферии будут относиться виды

[21] Бахтин М.М. Франсуа Рабле и народная культура средневековья и Ренессанса. М., 1990.
[22] Там же. С. 21-22.

агрессии с ослабленной иллокуцией. Например, непрямая, инструментальная, реактивная, пассивная, опосредованная, рациональная, слабая, невраждебная, подготовленная.

В итоге мы рассмотрели, на наш взгляд, наиболее полную классификацию речевой агрессии. Таким образом, мы еще раз подтвердили сложность структуры речевой агрессии как феномена языка.

2. Лингвистические средства выражения агрессии в русском языке

В речи агрессия может проявляться на разных языковых уровнях (лексическом, грамматическом и др.). В нашем исследовании ставится задача показать, какие лингвистические средства используются для выражения агрессии в русском языке. Лингвистический подход к изучению речевой агрессии предполагает ее исследование не только как феномена языка, но и изучение способов ее выражения, то есть «использование языковых средств для выражения неприязни, враждебности, манера речи, оскорбляющая чье-либо самолюбие»[23].

Основной целью нашей работы является создание автоматического классификатора, который мог бы выявлять признаки агрессии в текстах сетевых сообществ. Для обучения такого классификатора нам необходимо собрать большое количество текстов, содержащих агрессию. Чтобы найти такие сообщения, нам необходимо понимать, что такое речевая агрессия в целом и какие лингвистические средства выражения используются. В первом параграфе работы мы подробно рассмотрели речевую агрессию как лингвистический феномен, изучили ее с точки зрения разных наук и отраслей, определили основные особенности и отличия от других языковых явлений. В этом параграфе мы выделим основные лингвистические маркеры речевой агрессии и рассмотрим их на примерах. Эта информация нам также необходима для того, чтобы при поиске комментариев с речевой агрессией мы могли отделить их от тех комментариев, где ее нет. Такая

[23] Стилистический энциклопедический словарь русского языка / под ред. М. Н. Кожиной. М., 2006. 696 с.

работа повысит точность и репрезентативность наших текстовых данных и улучшит результаты обучения модели.

Исследователи выделяют два основных способа выражения агрессии: эксплицитный и имплицитный. Ю.В. Щербинина утверждает, что эксплицитная речевая агрессия представляет собой открытую форму агрессии, «план содержания которой соответствует плану выражения в пределах одного агрессивного высказывания»[24], а в качестве основных языковых средств В.И. Жельвис относит «стилистически маркированные языковые средства (экспрессивно-окрашенную лексику, инвективу, грубо-просторечные слова и словосочетания, жаргонную лексику и т.д.»[25], а также те виды высказываний, которые содержат прямую агрессию: угроза, оскорбление, порицание и др.

Имплицитная агрессия, наоборот, предполагает скрытое, неявное выражение агрессии. Такой тип агрессии выражается с помощью следующих средств: намеки («диффамация»), «искажение фактов», неправомерное выражение, ироничное замечание, скрытая угроза, как утверждает Е.И. Шейгал[26], а также «пассивные формы коммуникации: переживание, лишение слова, запрет на коммуникативные действия, прекращение контакта с собеседником или нежелание в него вступать», как утверждает К.Ф. Седов[27]. К имплицитным проявлениям агрессии также можно добавить иронию и сарказм.

2.1. Эксплицитный способ выражения агрессии

В своей работе Ю.В. Щербинина выделяет двенадцать основных форм проявлений языковой агрессии. Каждая форма имеет свой набор лингвистических средств для выражения агрессии. Например, оскорбление является одной из самых

[24] Щербинина Ю. В. Вербальная агрессия в школьной речевой среде// Дисс. … канд. филол. наук. М., 2001.
[25] Жельвис В.И. Поле брани. Сквернословие как социальная проблема в языках и культурах мира. М.:Ладомир, 1997. 330 с.
[26] Шейгал Е.И. Вербальная агрессия в политическом дискурсе // Вопросы стилистики: Антропоцентрические исследования. Саратов: Изд-во Сарат. ун-та, 1999. Вып. 28. С. 204-222.
[27] Седов К.Ф. Речевая агрессия в межличностном взаимодействии // Прямая и непрямая коммуникация. Саратов: Изд-во ГУНЦ «Колледж», 2003. С. 196-212.

распространенных форм проявлений речевой агрессии. Исследователь утверждает, что структура такого вида агрессии представляет собой использование местоимения (которое также может опускаться) в сочетании с эмоционально-оценочными словами, содержащими негативную семантику. Например, существительные с прямым отрицательным значением (*идиот, дурак, сволочь, дрянь, подонок и др.*) или с переносным отрицательным значением (*валенок, дубина, собака, псина, пень и др.*).

Для усиления агрессии к существительному могут добавляться прилагательные, образующие устойчивые словосочетания (*паршивая овца, змея подколодная и др.*). Еще одним языковым средством оскорблений является употребление оскорбительных восклицаний (*Вот идиот! Ну и тварь!*). В таких восклицаниях употребляются восклицательные местоимения и усилительные частицы (*какой, вот, что за, ну, же, да и др.*). Средством выражения оскорблений при наличии определенного контекста может быть неуместное использование обращений на *ты/Вы* или обидного прозвища (*жирдяй, очкарик и др.*).

Еще одной формой выражения речевой агрессии Ю.В. Щербинина называет враждебное замечание, целью которого является «изменение самочувствия адресата путем высказывания негативных суждений прежде всего о его личности в целом, а также о поступках, качествах и прочих манифестациях». Такое высказывание, в отличие от других, употребляется в форме целого предложения: *Ты меня бесишь*; *Ты мне не указ*. В ядре такого высказывания всегда стоит местоимение ты/Вы.

Угроза также является одной из форм языковой агрессии. Ее структура включает следующие элементы: адресант (угрожающий), адресат (тот, кому угрожают), содержание угрозы (действия, которые будут произведены при невыполнении адресатом определенных действий), возможный результат, средство достижения угрозы, которое может совпадать/не совпадать с действиями самого адресанта. Восклицательные и побудительные предложения часто используются в

качестве языковых средств угрозы (*Если ты не закроешь группу, я найду и убью тебя!*). Используются также императивы, категоричные побуждения (*Удали группу или я посажу тебя*). Констатация факта может использоваться как угроза (*Ты скоро сдохнешь*). Риторический вопрос также может служить способом выражения угрозы (*Хочешь, чтобы я голову тебе свернул?*).

Исследователь выделяет также такую форму агрессии как порицание, при чем в грубой форме. Для нее характерны восклицательные и вопросительные конструкции (*Ты мне нагрубил!*), использование усилительных частиц же и ведь (*Ты ведь не убрала после себя*). В молодежной среде часто используется сниженная речь (сленг, просторечия), например *офигел, охуел, заебал, достал, обнаглел и др.*

Насмешка также часто встречается среди речевой агрессии. Она может содержать как имплицитную, так и эксплицитную языковую агрессию. В первом случае она выражается с помощью иронии, во втором – с помощью сарказма. В качестве языковых средств для формирования насмешки используются слова с уменьшительно-ласкательным суффиксом в уничижительном значении (*дорогуша, заинька и др.*), а также отрицательное переосмысление положительно-оценочных слов (*Дорогая моя, что же ты натворила?*).

Среди других форм выражения речевой агрессии выделяют также грубое требование, в основе которого содержится желание оказать сильное воздействие на собеседника или избавиться от него, мотивируя его совершать какие-либо действия в интересах говорящего. Для данной формы характерно использование местоимений и *наречий никогда, нигде, никто, никак* и др. Для подчеркивания грубости высказывания можно встретить употребление таких слов, которые семантически содержат грубый отказ: *ну конечно, прямо сейчас, стану я и др.*

2.2. Имплицитный способ выражения агрессии

Стоит отметить, что часто речевая агрессия может выражаться имплицитно и без использования языковых единиц с негативным стилистическим оттенком. В

таком случае в качестве маркера могут выступать грамматические средства языка, которых гораздо меньше лексических. Исследователи это связывают с тем, что закрепление грамматических норм в языке происходит медленнее лексических. К основным грамматическим средствам речевой агрессии можно отнести в первую очередь императивные конструкции *(поговори мне еще тут)*. Причем функция императива в таких фразах не приглашение совершить некоторое действие, а, по мнению В. Ю. Апресян, угроза. Говорящий угрожает, что если не будут выполнены все его условия, то он от вербальной агрессии перейдет к агрессивным действиям.

Вопросные конструкции, где адресат выражает свое недовольство невнимательностью, несоответствующим поведением, плохой сообразительностью (*ты понимаешь, что делаешь? хули ты пропагандируешь пидорасов? ты сука ебнутая, ты че хуйню несешь???)* также могут быть средством выражения имплицитной агрессии. Такие вопросы, как правило, используются в ограниченном количестве ситуаций. Например, для того чтобы упрекнуть своего собеседника в чем-то. При этом в таких речевых конструкциях явно прослеживается, что адресант не просто высказывается свое недовольство, но и заставляет адресата чувствовать себя униженным.

Лексические средства выражения имплицитной агрессии представлены гораздо шире в русском языке, чем грамматические. Среди основных лексических средств выражения речевой агрессии можно выделить использование прагматически окрашенных синонимов вместо существующих реальных, при чем в переносном смысле. Например, русский глагол движения *пихать* в агрессивной ситуации можно использовать так: *куда ты запихнул куртку? Я твои грязные руки запихну в жопу.* Из этих примеров видно, что сам глагол не несет в себе негативной коннотации (поэтому мы его относим к имплицитным средствам выражения агрессии), но оказавшись в другом контексте, где объектом этого действия оказывается лексема с эксплицитно языковой агрессией (*жопа, пизда, ебло и т. д.),* то выражение сразу приобретает все признаки речевой агрессии.

