МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

СӘТБАЕВ УНИВЕРСИТЕТІ

Институт кибернетики и информационных технологий
Кафедра «Программной инженерии»

Сымагулов Адилхан

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

На соискание академической степени магистра

Название диссертации	Выявление криминогенного содержания в текстах на естественном языке
Направление подготовки	6М100200 – Системы информационной безопасности

Научный руководитель
доктор тех. наук, проф.
Р. И. Мухамедиев
«__» _____ 2020 г.

Оппонент
Доктор PhD, заведующий кафедры
Н.Т. Дузбаев
«__» _____ 2020 г.

ДОПУЩЕН К ЗАЩИТЕ
Заведующий кафедрой ПИ
Доктор PhD
_____ М. Тұрдалыұлы
«__»_____ 2020 г.

Нормоконтроль
лектор
_____ Ж.М.Алибиева
«__» _____ 2020 г.

Алматы 2020

Институт информационных и телекоммуникационных технологий

Кафедра Программная инженерия

Специальность: 6М100200 – Системы информационной безопасности

УТВЕРЖДАЮ
Заведующий кафедрой ПИ
Доктор PhD
_____ М. Тұрдалыұлы
«___» _____ 2020 г.

ЗАДАНИЕ
на выполнение магистерской диссертации

магистранту Сымагулову Адилхану

Тема диссертации: «Выявление криминогенного содержания в текстах на естественном языке»

Срок сдачи законченной диссертации «___» _____

Исходные данные к магистерской диссертации. Дан набор новостных текстов. Поставленные цели и задачи диссертационной работы направлены на анализ методов и создание модели машинного обучения для работы с текстовыми данными.

Перечень подлежащих разработке в магистерской диссертации вопросов или краткое содержание магистерской диссертации: а) определение требований – анализ целей, задач и назначения разработки; б) исследование подхода, реализуемого в модели классификации текстов; в) сбор и предобработка данных; г) создание, увеличение, очистка криминогенного словаря; д) эксперименты по оценке качества модели.

ГРАФИК
подготовки магистерской диссертации

Наименование разделов, перечень разрабатываемых вопросов	Сроки представления научному руководителю	Примечание
Раздел 1. Подход		
Раздел 2. Основные эксперименты		

Консультации по проекту с указанием относящихся к ним разделов проекта

Раздел	Консультант, (уч. степень, звание)	Сроки	Подпись
Нормоконтроль	Ж.М.Алибиева, Лектор кафедры программная инженерия		

Дата выдачи задания "____"_____ 2020 г.

Заведующий кафедрой _____ М. Тұрдалыұлы

Научный руководитель _____ Р. И. Мухамедиев

Задание принял к исполнению магистрант _____ А. Сымагулов

Дата "_____"_____ 2020 г.

АННОТАЦИЯ

В данной работе рассматривается применение словарного метода в задаче классификации текстов на естественном языке для класса «криминогенный». Определение криминогенности также приведено в тексте данной работы. Вручную создан криминогенный словарь из 274 слов, который затем автоматически увеличен с помощью методов машинного обучения до 2286 семантически близких к исходным слов. Построена модель классификации криминогенных текстов. Проведены эксперименты на двух выборках новостных текстов размером в пять тысяч и одну тысячу текстов.

Сделаны выводы о том, что использование увеличенного криминогенного словаря на выборке из пяти тысяч текстов порождает криминогенный класс более чистый и большего размера по сравнению с исходным не увеличенным словарем. В свою очередь использование исходного не увеличенного словаря на выборке из одной тысячи текстов порождает криминогенный класс более чистый и большего размера по сравнению с увеличенным словарем. Выводы о таком поведении модели приведены в конце раздела 2.5 данной работы.

АҢДАТПА

Бұл жұмыста мәтіндерді криминогендік класс үшін табиғи тілде жіктеу міндетінде лексика әдісін қолдануды қарастырамыз. Сондай-ақ, криминогендік анықтама осы жұмыстың мәтінінде келтірілген. 274 сөзден тұратын криминогендік сөздік қолмен жасалды, ол автоматты оқыту әдісімен бастапқы сөздерге семантикалық жақын 2286 сөзге дейін көбейтілді. Криминогендік мәтіндердің классификациясы үшін үлгі жасалды. Бес мың және бір мың мәтіндік жаңалықтар мәтіндерінің екі үлгісі бойынша эксперименттер жүргізілді.