В своей работе М.Я. Гловинская называет такой способ выражения агрессии гиперболой[28]. Она отмечает, что глаголы, которые в первичном значении нейтральны, предполагают, что действие, которое они обозначают, требует больших физических усилий. Подобные глаголы в своих «переносных» значениях имеют двухслойную семантическую структуру: в ассерции у них находится негативная оценка действия, иными словами – агрессия, а в пресуппозиции – его нейтральное описание. Кроме того, под отрицанием агрессия исчезает, а указанное действие остается: *я не ору, а кричу/громко разговариваю*.

Например, *тащить, запихивать, тянуть* используются, когда речь идет о тяжелых объектах, которые нужно перемещать либо по плоскости, либо в неподходящие по размеру емкости. Но когда затрачиваемое усилие на самом деле не выходит за границы нормы, то подобные глаголы используются уже не буквально, а гиперболически (*засунь свой язык в пизду и не вякай*) и приобретают скрытую агрессию.

Имплицитная агрессия также может выражаться с помощью частиц, утверждает В. Ю. Апресян. Они используются в сочетании с такими синтаксическими конструкциями, главной целью которых является выражение непрямой агрессии. Их мы подробно рассматривали ранее в нашей работе: псевдо-императив и вопрос.

1. Использование частицы где. Иногда в сочетании с частицей *уж*. *Ср. Где уж тебе знать о морали человеческой педерастка; где уж мне тебя понять.* В данных примерах адресант выражает иронию по отношению к адресату и скрыто его обвиняет в неоправданно высокой самооценке.

2. Использование частиц *тоже мне* и *еще*. *Ср. Тоже мне самая умная нашлась тут; тоже мне мать называется; тоже мне защитница геев нашлась.* В этих примерах адресант скептичен по отношению к адресату или третьему лицу,

[28] Гловинская М.Я. Гипербола как проявление речевой агрессии // Сокровенные смыслы : сборник статей в честь Н. Д. Арутюновой. М., 2004. С. 69.

считает их несоответствующими тем требованиям, которым он должен соответствовать, по мнению говорящего.

3. Использование частиц *эх* в сочетании с местоимением *ты/вы*. *Ср. Эх ты, а еще друг называется. Говорящий в данном случае выражает мягкий упрек.*

4. Использование сочетание частиц *как же*. *Ср. Я тебе помогу. – Поможешь, как же. Здесь говорящий сомневается в действиях адресата.*

5. Использование частицы *да*. *Ср. Да вы все наркоманы что ли; Лена да дебилка ты.* Здесь адресант раздражен тем, что адресат не знает того, что кажется говорящему очевидным.

6. Использование частиц *только/тебе* в сочетании с псевдо-императивом или личной формой глагола. *Ср. Только попробуй не ответить мне. Мало не покажется; ты только попробуй сказать мне что-то*. Говорящий предупреждает адресата о том. Что ему не следовало бы делать, ведь иначе последует какое-либо наказание.

В нашей работе мы рассмотрели эксплицитный и имплицитный способы выражения речевой агрессии в русском языке. Мы заметили, что для каждого способа характерны свои средства выражения агрессии. Например, для эксплицитной агрессии характерно использование таких слов, которые содержат агрессию в своем основном лексическом значении (*Ср. Дура - неумная, глупая женщина.* Основное и единственное значение этой лексемы именно бранное, агрессивное). Для выражения скрытой агрессии используются, наоборот, те слова, которые в основном своем значении нейтральны, но попадая в определенный контекст или в сочетании со словами, содержащими эксплицитную агрессию, приобретают негативную коннотации (*Ср. Я постараюсь запихнуть в свой чемодан максимум вещей – я тебе твои корявые руки в задницу запихну*). Мы также выяснили, что частицы могут придавать агрессивное значение выражениям.

Таким образом, в русском языке можно выделить два основных способа выражения агрессии: имплицитный и эксплицитный. Главное их различие – это то,

что эксплицитная агрессия выражается при помощи основных средств языка, а имплицитная, наоборот, неосновных. Сегодня исследователи также продолжают работу над изучением речевой агрессии как феномена языка.

3. Статус оскорблений и проблема киберагрессии

Задачей нашей работы является создание автоматического анализатора, который позволял бы выделить среди комментариев социальных сетей те, в которых есть речевая агрессия. Эта тема сейчас наиболее актуальна, так как социальные сети появились в нашей жизни не так давно и еще не сформировалась общая культура поведения в сети интернет.

Сегодня в социальных сетях очень часто мы можем встретить оскорбления, но, если в офлайн-жизни они регулируются ч.1 ст.5.61 КоАП РФ и предполагает административное нарушение, то в сети интернет не все однозначно. Только в 2020 году были приняты поправки к этой статье, которые регулируют оскорбления в интернете. Так, например, оскорбления в мессенджерах считаются личными, а в социальных сетях – публичными, соответственно и наказания у них разные. Долгое время виртуальную агрессию (или киберагрресию) не принимали как потенциально опасную, но это совсем не так.

Прежде чем перейти к рассуждениям о том, почему виртуальная агрессия страшнее реальной, определим, что такое киберагрессия.

Данный термин был введен в употребление в 2007 году доктором философии Д. Шабро. Под киберагрессией понималась форма девиантного (отклоняющего) поведения в интернет-среде, которая может выражаться в оскорблениях, унижениях, издевательствах, разоблачениях, манипулировании, агрессивных нападках, преследованиях через коммуникативные технологии[29].

[29] Ярец А.Д. Разновидности конфликтов и агрессии в интернет-коммуникации // Идеи. Поиски. Решения: сборник статей и тезисов XIII Международной научно-практической конференции. Режим доступа : http://elib.bsu.by/handle/123456789/241180

Рассмотрим мнения зарубежных авторов: «киберагрессия – это умышленное совершение действий оскорбительного, унижающего или нежелательного характера (угрозы, преследование, домогательство, разглашение конфиденциальной информации) по отношению к лицу или группе, осуществляемое с помощью информационно-коммуникативных средств»[30]. Другое определение очень похоже, но в нем уточняются средства передачи такого вида агрессии в отношении жертвы: «киберагрессия – это действия, направленные на причинение вреда лицу или группе, выполняемые с помощью профессионального компьютера, мобильного телефона и других электронно-коммуникативных устройств, посредством электронной почты, социальной сети, мгновенных сообщений, блогов, игр в режиме онлайн»[31].

Отметим, что в ранних работах по данной проблематике использовался не термин «киберагрессия» (cyberaggression), а термин «кибербуллинг» (cyberbullying), означающий интернет-травлю в форме угроз, клеветы, шантажа. До сих пор некоторые авторы используют их как синонимы. На самом деле кибербуллинг является разновидностью киберагрессии[32]. Отечественные исследователи придерживаются аналогичной позиции[33].

Основными формами агрессии выделяют[34] **троллинг** (размещение в интернете провокационные сообщения для того, чтобы вызвать негативную реакцию или конфликт между участниками), **флейринг** (публичные оскорбления в

---

[30] Schoffstall C., Cohen R. Cyber-Aggression: The Relation between Online Offenders and Offline Social Competence // Social Development. 2011. Vol. 20, iss. 3. P. 586–604

[31] Willard N. E. Cyberbullying and Cyberthreats : Responding to the Challenge of Online Social Aggression, Threats, and Distress. Champaign, Illinois : Research Press, 2007. 320 p.

[32] Солдатова Г. У., Ярмина А. Н. Кибербуллинг : особенности, ролевая структура, детско-родительские отношения и стратегии совладания // Национальный психологический журнал. 2019. № 3 (35). С. 17–31.

[33] Черенков Д. А. Девиантное поведение в социальных сетях: причины, формы, следствие // Nauka-rastudent. ru. 2015. № 07. Режим доступа : https://readera.org/14330143, Шаров А. А. Специфика девиантной активности молодежи в интернет-среде // Учен. записки. Электронный научный журнал Курского государственного университета. 2019. № 3. С. 255–261.

[34] Киберугрозы, киберагрессия, кибербуллинг: различия в восприятии, оценке и поведении у разных групп населения Российской Федерации. Режим доступа : https://raec.ru/activity/analytics/9880/.

интернете между участниками в равных позициях, разжигание спора), **хейтинг** (негативные комментарии и сообщения, иррациональная критика в адрес конкретного человека, часто без обоснования своей позиции), **кибербуллинг** (агрессивные, умышленные, повторяющиеся и продолжительные во времени действия, совершаемые группой лиц или один лицом использованием электронных форм контакта в отношении жертвы, которой трудно защитить себя) и **кибersталкинг** (использование электронных средств для преследования жертвы через повторяющиеся сообщения, вызывающие тревогу и раздражения). Таким образом, основное отличие киберагрессии от простой агрессии – это интернет, который является идеальной средой для порождения оскорблений. Почему это происходит, рассмотрим далее.

От традиционных оскорблений люди могут скрыться, так как они, чаще всего, проявляются в каких-то ситуациях. Например, в автобусе человек не уступил пожилой женщине место, и она назвала его неблагодарной свиньей, или в громком споре со знакомым кто-то обозвал своего оппонента мразью. Ситуация прошла, но человек может просто больше не встречаться со своим обидчиком, чтобы не подвергать себя дальнейшим унижениям. Кроме того, может его проучить, подав соответствующую жалобу в суд. Киберагрессия может найти где угодно. Единственный метод защиты – это выкинуть телефон и удалиться из социальных сетей, но это неправильная логика.

Почему же в интернете люди чаще оскорбляют друг друга и подвергаются киберагрессии? По данным последних исследователей, почти половина подростков (49%) совершали агрессивные действия в интернете и примерно столько же (51%) сами становились жертвами[35]. Как отмечает С.И. Коданева[36], в киберпространстве люди чувствуют себя более раскованными в эмоциях, словах и поведениях,

[35] Calpbinici, Arslan, 2019 *Calpbinici P., Arslan F.T.* Virtual behaviors affecting adolescent mental health: The usage of Internet and mobile phone and cyberbullying // Journal of Child and Adolescent Psychiatric Nursing. 2019. Vol. 32, N 3. P. 139–148.

[36] Коданева С.И. Кибрбуллинг: причины явления и методы предупреждения // Социальные новации и социальные науки. М., 2020. №1. С. 149-159.

раскрывают свое истинное «я». В социальных сетях можно говорить и делать то, что сложно делать в физическом мире, включая издевательства и оскорбления. Обидчик как бы находится под «защитой» (не может получить мгновенную реакцию, находится на расстоянии).

В 2017 году ассоциация электронных коммуникаций подготовила большое исследование на тему киберагрессии. Они определили классификацию интернет-рисков, виды агрессии в интернете, характеристики буллинга и кибербуллинга, их особенности, поводы, мотивы, продемонстрировали опыт столкновения с разными видами онлайн-агрессии (разных возрастных групп) и др.

На рис. 1 показаны основные виды онлайн-агрессии. Мы видим, что подростки 14-17 лет чаще всего становятся свидетелями агрессивного онлайн-поведения (46%), а 44 % получали агрессивные сообщения в свой адрес, жертвами буллинга стали 48 %, 23% получали угрозы физической расправы, младшие подростки чаще всего становятся свидетелями или жертвами агрессивной коммуникации.

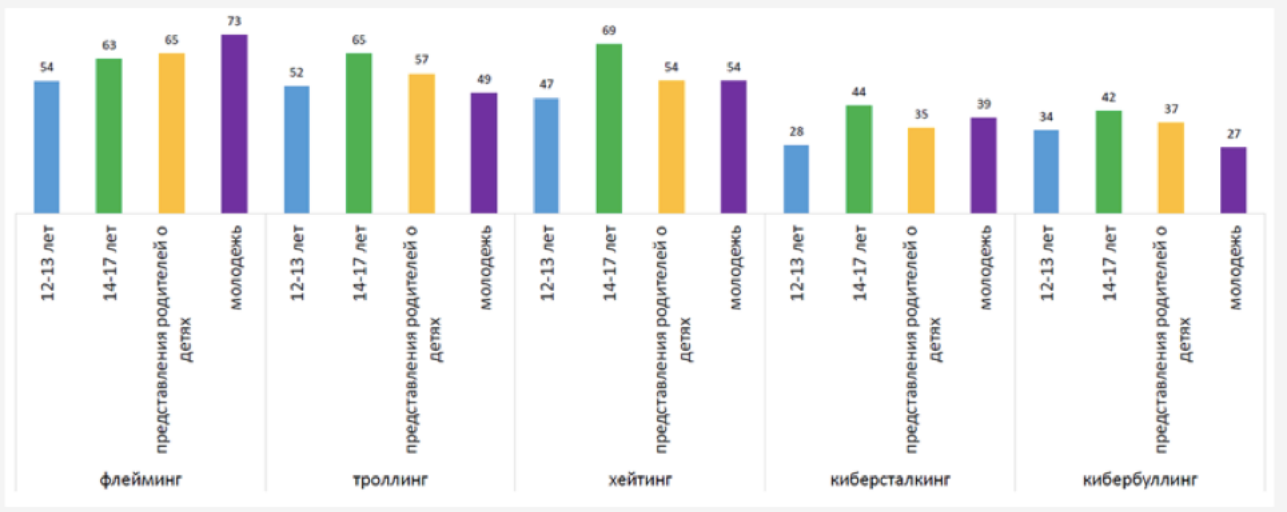

Рисунок 1. Виды онлайн-агрессии

Флейминг, троллинг и хейтинг – самые распространенные виды онлайн-агрессии, с которой старшие подростки встречаются чаще, чем все остальные.

Среди поводов для киберагрессии внешность и личностные особенности занимают лидирующие позиции во всех возрастных группах (рис. 2): например, у старших подростков внешность является поводом для киберагрессии в 64% случаев, личностные особенности – 60%, увлечения и хобби – 50%, сексуальная ориентация и национальная принадлежность – по 44 %.

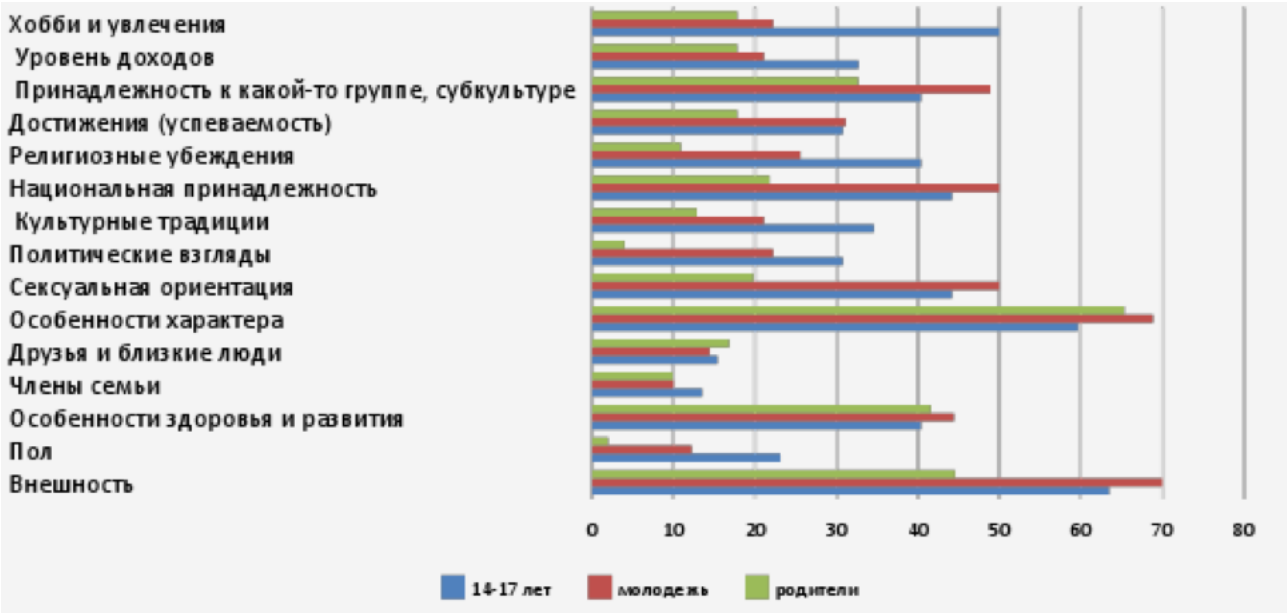

Рисунок 2. Основные поводы для киберагрессии

Среди основных мотивов, которые показаны на рис. 3, мы видим, что подростки предпочитают вести себя агрессивно чаще онлайн, чем офлайн, поскольку в онлайн-пространстве их привлекает безнаказанность (46 %), анонимность (33 %), простота и скорость (39 %). 33% подростков отметили, что выражение своего мнения в виртуальном пространстве менее болезненно, а 31%

отметили страх наблюдения реакций другого.

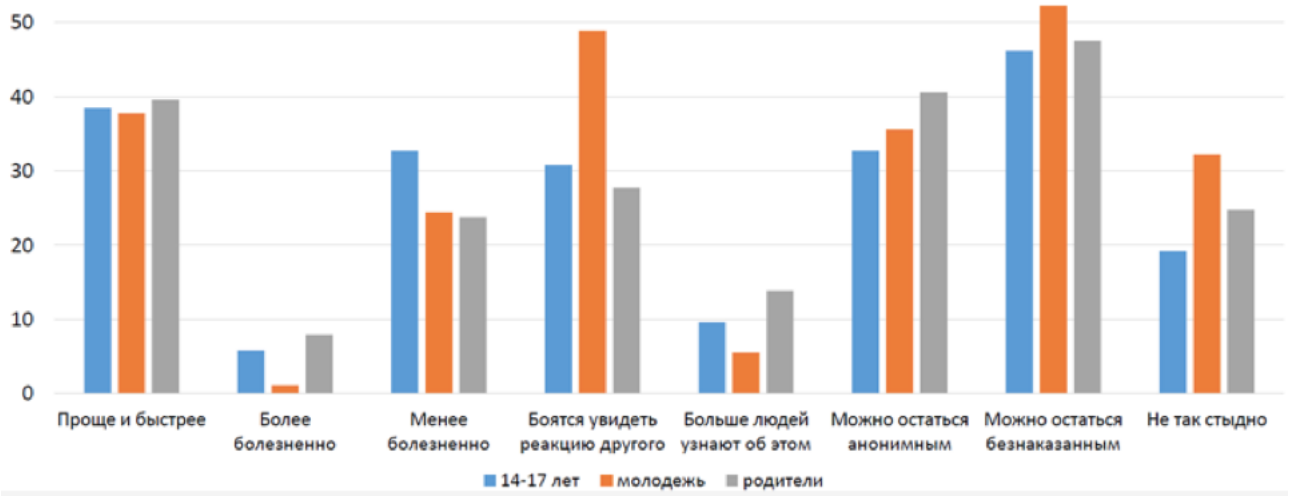

Рисунок 3. Мотивы кибераггрессии интернет-пользователей

В современном мире речевая агрессия осуждается (особенно это видно в условиях роста толерантности), а в своих экстремальных формах даже заслуживает наказания. Сегодня в некоторый странах за выражение ненависти в адрес людей, связанные с их особенностями, предусматривается наказание, оформленное на законодательном уровне, так как подобное поведение приравнивается к дискриминации.

Однако на пути борьбы с вербальной агрессией существует целый ряд трудностей. Например, невозможно составить универсальный для всех культур и языков список запрещенных слов, потому что то, что в одной культуре осуждается, то в другой возвышается или вообще имеет нейтралитет. Например, слово fuck в американской культуре является одним из самых распространенных ругательств. В русском языке оно имеет перевод как «черт возьми», но в то же время современная русская молодежь относится к слову fuck совсем не так, как американские сверстники. Для русскоязычного населения это слово вполне может присутствовать в обыденной речи.