Бес мың мәтіннің үлгісінде кеңейтілген криминогендік сөздікті пайдалану бастапқы кеңейтілген сөздікпен салыстырғанда таза және үлкенірек криминогендік класс тудырады деген қорытындыға келді. Өз кезегінде, мың мәтіннің үлгісінде түпнұсқа үлкейтілмеген сөздікті пайдалану, кеңейтілген сөздікпен салыстырғанда таза және үлкенірек криминогендік сынып тудырады. Модельдің осы әрекеті туралы қорытынды осы қағаздың 2.5 бөлімінің соңында келтірілген.

SUMMARY

This paper considers the application of the vocabulary method in the task of classifying texts in natural language for the class "criminogenic". The definition of criminogenicity is also given in the text of this paper. A 274-word criminogenic dictionary has been manually created, which is automatically increased by means of machine learning methods to 2286 semantically close to the source words. The machine learning model of classification of criminogenic texts is constructed. Experiments on two samples of news texts with the size of five thousand and one thousand texts were conducted.

The conclusions are made that the use of an enlarged crime dictionary on a sample of five thousand texts generates a crime class that is cleaner and larger in size than the original non-enlarged dictionary. In its turn, the use of a source dictionary on a sample of one thousand texts generates a criminal class that is cleaner and larger than the enlarged dictionary. Conclusions about such behavior of the model are given at the end of Section 2.5 of this paper.

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей диссертации применяют следующие обозначения и сокращения с соответствующими определениями:

Обозначение/сокращение	**Определение**
Гистограмма	Столбчатая диаграмма
Датасет	Набор данных
Парсинг	Сбор информации с интернет-сайтов
РФ	Российская Федерация
Сервер	Устройство для обработки/хранения информации
СМИ	Средства Массовой Информации
Тэг	Идентификатор при категоризации
CBOW	Continuous Bag of Words
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol

ВВЕДЕНИЕ

На данный момент, во время стремительного развития информационных технологий новостные агентства, выполняющие роль СМИ (средства массовой информации) в нашем обществе, занимают лидирующие позиции в том, откуда мы черпаем информацию о происходящем в мире и вокруг нас. СМИ имеют огромное влияние на то, как общество живет, думает и рассуждает. Огромное и быстрорастущее количество новостных материалов делает невозможным их ручной анализ. Своевременная автоматическая обработка новостных текстов с целью «отлова» подозрительных, провокационных и иных деструктивных материалов становится необходимым условием качественного развития каждого члена нашего общества [1].

Известно, что акты насилия, будь то на телеэкранах или в сводках новостей способны нанести вред психологическому здоровью человека [2, 3, 4]. Чаще всего негативное влияние на человека несут в себе новостные тексты о совершенных преступлениях, порой имеющие детальное описание ужасов, зверств, аморального поведения преступников, описывающие мошеннические схемы или т.н. легкую наживу и т.п, такие тексты, а также любые другие, способные нанести вред психологическому здоровью человека — являются криминогенными, так как из-за их пагубного влияния, в будущем они могут поспособствовать совершению новых преступлений. В работе [5] авторы проводят исследование взаимосвязи между мониторингом СМИ родителями и поведением их детей-подростков, отмечая в выводах, что ограничительный и активный мониторинг СМИ родителями тесно связан с социальным поведением детей. Вышесказанное позволяет говорить о том, что родительское фильтрование информации, которую потребляют их дети положительно сказывается на том, как дети мыслят и соответственно действуют. Такое или схожее воспитание/коррекцию автор данной работы считает применимым не только в рамках семьи, но и в государстве.

В связи с вышесказанным создание модели классификации новостных текстов позволит решить задачу предотвращения негативного влияния СМИ на человека, оградив его от деструктивной информации, что, несомненно положительно скажется на жизни индивида и всего государства.

Целью данной работы является создание модели классификации текстов, которая послужит своеобразным фильтром для новостных текстов криминогенного характера, а также будет достаточно динамична для того, чтобы на её основе можно было создавать новые фильтры. Создание новых фильтров возможно при создании новых тематических словарей.

Автор данной работы не ставит перед собой цель закрыть глаза всем читателям на проблемы общества и/или лишить служителей новостных агентств возможности освещать события, происходящие в стране. Работа новостных агентств является важной в условиях качественного развития, как индивида, так и целого государства. В связи с вышесказанным целью автора данной работы

является лишь предоставление возможности каждому индивиду оградить себя от информации, которая способна причинить ему вред.