В современной коммуникации можно также встретить такие случаи, когда в пределах одной этнической группы, но в разных социальных подгруппах можно

встретить совершенно полярное отношение к бранному слову. Например, то, что является для студентов приемлемым, для преподавателя считается недопустимым, для рыночной торговки вообще каждодневным и обыденным. К объяснению данной ситуации хорошо подойдет пример В.И. Жельвис: «в языке современного российского телевидения еще совсем недавно безоговорочно табуированные *жопа* и *говно* практически получили права гражданства»[37].

Существуют также примеры, когда в пределах одной подгруппы отношение к запрещенным словам может в течение времени меняться на противоположное. Особенно это стало проявляться с распространением толерантности в обществе. Например, сегодня слово *пидорас* уже не относится только к представителям ЛГБТ сообщества, как это было изначально. Конечно, ненависть и агрессия к секс-меньшинствам осталась, но в связи с распространением толерантности уровень ее с каждым годом снижается. Еще одним примером может служить тот факт, что в начале 2000-х было очень много шуток в адрес блондинок. Существовало мнение, что если женщина имеет светлый оттенок волос, то она не отличается особо умом и сообразительностью. Сегодня такое мнение очень сложно встретить (по крайней мере в открытых источниках), потому что феминистическое движение привело к тому, что подобное поведение и отношение к блондинкам приравнивалось к харассменту.

В обществе, где процветает свобода слова и отсутствует цензура, очень остро воспринимается запрет на использование определенной группы слов, поэтому необходимо искать другие способы борьбы с речевой агрессии. Например, ее предотвращение и развитие толерантности в обществе, потому что запрет на использование агрессии никак не изменит позицию агрессора, и даже может иметь обратный эффект. Поэтому просто запретить ему высказывать свое мнение не

[37] Жельвис В.И. Поле брани. Сквернословие как социальная проблема в языках и культурах мира. М.:Ладомир, 1997. 330 с.

является целесообразным и не снижает агрессивность общества. Поэтому действовать необходимо не «кнутом», а другими, более толерантными способами.

Таким образом, мы еще раз убедились в том, что бороться с киберагрессией нужно, как и с любой другой преступностью. Она доставляет жизненные неудобства очень большому количеству подростков и молодежи, так как именно они являются активными пользователями интернета. В то же время их не окрепшее психологическое здоровье не позволяет правильно реагировать на киберагрессию. Сегодня оскорбления в интернете наказуемы, но иногда лучше предупредить, чем бороться, именно поэтому главной целью нашей работы является создание автоматического классификатора для выявления агрессии в текстах сетевых сообществ.

4. Выводы к главе I

В этой главе мы подробно рассмотрели речевую агрессию как лингвистический феномен, проанализировав научные работы разных исследователей. Мы выяснили, что речевая агрессия – это сложное явление, для изучения которого необходимо применять разные подходы, в том числе разных научных отраслей. В этом мы опирались на работы К.Ф. Седова, В.Я. Апресян, Ю.В. Щербининой, В.И. Жельвис и др.

Нами была приведена наиболее полная классификация видов речевой агрессии, учитывающая около 10-ти бинарных оппозиций, к каждой из которых нами были подобраны собственные примеры. Полученная информация обязательно пригодится в практической части работы, так как мы будем не только проверять работоспособность классификатора, но и его реакцию на особые виды агрессии.

Мы также выяснили, что такое киберагрессия, чем она отличается от привычной нам речевой агрессии. Проанализировав социологическое исследование, посвященное изучению киберагрессии, мы определили, почему она

возникает, в каких контекстах ее можно встретить, для каких возрастных групп она характерна, а также определили, что проблема киберагрессии является актуальной на сегодняшний день. Все это определяет важность и актуальность нашей работы для научного сообщества. Следующий этап связан со сбором датасета, состоящего из агрессивных комментариев, поэтому важно понимать, из каких источников мы можем получить необходимый для обучения материл, что заметно облегчит его поиск и сбор. Кроме того, данное исследование еще раз подтвердило мысль о том, что киберагрессия является важной современной проблемой, требующей внимания и решения, так как она способна нанести непоправимый вред психическому здоровью интернет-пользователей и привести к суициду.

Глава II. Использование сверточной нейронной сети для выявления агрессии

1. Сверточные нейронные сети

Изначально сверточный нейронные сети (CNN) успешно зарекомендовали себя в процессах распознавания и классификации изображений. Они настолько хорошо справлялись с поставленной задачей, что исследователи по всему миру стали использовать их для классификации различных данных (в том числе и текстовых), так как они устроены подобно зрительной коре головного мозга, т. е. способны концентрироваться на небольшой области и выделять в ней важные особенности. На сегодняшний день принадлежность текста к какому-либо классу является актуальной задачей, поскольку данных становится все больше, и старые способы в некоторых случаях уже не справляются[38]. В основном этот метод обработки текстовых данных применяется для фильтрации документов, распознавания спама, классификации новостей и др.

Самым распространенным способом классификации является «способ на основе описаний объектов с использованием признаков, в котором каждый объект характеризуется набором числовых/нечисловых признаков»[39]. Важно отметить, что такой способ не является универсальным, потому что некоторые типы данных могут содержать «скрытые» признаки. Например, даже ребенок может определить, что находится перед ним: цветок или дерево, но у машины с такой задачей могут возникнуть трудности. Это происходит потому, что мы умеем определять «скрытые» признаки этих предметов: у дерева есть ствол, крона, а цветок состоит из стебля и соцветия, собранного из лепестков. Для таких сложных случаев исследователи используют глубокое обучение, в основе которого находится набор

[38] Воробьев Н.В., Пучков Е.В. Классификация текстов с помощью сверточных нейронных сетей // Молодой исследователь Дона.Ростов-на-Дону. 2017. №6. Режим доступа : https://mid-journal.ru/upload/iblock/8ed/1.-vorobev_-puchkov.pdf

[39] Ле Мань Ха. Свёрточная нейронная сеть для решения задачи классификации // Труды МФТИ. 2016. Том 8 № 3. С. 91-97

алгоритмов машинного обучения, пытающиеся моделировать высокоуровневые абстракции в данных, выделять как раз эти самые «скрытые признаки»[40].

Повышенное внимание лингвистов к нейронным сетям в целом обусловлено несколькими причинами. Во-первых, их использование для решения лингвистических задач повышает качество их решения, во-вторых, снижает трудоемкость при работе с текстовыми данными, и в-третьих, открывает возможность решать новые задачи. Например, создание чат-ботов, в основе которых лежат не набор правил и шаблонов, а набор размеченных реальных запросов клиентов, по которым бот обучается, запоминает и в дальнейшем без труда распознает похожие запросы. Это связано с тем, что, чем больше «обязанностей» у бота, тем больше правил для обработки запросов должен прописать лингвист, но не всегда сразу удается определить точные шаблоны, так как сложно предугадать, как именно будет общаться человек с ботом. Кроме того, написание правил – это огромный пласт работы, поэтому в таких случаях также могут прийти на помощь нейронные сети.

Несколько лет назад коллектив авторов из Intel и Carnegie-Mellon University в своей работе признали, что CNN справляются с поставленной задачей классификации текста намного лучше, чем RNN, которые использовались до этого на протяжении последних лет[41]. А в 2015 году Крис Маннинг в своей работе определил круг применимости нейронных сетей. К ним он отнес задачи классификации, последовательности и снижения размерности[42]. Таким образом, для решения нашей задачи мы приняли решение использовать именно CNN, так как они отлично зарекомендовали себя на практике.

---

[40] Bengio Y. Learning deep architectures for AI // Foundations and Trends in Machine Learning, 2009. Режим доступа : https://www.researchgate.net/publication/215991023_Learning_Deep_Architectures_for_AI

[41] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. Реим доступа : https://arxiv.org/abs/1803.01271

[42] Christopher D. Manning. Computational linguistics and deep learning. Computational Linguistics, Volume 41. Issue 4. 2016. Режим доступа: http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00239#.WQH8MBhh2qA

В своем исследовании для обучения сверточной нейронной сети мы выбрали конвейер spaCy, потому что он прост в использовании, не требует серьезного технического оснащения и имеет официальный русскоязычный конвейер, созданный для решения NLP задач.

Мы уже выяснили, что CNN на сегодняшний день (а также ее модификации) считаются лучшими по точности и скоростями алгоритмами, поэтому рассмотрим далее ее архитектуру. Она имеет сложную структуру, состоящую из нескольких слоев, которые показаны на рис. 4. Конвейер пропускает текст комментария через СНС, состоящую из разных видов слоев: сверточные (convolutional) слои, субдискретизирующие (subsampling, подвыборка) слои и слои «обычной» нейронной сети – персептрона, после чего генерируется вывод.

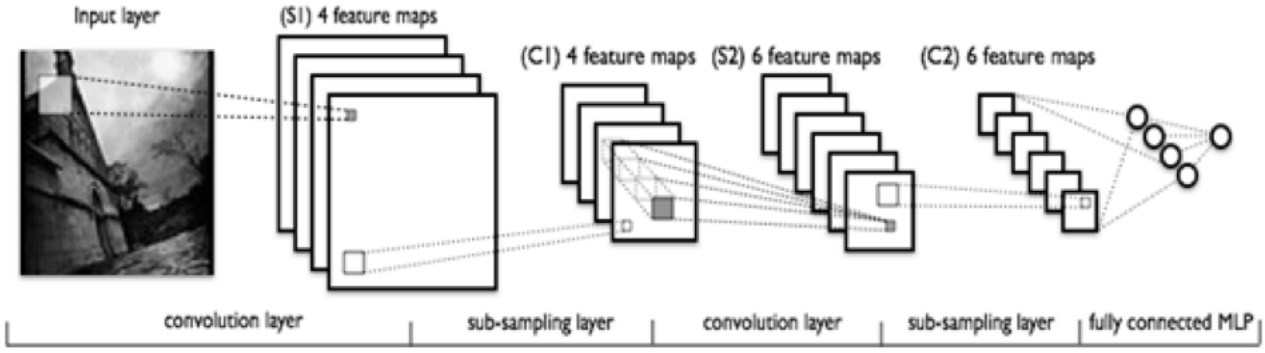

Рисунок 4. Схема сверточной нейронной сети

Вывод может быть представлен в виде класса или вероятность классов, которые лучше описывают данный комментарий. В нашей работе мы будем использовать оба формата вывода.

Свое название технология получила из-за наличия операции свертки, в которой каждый фрагмент изображения умножается на матрицу (ядро) свертки поэлементно, а результат суммируется и записывается на выходе. В архитектуру изначально были заложены знания о компьютерном зрении, где пиксели изображения очень связаны с соседними. Отличные результаты сверхточных нейронных сетей в области компьютерного зрения заставили исследователей

искать применение данного метода для решения других задач, поэтому в последнее время они активно используются при работе с текстовыми данными.

Несмотря на сложную архитектуру нейронных сетей, каждый слой играет важную роль для обучения и выполняет определенную функцию. Рассмотрим каждый из них подробнее.

Полносвязный слой. В этом слое каждый нейрон соединяется со всеми нейронами, выделенными на предыдущем, входном уровне (т.е. оригинал данных), каждая связь которого имеет свой весовой коэффициент.

Сверточный слой. В этом слое нейрон соединяется лишь с ограниченным количеством нейронов предыдущего уровня. Сверточный слой действует также, как и сама свертка, где используется ядро в виде матрицы весов небольшого размера, которую «двигают» по всему обрабатываемому слою. Нейроны с одинаковым весом, объединяются в карты признаков, а каждый нейрон карты признаков связан с частью нейронов предыдущего слоя. В итоге каждый нейрон выполняет свертку некоторой области предыдущего слоя (определяемой множеством нейронов, связанных с данным нейроном).

Субдискретизирующий слой. На данном слое происходит уменьшение размерности с использованием метода выбора максимального элемента, суть которого заключается в том, что вся карта признаков, полученная на предыдущем слое, разделяется на ячейки. В итоге, из этих ячеек выбираются только те, которые имеют максимальное значение.

Dropout слой. Данный слой необходим для борьбы с возможным переобучением в нейронных сетях. На этом этапе каждый нейрон выбрасывается с некоторой вероятностью. В итоге получается прореженная сеть, по которой и производится обучение. Оставшимся нейронам назначается вес и градиентный шаг, а все выброшенные ранее нейроны возвращаются.

Исследователи выделяют несколько подходов для классификации текстов с помощью нейронных сетей. В нашем исследовании мы выбрали подход с

использованием кодирования слов, показанный на рис. 5. Каждому слову в тексте сопоставляется вектор определенной длины, а из полученных векторов для каждого объекта выборки создается матрица, которая аналогично изображениям подается на вход сверточной нейронной сети.

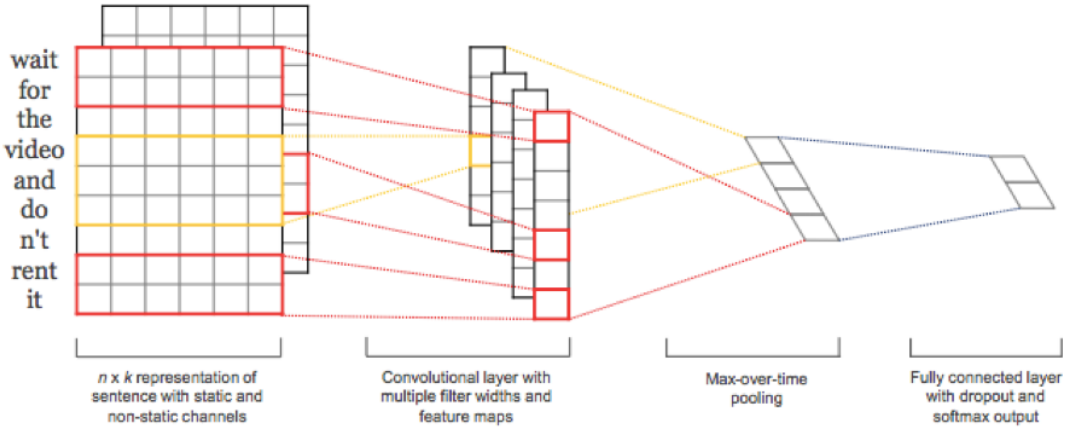

Рисунок 5. Схема подхода с использованием кодирования слов

Существует также посимвольный подход для классификации текстов, который изображен на рис. 6. В настоящем подходе если представить алфавит как упорядоченный набор символов, то каждому символу назначается вектор определенной длины и порядковый номер элемента (единица, если позиция этого элемента равна порядковому номеру символа в алфавите, ноль – для всех остальных случаев). Если в текстовых данных появляется символ, который не встречается в алфавите, то он также кодируется вектором определенной длины,

состоящий только из нулей.

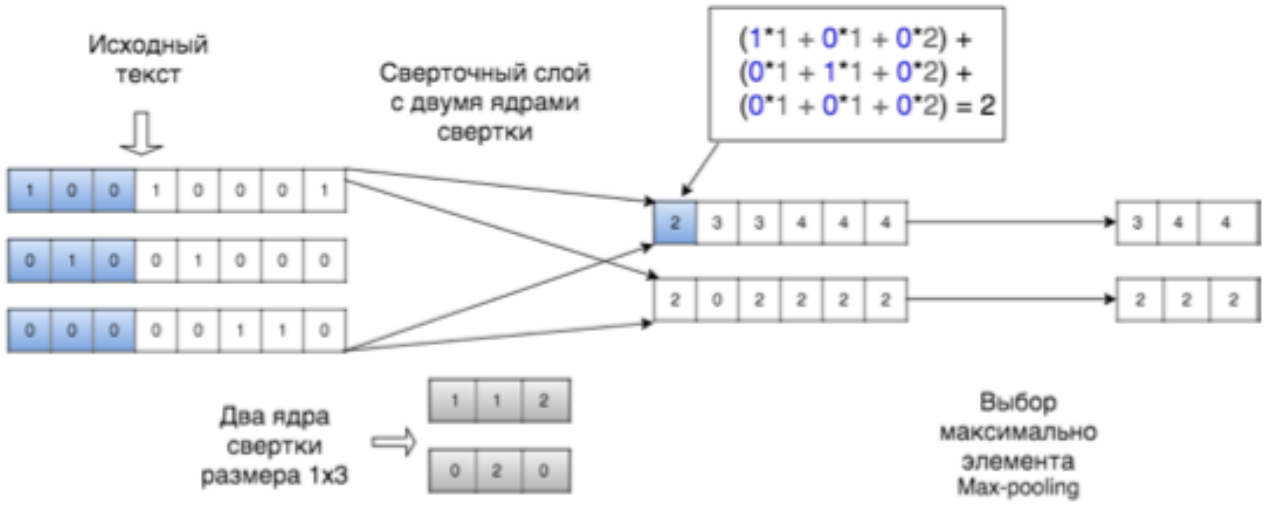

Рисунок 6. Схема посимвольного подхода

Таким образом, применение CNN успешно зарекомендовали себя для решения задач классификации не только изображений, но и текстов. Многие исследователи активно их используют в своих работах, отмечая преимущества сверточных нейронных сетей. Далее в нашей работе мы расскажем, как можно использовать их для классификации текстов, содержащих речевую агрессию.

2. Источники текстовых данных

Для исследования мы решили взять комментарии из социальной сети ВКонтакте, являющейся самой популярной в русскоязычном сегменте. Ее используют ежедневно миллионы пользователей для общения, обмена впечатлениями, демонстрации фотографий и, конечно, комментирования. В СМИ неоднократно появляется информация о том, что данная социальная сеть разрабатывает способы борьбы с киберагрессии. В 2019 году они на один день провели тестирование своей системы. Суть ее заключается в том, что, когда пользователь оставляет агрессивный комментарий, то соц. сеть предупреждает его о том, что сообщение может обидеть человека и предложит написать что-нибудь хорошее. Эксперимент продлился один день, но это вызвало неоднозначную

реакцию. Кому-то очень понравилась эта идея, а кто-то в шутку, наоборот, начал оставлять еще больше агрессивных, чтобы «проверить», как работает алгоритм.

В 2020 году было объявлено, что разработчики ВК снова тестируют свою нейросеть, нацеленную на распознавание хейтспича (враждебных высказываний). Технология призвана ускорить время их обработки и помогает пользователям реже сталкиваться с агрессией в свою сторону. Авторы нейросети утверждают, что за первую половину 2020 года на платформе было удалено 520 тысяч единиц контента на тему разжигания вражды и ненависти, а также заблокировано 1340 профилей и 2470 сообществ, распространяющих враждебные высказывания. Датасет сейчас активно пополняется, поэтому обучение нейросети идет до сих пор.

Несмотря на эти нововведения, в ВКонтакте до сих пор можно найти агрессивные комментарии, причем не только в маленьких группах, но и в пабликах с миллионной аудиторией. В течение нескольких месяцев мы просматривали сообщества, посвященные различным тематикам, изучали их комментарии. Основным источником материала для нашего исследования были те сообщества, где хотя бы половина комментариев под записью были с признаками речевой агрессии. Это позволило потом быстро и качественно пополнить наш датасет для обучения. Этот предварительный анализ и поиск источников занял много времени, но в итоге мы смогли выделить основные виды сообществ, в которых речевую агрессию можно встретить чаще всего:

1. *Сообщества, посвященные одним из самых обсуждаемых тем: ЛГБТ, феминизм, бодипозитив, внешность и др.*

Эта группа сообществ лидирует по количеству агрессивных комментариев пользователей. В предыдущей главе мы выяснили, что тема внешности, особенности характера и сексуальная ориентация являются самыми частыми поводами для киберагрессии. Конечно, со временем агрессивных комментариев становится меньше, люди становятся толерантнее и не боятся высказывать свою точку зрения.

2. Сообщества, освещающие деятельность иностранных знаменитостей.