Для достижения поставленной цели необходимо решить набор задач, одна из которых — это задача классификации. Задача классификации состоит в том, чтобы разбить объекты на уже известные классы по каким-либо признакам, которые соответствуют имеющимся классам, эта задача обычно может вытекать из результата решенной задачи кластеризации. Задача кластеризации представляет собой разделение объектов на неопределенные группы, называемые кластерами. В каждом кластере должны находиться похожие/близкие друг к другу объекты, а объекты из разных групп должны быть различны/не похожи друг на друга. Отличаются задачи кластеризации и классификации тем, что при кластеризации неизвестны ни количество групп, ни их названия, в то время как при классификации количество групп и их названия известны заранее [6]. На рисунке 1 приведен пример классификации для одного класса, то есть отделение определенного класса, отмеченного красным овалом, от всех остальных, отмеченных фиолетовым овалом.

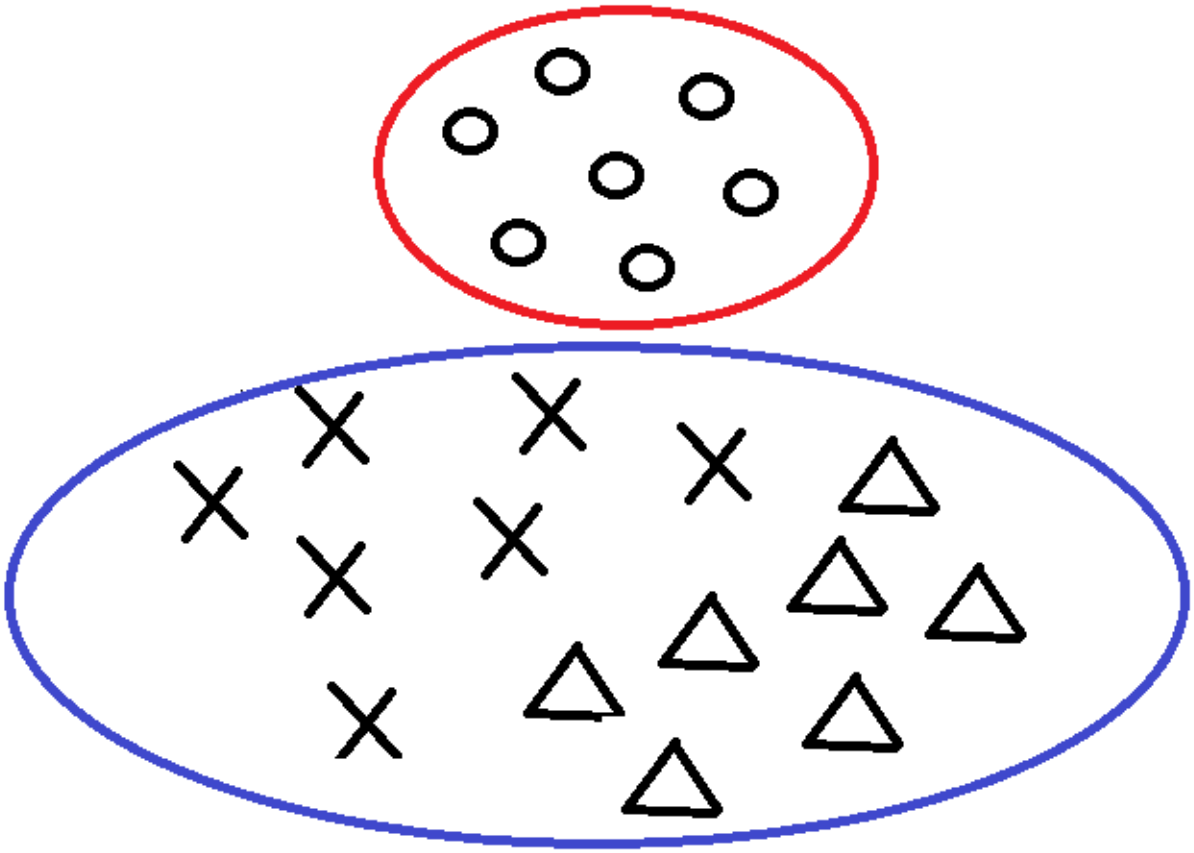

Рисунок 1. Классификация для одного класса

Так же одной из задач для достижения поставленной цели является подготовка текстов к работе с ними, эта задача называется предобработкой текстов, в нее входят такие задачи, как: токенизация, очистка от стоп-слов, лемматизация, морфологическая разметка и другие.