Данные сообщества посвящены деятельности зарубежных артистов. Ежедневно администраторы выкладывают новостные посты, а пользователи активно их комментируют. Мы заметили, интересное поведение интернет-пользователей в таких сообществах. Их комментарии можно разделить на два вида: первые – это агрессия в сторону знаменитости, о которой написан пост, и вторая – это агрессия между участниками разных фанатских сообществ. Между ними, конечно, можно встретить несколько неагрессивных комментариев, но они составляют значительно меньшую часть по сравнению с другими. Большинство пользователей, которые оставляют комментарии – это фейковые страницы, возможно, созданные специально для того, чтобы провоцировать агрессию у других пользователей. Такие типы сообществ доказывают, что киберагрессия очень удобна для обидчиков. Их разделяет огромное расстояние, что позволяет легко и безнаказанно оскорблять своих некумиров или других пользователей. А анонимность позволяет скрыть свою настоящую личность.

3. Публичные оскорбления в адрес ЛГБТ сообществ Лены Климовой.

Лена Климова является создателем группы, посвященной поддержке ЛГБТ сообществу среди подростков. В ее адрес неоднократно приходили оскорбления и недовольства от противников ЛГБТ-сообществ, причем все они носили личностный характер. Все скриншоты этих оскорблений Лена опубликовала на своей странице в альбоме «Красивые люди, и что они пишут». В нем содержится ок. 500 комментариев. Именно их мы и возьмем для своего датасета. Так как они представлены в виде картинок, то собирать будем вручную, так как в группе они уже не сохранились.

4. Частные новостные сообщества с открытыми комментариями.

Иногда в мире происходят яркие события, которые вызывают огромный резонанс у пользователей. Соответственно, все активные обсуждения происходят

именно в комментариях новостных сообществ, под информационным постом. Комментаторы могу как оскорблять как главного героя новостного события, так и друг друга, если кто-то будет иметь иное мнение по поводу произошедшей ситуации.

3. Сбор текстовых данных

После того, как мы определились с источником сбора данных, мы приступили непосредственно к парсингу комментариев. У ВК есть возможность автоматически скачать через API всю информацию. Единственный минус – это ограничения в количестве комментариев, поэтому многомиллионные паблики нам не подходят. Под некоторыми записями больше тысячи комментариев и система не разрешает скачать даже 100 из них, поэтому сам сбор текстовых данных также растянулся на несколько недель, потому что не все записи сообществ подходили под этот критерий, что увеличивало время поиска подходящих записей и сбора комментариев. На этом этапе мы написали программу для автоматического парсинга текстов комментариев с использованием API на языке Python.

Сама программа имеет простую структуру. Сначала необходимо вызвать функцию авторизации и ввести логин и пароль от учетной записи ВК и запустить сессию с помощью метода **vk_session.get_api().** Далее используя метод **wall.getComments()**, мы ищем необходимую нам запись на стене с помощью переменных **owner_id** (идентификационный номер сообщества) **post_id** (идентификационный номер записи). С помощью метода **comments_strings.append()** и указания типа переменной «**text**» происходит парсинг текста комментария. Все полученные данные автоматически сохранялись в текстовом файле, а каждый комментарий был записан с новой строки, что в дальнейшем нам упростило задачу предварительной обработки текстовых данных. В итоге был получен текстовый файл с текстами комментариев:

я вас ненавижу. Вы отвратительны и если бы в моих руках была бы власть, я бы вас расстрелял. Надеюсь что ваш проект запретят и вас изолируют от общества. Вы никому не нужны, кроме уродцев вроде вас.

чтоб ты сдохла продажная шлюха европы.

Ты чем, сука занимаешься? Тебя посадят.

слышь тварь где ответ.

Параллельно сбору происходил также первичный анализ материала. Мы просматривали все собранные комментарии и удаляли в них ненужные элементы (например, эмодзи или ники пользователей).

Важно отметить, что для обучения качественной модели должны быть использованы не только агрессивные комментарии, но и те, где агрессии нет. Для этого весь собранный ранее материал проверялся, а комментарии вручную распределялись по разным категориям: есть агрессия (метка «ag») и нет агрессии (метка «no_ag»). Тексты каждой категории заносились из общего файла в два разных, в соответствии с обозначенными ранее метками. Во вторую группу могли попадать не только нейтральные тексты, но и положительные, отрицательные, имеющие признаки радостного или грустного настроения. Важно отметить, что мы выявляем именно агрессию в тексте, поэтому анализ остальных настроений не был нашей целью. Именно по этой причине мы обозначаем вторую группу не как «нейтральные», а как те, в которых нет агрессии.

Таким образом мы получили два текстовых файла: первый содержал тексты комментариев с агрессией, а второй – без нее. Всего у нас было собрано 1109 агрессивных комментариев и 1053 без агрессии. Такой небольшой объем датасета объясняется тем, что работа по его сбору была только частична автоматизирована (парсинг комментариев). Необходимо было тщательно просматривать весь материал, удалять ненужные элементы и распределять комментарии по категориям, что является трудозатратной деятельностью.

4. Предварительная обработка и преобразование текстовых данных

Любой рабочий процесс анализа данных начинается с их загрузки в рабочую директорию. Мы провели эту процедуру с помощью функций Python, где указали путь к датасету, определили соотношение обучающих и тестовых данных (80% на 20%), а также количество отбираемых записей (установили лимит 0).

Для повышения точности результатов классификатора любые текстовые данные необходимо пропустить через конвейер **(pipeline)** предобработки:

1. Токенизация используется для разбиения текста на единицы;
2. Удаление стоп-слов позволяет уменьшить «шум» текста;
3. Лемматизация предполагает приведение слов к нормальной форме;
4. Векторизация текстов – перевод текста в числовое представление, более понятную для машины форму.

Для решения указанных задач сегодня существует несколько библиотек (NLTK, TextBlob, spaCy и другие). В нашем исследовании мы будем использовать именно последнюю, так как ее легче интегрировать в наш классификатор. Рассмотрим далее подробно каждый этап.

Токенизация. Это процесс разбиения текста на части. В данной работе для токенизации мы будем использовать встроенный конвейер библиотеки spaCy (**pipeline**).

Удаление стоп-слов. Стоп-слова – это слова, которые не несут смысловой нагрузки для машины. С помощью атрибута токенов **.is_stop** мы очистили токенизированный ранее текст от стоп-слов.

Лемматизация. В нашем конвейере лемматизация происходит автоматически с помощью атрибута **.lemma_.** Стоит отметить, что для русскоязычной модели также необходимо дополнительно установить в рабочую директорию pymorphy2.

Векторизация. Важный этап для обучения классификатора. Машина не знает ни один естественный язык, поэтому мы должны перевести текст в удобный для нее формат – числовой массив. Каждый токен имеет свой уникальный вектор. Векторизация токенов необходима для оценки сходства слов, классификации текстов и т. д. В spaCy токены векторизуются в виде плотных массивов, в которых для каждой позиции определены ненулевые значений. В данной работе векторизация выполняется автоматически с помощью вызова **nlp()** и использования атрибута **.vector.**

Далее подготовленные файлы мы перебираем в наборе данных и загружаем их в список. На этом этапе мы создаем структуру каталогов данных, ищем и открываем текстовые файлы. Содержимое и словарь меток (это стандартный формат для обучения модели в spaCy, в нашем случае метками являются «ag» и «no_ag») добавляем в виде кортежа в список комментариев.

На этом этапе стоп-слова удаляются из обучающей выборки не сразу, потому что это может снизить качество классификатора. Затем необходимо разделить данные для обучения и теста и вернуть два списка комментариев. Каждый комментарий имеет заданную ранее метку («ag» и «no_ag») и значения true и false:

```
('Первый норм чувак, чисто пацанское лицо)\n',
{'cats': {'ag': False, 'no_ag': True}}),
('эту мразь на кол нужно посадить.\n',
{'cats': {'ag': True, 'no_ag': False}}),
('Чего ты, сука, пытаешься добиться? ты башкой то своей подумала, мразь, что ты делаешь? Вот представь, залупа, что будет, если все люди на земле будут гомосеками? Человечество вымрет! У тебя похоже болезнь какая-то психологическая, или тебя батя в детстве пиздил мало, чтоб всю хуйню из твоей башки выбить? Всем сразу станет лучше, если таких как ты не будет, по крайней мере в России. Или тебе США деньги за это платит? Не смей игнорировать это сообщение, я жду развернутого ответа, мразь\n',
```

```
{'cats': {'ag': True, 'no_ag': False}}),
('Уиллушка нормас позу принял, чего стесняться. Это не отменяет
его *издатость, как актера\n',
{'cats': {'ag': False, 'no_ag': True}}),
('Так себе реконструкция...\n', {'cats': {'ag': False, 'no_ag':
True}}),
('Если сама живет, то круто. Еще и работает\n',
{'cats': {'ag': False, 'no_ag': True}})
```

Теперь текст стал машиночитаем, поэтому мы можем начать работу над его обучение классификатора на этих данных.

5. Обучение классификатора

Ранее мы отмечали, что spaCy может делать предварительную обработку текста с помощью конструктора **nlp().** Используемый для этого конвейер находится в файле JSON, связанным с уже существующей русскоязычной моделью.

У конвейера существует встроенный компонент **textcat** (сокр. **TextCategorizer**), который позволяет назначать текстовым данным категории и использовать их в качестве обучающих данных нейронной сети. Чтобы с помощью этого инструмента обучить модель, необходимо выполнить следующие действия:

1. добавить **textcat** в существующий конвейер;
2. добавить в **textcat** необходимые для обучения валидные метки (в нашем случае это «ag» и «no_ag»);
3. загрузить, перемешать и разделить на части данные для обучения;
4. обучить модель, при этом оценивать каждую итерацию;
5. использовать обученную модель, чтобы предсказать наличие речевой агрессии в текстах, которые не входили в обучающую выборку;
6. сохранить обученную модель для дальнейших тестов на реальных примерах.

Существование этого компонента помогает нам регулировать классификатор под свои требования, что облегчает дальнейшее создание классификатора.

Ранее мы отмечали, что CNN можно использовать для классификации разных типов данных. Для того, чтобы сеть понимала, что она будет обрабатывать текст, необходимо указать при строительстве конвейера компонент **textcat**, для обучения которого создается отдельный цикл. В **textcat** также необходимо задать метки, которые использовались для разметки наших текстовых данных: «ag» – для комментариев с речевой агрессией и «no_ag» – для комментариев без речевой агрессии.

В библиотеке spaCy мы не строим конвейер с нуля, а используем уже готовый. В их коллекции есть также мультиязычный конвейер, но мы будем использовать тот, который предназначен именно для русского языка, тем более что существует официальная модель от spaCy, поддерживающая токенизацию и ряд других базовых операций (она называется просто – Russian**()**).

Сначала мы загружаем встроенный конвейер в рабочую директорию и проверяем, доступен ли компонент **textcat**. Если он доступен, то мы переходим к циклу его обучения, самой важной части нашего исследования. Для того, чтобы начать цикл обучения, необходимо сгенерировать для него пакеты данных (та небольшая часть данных, которая участвует в обучении). Пакетная обработка данных позволяет сократить объемы памяти, которая используется во время обучения и в целом оптимизировать процесс обучения.

В итоге мы получаем готовый для дальнейшего тестирования классификатор, обученный на наших данных с использованием официального русскоязычного встроенного конвейера spaCy, который должен успешно определять наличие или отсутствие агрессии в тексте комментариев.

6. Тестирование классификатора и оценка результатов

Любую обученную модель необходимо обязательно протестировать. Для этого используют стандартные метрики: точность и полноту, с помощью которых можно вывести значение F-меры. Перед тем, как определить полноту и точность нашей модели, необходимо определить метрики, по которым эти параметры будут оцениваться:

1. *Истинно агрессивные (истинно-положительное решение, TP)* – количество комментариев, которые классификатор верно определил как агрессивные.

2. *Ложноагрессивные (истинно-отрицательное решение, TN)* – количество комментариев, в которых не было агрессии, но классификатор определил, что есть.

3. *Истинно неагрессивные (ложно-положительное решение, FP)* – количество комментариев, которые модель правильно предсказала как неагрессивные.

4. *Ложнонеагрессивные (ложно-отрицательное решение, FN)* – количество комментариев с агрессией, которые классификатор определил как неагрессивные.

На основе четырех описанных статистических данных мы вычисляем две метрики: точность и полноту, которые являются показателями эффективности данной модели:

Точность – отношение количества истинно агрессивных комментариев к количеству тех элементов, которые модель определила как агрессивные (и истинные и ложные). Максимальный возможный результат точности – 1,0. Он говорит о том, что наша модель верно определила все возможные комментарии с агрессией как агрессивные.

Полнота – отношение количества истинно агрессивных комментариев к сумме истинно агрессивных и ложнонеагрессивных (т. е. тех, которые модель предсказала как неагрессивные, хотя на самом деле в них была агрессия).

Далее приведем полные версии формул, которые мы использовали для определения метрик:

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad \boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad \boldsymbol{F} = \frac{\boldsymbol{2} * \boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}}$$

Обозначив количество итераций и указав лимит обучения, мы получили следующую оценку:

```
Начинаем обучение
  Loss      Prec.     Rec. F-score
2.201556 0.863    0.863    0.863
0.783982 0.868    0.902    0.885
0.294236 0.873    0.941    0.906
0.093111 0.887    0.922    0.904
0.042186 0.885    0.902    0.893
0.020107 0.825    0.922    0.870
```

Мы видим, что с каждой итерацией обучения меняются показатели полноты и точности, но они обычно растут в самых первых итерациях, но потом держатся на одном уровне, но значение функции потерь стремительно уменьшается. В итоге мы получили наилучший вариант полноты – 0. 887, точности – 0.941. Исходя из этого можно сделать вывод, что наша модель неплохо справляется с поставленной задачей. Стоит отметить, что сентимент-анализ – это достаточно сложный процесс, в котором очень трудно получить стопроцентный (или приближенный к этому) результат. В целом мы остались довольны такими итогами. Эти показания можно также улучшать путем пополнения датасета.

После всех пройденных этапов мы получили обученную модель, которую протестировали на реальных комментариях, взятых из социальной сети ВКонтакте, чтобы определить, насколько успешно она справится со своей задачей. В переменную вводим текст любого реального комментария:

Значение этой переменной мы передаем модели, чтобы сгенерировать прогноз и вернуть его значение пользователю. В функции мы передаем входные данные в загруженную и обученную ранее модель и генерируем предсказание. Затем проверяем оценки каждой категории («ag» и «no_ag») и сохраняем ту, у которой более высокий прогноз. В итоге нами были получены следующие результаты.

Мы проверили случайно отобранные комментарии, взятые из социальной сети ВКонтакте, на нашем классификаторе. Нами были взяты тексты разные по размеру и составу. Среди комментариев для эксперимента также были выбраны несколько текстов без речевой агрессии, чтобы убедиться, что классификатор не будет находить агрессию там, где ее нет.

1. Текст комментария: Уроды сраные на кол и в огонь утомили извращюги их стремление ,или стремление сильных мира нас ими изгноить в бездну!!

Предсказание: Осторожно, агрессия

Score: 0.678

2. Текст комментария: стремно видеть в коммах как все судят человека только по ее волосам и как она сидит мда:^)

Предсказание: Агрессии нет

Score: 0.999

3. Текст комментария: еужели только мужики могут ходить с волосатыми подмышками? Русобляди, когда вы ума наберетесь. ужас кошмар отврат мерзость ничтожность никчемность тупость нелепость глупость низость нечестие грязь

Предсказание: Осторожно, агрессия

Score: 0.947

4. Текст комментария: Очередная пиздофемка с небритым ебалом.

Предсказание: Осторожно, агрессия

Score: 0.952

5. Текст комментария: Что за ущербная попытка выглядеть как Гари Стайлс. Видимо фанатки дрочащие на его ебло закончились

Предсказание: Осторожно, агрессия

Score: 0.947

Классификатор также справляется с имплицитными способами выражения агрессии:

Текст Комментария: Стройные, спортивные, модные, стильные – вы все вызываете рвотный рефлекс. Хочется загнать всех в одну комнату и пустить газ.

Предсказание: Осторожно, агрессия

Score: 0.981

Но иногда с имплицитной получается не всегда хороший результат:

Текст комментария: Королева рехабов, немытости, небритых вонючих подмых, бездарности, безголосости, татух-партаков

Предсказание: Агрессии нет

Score: 1.000

Или если встречаются зацензуренные варианты:

Текст комментария: Как не крути, х*й и п*зда - только одни

Предсказание: Агрессии нет

Score: 1.000

Эта ситуация исправима, так как необходимо в дальнейшем добавить больше таких «скрытых» вариантов, чтобы классификатор понял, что такой способ выражения также приравнен к агрессии.

Классификатор также неплохо справляется с длинными комментариями:

1. Текст комментария: Тебе почти 30, а имидж инфантила переростка, судя по всему отражает его личность. Уже надоедает смотреть на бича из гетто, обосраные штаны и манеру поведение негрилы. Все же голос хороший, надеялся изменится с возрастом, станет более стильным, элегантным. Деревню все же из человека не вывезешь.

Предсказание: Осторожно, агрессия

Score: 0.511

2. Текст комментария: Внешность - это не только мейк, если бы не красился и не накачивал филеры в губы, как какая-нить милфа из провинции, был бы обычным симпатичным парнем, но слово "обычный" сейчас хуже всратого.

   Предсказание: Агрессии нет

   Score: 0.990

Во втором комментарии нет агрессии, здесь человек объясняет свое мнение, при этом используя ненормативную лексику, но классификатор смог определить отсутствие речевой агрессии в тексте.

Неагрессивные варианты классификатор также успешно определяет:

Текст комментария: Надеюсь на этот раз у нее все получится, за прошлый альбом конечно обидно было, хотя он был достоин топ 20

Предсказание: Агрессии нет

Score: 0.998

И даже в тех случаях, где используются аллегория, мы получаем правильный результат:

Текст комментария: Кабан побежала на воронежский базар за перекисью чтобы обезцветить свою мохнатку. Надеюсь сожжет все там к херам дабы не плодилась.

Предсказание: Осторожно, агрессия

Score: 0.858

Иногда бывает, что личностная агрессия может проявляться не прямым способом, а через восхваление противоположного человека, но даже здесь мы видим, что классификатор справился:

Текст комментария: настоящий талант в отличие от бездарного растлителя беззащитных младенцев

Предсказание: Осторожно, агрессия

Score: 0.787

Если встречаются комментарии, содержащие новые, никому не известные слова, то классификатор также их правильно определяет.

```
Текст комментария: Срагоебы вы ценник этого белья видели икаю когда фанаты леди сраки рекламирующей блевотное розовое печенье с опарышами и такие же палетки для трансов срага бьюти после использования которых развиваются кожные заболевания кудахчат о брезгливости
Предсказание: Осторожно, агрессия
Score: 0.725
```

Прогноз нашей модели и классификатора действительно совпадают с ожиданиями. Для наглядности мы также вывели параметр Score, где можно увидеть коэффициент вероятности настроения (агрессивного или нет).

Таким образом, мы создали эффективную модель для обучения классификатора и сам классификатор, который показывает высокие результаты при решении поставленной задачи.

7. Выводы к главе II

В этой главе мы рассмотрели практическую сторону нашего исследования, где попробовали создать классификатор на основе сверточных нейронных сетей, обученный на собственном датасете. Мы выяснили, что такой вид нейронных сетей наилучшим образом подходит для классификации текстов. Стоит отметить, что они не универсальны и для других задач компьютерной лингвистики они могут быть менее эффективными. Прежде чем приступить к обучению, мы проанализировали практический опыт некоторых исследователей, которые для решения зада определения тональности текста использовали нейронные сети, что определило наш окончательный выбор.