Токенизация представляет собой разбиение текстов на так называемые токены, токенами для текстов являются их слова, разбиение текстов на отдельные слова позволяет производить работу, как с отдельными словами, так и с полным текстом, собрав токены воедино. На рисунке 2 можно ознакомиться с тем, как решается задача токенизации на примере одного предложения, которое заключено в квадратные рамки. Результат токенизации представляет собой отдельные слова из предложения, которые заключаются в кавычки и имеют свой индекс, ведущий отсчет от нуля и далее.

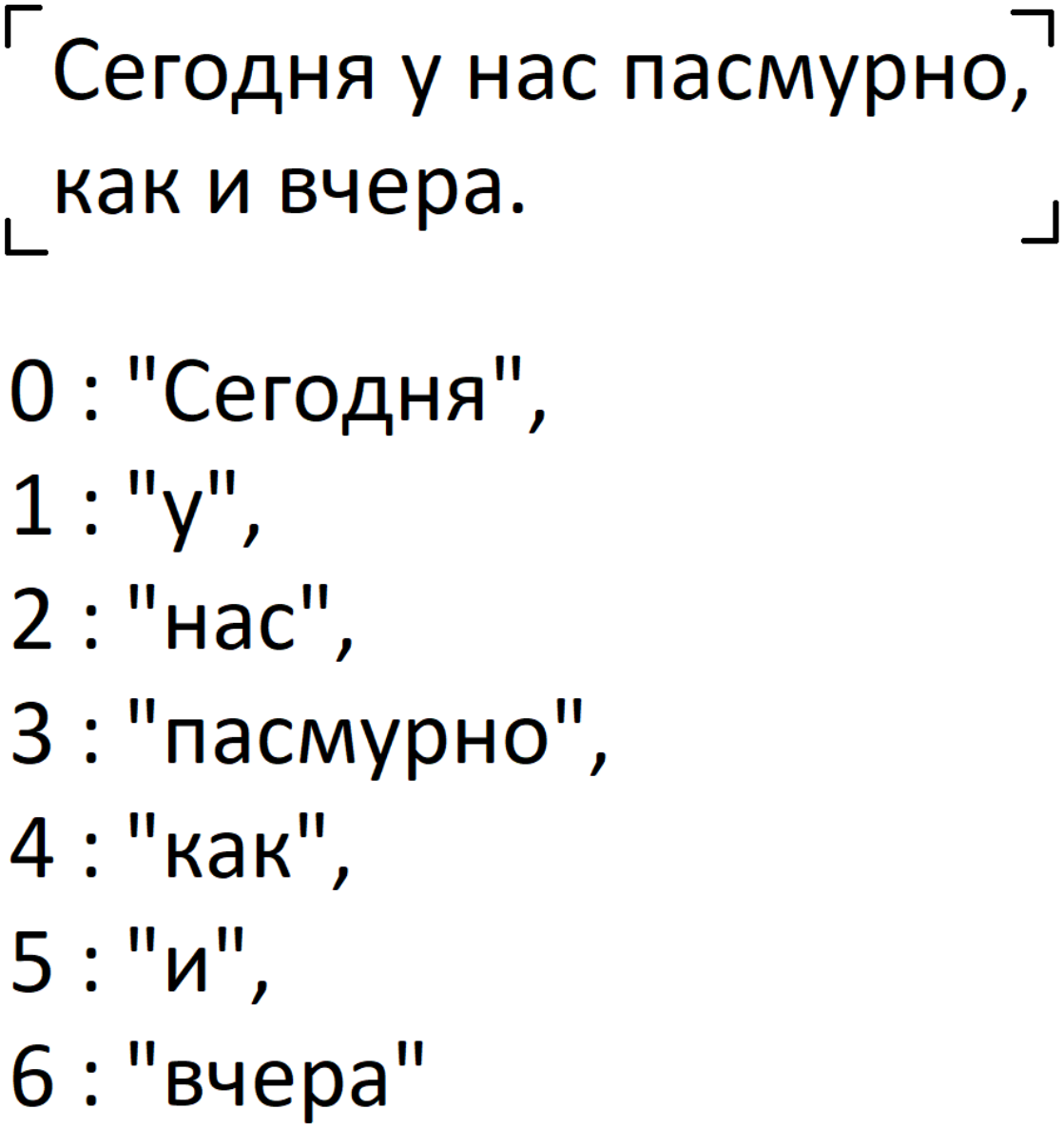

Рисунок 2. Пример токенизации

Очистка от стоп-слов подразумевает удаление из текста тех слов, которые не несут в себе смысла для компьютера, хотя и обладают смыслом для человека, например такие слова, как «что, под, над, и, в, я, ты» и тому подобные слова. На рисунке 3 приведен пример того, какие слова остаются после проведенной очистки от стоп-слов. Можно заметить, что исходный смысл предложения сохранился и его можно передать в краткой форме, изменив порядок оставшихся слов - «Вчера, сегодня пасмурно».