В течение долгого времени мы изучали основные источники нашего материала и собирали датасет. В итоге у нас получился набор текстовых данных,

объемом около 2 000 комментариев, где половина из них содержит речевую агрессию, а вторая – нет.

В результате нашей практической работы был получен классификатор, который способен успешно автоматически выявлять агрессию в текстах сетевых сообществ.

Заключение

В данной работе представлена попытка автоматизировать выявление агрессии в текстах сетевых сообществ.

В теоретической части работы были проанализированы лингвистические особенности речевой агрессии и определена значимость проблемы киберагрессии. Мы выяснили, что для полного изучения речевой агрессии, ее необходимо рассматривать с точки зрения разных наук и их отраслей, так как это достаточно сложное явление, затрагивающее различные сферы. Данная теоретическая информация необходима для понимания объекта нашего изучения, его особенностей и структуры. Без этой информации невозможно было бы достичь цели нашего исследования.

В практической части мы подробно описали создание классификатора, который определяет наличие речевой агрессии в текстах сетевых сообществ, определили причины использования именно сверточных нейронных сетей и конвейера spaCy. В результате нами был собран датасет, состоящий из комментариев, и классификатор, которые успешно справляются со поставленной задачей.

Цель данного исследования была достигнута, и мы видим несколько векторов дальнейшего его развития. Прежде всего можно увеличить размер датасета, для того чтобы улучшить эффективность классификатора. Чем больше примеров знает классификатор, тем с большим количеством комментариев он может справляться. На наш взгляд, нет предельного количества комментариев, которые должны быть в датасете. Язык имеет тенденцию к изменениям, появлениям новых слов, особенно это характерно для лексики интернет-пространства, поэтому необходимость в пополнении нашего датасета будет всегда.

В дальнейшем разработанный классификатор также можно использовать и для других исследований, связанных с классификацией текстовых данных, так как мы выяснили, что сверточные нейронные сети лучше всего справляются именно с

этой задачей. Его структура позволяет использовать свою систему меток для разметки собранного датасета и, например, может быть применена для определения тональности текста (в частности, отзывов). Если для обучения модели предоставить качественные текстовые данные, то результаты могут быть более высокими. Кроме того, он подходит практически для всех языков, так как его архитектура основана на конвейере библиотеки spaCy, в арсенале которой представлены также конвейеры для других языков, в том числе мультиязычный.

Таким образом, в нашей работе мы подняли проблему киберагрессии, отметили, что в реальных условиях есть необходимость искать способы борьбы с ней и предложили свой вариант, в основе которого лежит классификатор, обученный на реальных данных и способный определять содержание речевой агрессии в тексте. Надеемся, что наше исследование может вызвать интерес к затронутой теме, а разработанная нами программа позволит своевременно выявлять агрессивные комментарии и, таким образом, будет способствовать благоприятному общению в социальных сетях.

Список используемой литературы

1. Апресян, В. Ю. Имплицитная агрессия в языке / В. Ю. Апресян // Компьютерная лингвистика и интеллектуальные технологии: тр. Междунар. конф. «Диалог 2003». – М. : Наука, 2003. – С. 32-35.
2. Бахтин, М. М. Франсуа Рабле и народная культура средневековья и Ренессанса / М. М. Бахтин. – М. : Художественная литература, 1990. – 544 с.
3. Воробьев, Н. В., Пучков Е. В. Классификация текстов с помощью сверточных нейронных сетей [Электронный ресурс] / Н.В. Воробьев, Е.В. Пучков // Молодой исследователь Дона. – Ростов-на-Дону., 2017. – №6. Режим доступа : https://mid-journal.ru/upload/iblock/8ed/1.-vorobev_-puchkov.pdf (дата обращения : 1.04.2021).
4. Воронцова, Т. А. Речевая агрессия : автореф. дис. ... д-ра филол. наук / Т. А. Воронцова. – Челябинск, 2006. – 43 с.
5. Гловинская, М.Я. Гипербола как проявление речевой агрессии // Сокровенные смыслы: сб. статей в честь Н. Д. Арутюновой. – М., 2004. – С. 69-76.
6. Горелов, И. Н. Основы психолингвистики / И. Н. Горелов, К. Ф. Седов. – М., 2001. – 149 с.
7. Енина, Л. В. Катартический характер речевой агрессии в сверхтексте лозунгов и источники ее смягчения / Л. В. Енина // Вопросы стилистики: Антропоцентрические исследования. – Саратов, 1999. – Вып.28. – С. 103-107.
8. Енина, Л. В. Речевая агрессия и речевая толерантность в средствах массовой информации / Л. В. Енина // Российская пресса в поликультурном обществе: толерантность и мультикультурализм как ориентиры профессионального поведения. – М., 2002. – С. 104-110.
9. Жельвис, В. И. Поле брани. Сквернословие как социальная проблема в языках и культурах мира / В. И. Желвис. – М. : Ладомир, 1997. – 330 с.

10. Закоян, Л. М. Речевая агрессия как предмет лингвистических научныхисследований / Л. М. Закоян // Полилингвиальность и транскультурные практики. – 2008. – №2. – С. 46-52.

11. Карпенко, Л. А. Психология. Словарь / Под общ. ред. А. В. Петровского, М. Г. Ярошевского. – М.: Политиздат, 1990. – 494 с.

12. Киберугрозы, киберагрессия, кибербуллинг: различия в восприятии, оценке и поведении у разных групп населения Российской Федерации [Электронный ресурс] – Режим доступа : https://raec.ru/activity/analytics/9880/ (дата обращения : 4. 05. 2020 г.).

13. Коданева, С. И. Кибрбуллинг: причины явления и методы предупреждения / С. И. Коданева // Социальные новации и социальные науки. – М.: ИНИОН РАН, 2020. – №1. – С. 149-159.

14. Курьянова, И. В. Маркеры речевой агрессии в интернет-коммуникации при исследовании текстов экстремистской направленности / И. В. Курьянова // Вестник МГЛУ. Гуманитарные науки. – М., 2018. – С. 29-38.

15. Кусов, Г. В. Оскорбление как иллокутивный лингвокультурный концепт: Автореф. дис… канд. филол. Наук / Г. В. Кусов. – Волгоград, 2004. – 27 с.

16. Ле Мань Ха. Свёрточная нейронная сеть для решения задачи классификации / Ле Мань Ха // Труды МФТИ. – М., 2016. – Том 8. – № 3. – С. 91-97.

17. Михальская, А. К. Русский Сократ: Лекции по сравнительно-исторической риторике. − М.: Изд. центр. «Academia», 1996. – 192 с.

18. Николаева, Т. М. О принципе «некооперации» и/или о категории социолингвистического воздействия / Т. М. Николаева // Логический анализ языка: Противоречивость и аномальность текста. – М., 1990. – 167 с.

19. Седов, К. Ф. Агрессия как вид речевого воздействия / К. В. Седов // Прямая и непрямая коммуникация. – Саратов: «Колледж», 2003. – С. 110-113.

20. Солдатова, Г. У. Кибербуллинг : особенности, ролевая структура, детско-родительские отношения и стратегии совладания / Г. У. Солдатова, А. Н. Ярмина // Национальный психологический журнал. – М., 2019. – № 3. – С. 17–31.

21. Стернин, И. А. Введение в речевое воздействие / И. А. Стернин. – Воронеж, 2001. – 227 с.

22. Стилистический энциклопедический словарь русского языка / под ред. М. Н. Кожиной. – М., 2006. – 696 с.

23. Тиллабаева, А. А. Речевое поведение интернет-пользователей в ситуации конфронтационного общения / А. А. Тиллабаева, В. А. Шульгинов // Слово.ру: балтийский акцент. – Калининград. – 2020. – Т. 11. – №4. – С. 45-57.

24. Черенков, Д. А. Девиантное поведение в социальных сетях: причины, формы, следствие [Электронный ресурс] / Д. А. Черенков // Nauka-rastudent. Ru. – 2015. – № 07. Режим доступа : https://readera.org/14330143 (дата обращения : 3.05.2021).

25. Шаров, А. А. Специфика девиантной активности молодежи в интернет-среде / А. А. Шаров // Учен. записки. Электронный научный журнал Курского государственного университета. – 2019. – № 3. – С. 255–261.

26. Щербинина, Ю. В. Русский язык: Речевая агрессия и пути ее преодоления. – М., 2012. – 224 с.

27. Ярец, А. Д. Разновидности конфликтов и агрессии в интернет-коммуникации [Электронный ресурс] / А. Д. Ярец // Идеи. Поиски. Решения: сборник статей и тезисов XIII Международной научно-практической конференции преподавателей, аспирантов, магистрантов, студентов. – Минск: БГУ, 2020. – С.201-214. Режим доступа: http://elib.bsu.by/handle/123456789/241180 (дата обращения: 20.05.2021).

28. Bai, S., Kolter, J. Z., & Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. Режим доступа : https://arxiv.org/abs/1803.01271 (дата обращения : 23.03.2021).

29. Bengio, Y. Learning deep architectures for AI // Foundations and Trends in Machine Learning. Режим доступа: https://www.researchgate.net/publication/215991023_Learning_Deep_Architectures_for_AI (дата обращения: 7.05.2021).

30. Calpbinici, Arslan. Virtual behaviors affecting adolescent mental health: The usage of Internet and mobile phone and cyberbullying // Journal of Child and Adolescent Psychiatric Nursing. – 2019. – Vol. 32. – N 3. – P. 139-148.

31. Christopher, D. Manning. Computational linguistics and deep learning. Computational Linguistics. – 2016. – Vol. 41. – Issue 4. Режим доступа : http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00239#.WQH8MBhh2qA (дата обращения : 4.04.2021).

32. Schoffstall, C., Cohen, R. Cyber-Aggression: The Relation between Online Offenders and Offline Social Competence // Social Development. – 2011. – Vol. 20. – Issue 3. – P. 586–604.

33. Willard, N. E. Cyberbullying and Cyberthreats : Responding to the Challenge of Online Social Aggression, Threats, and Distress. Champaign, Illinois : Research Press, 2007. – 320 p.