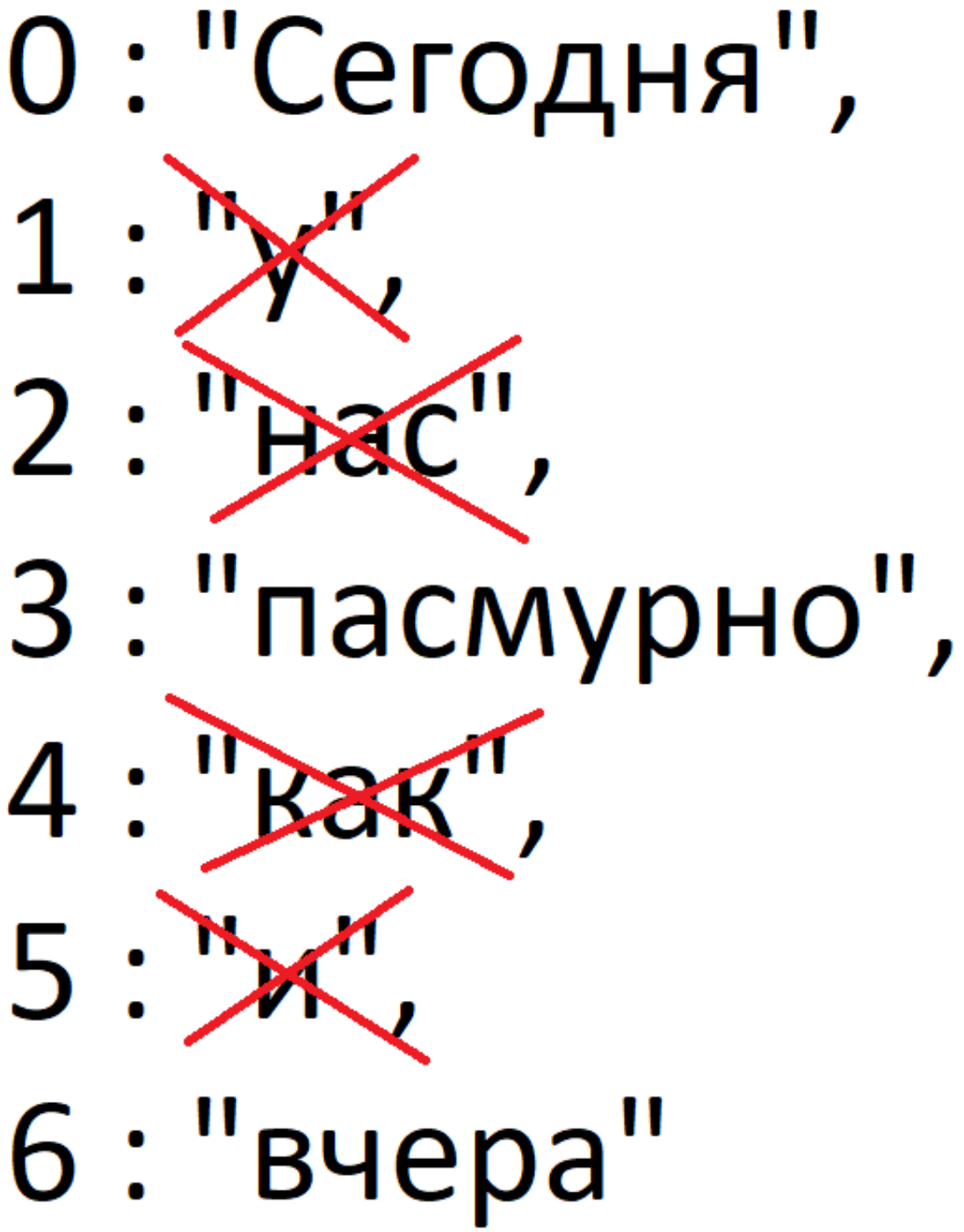

Рисунок 3. Пример очистки от стоп-слов

Задача лемматизации сводится к тому, чтобы привести слово к его словарной/начальной форме, например: слово “Автомобилем” превратится в слово “Автомобиль”, это позволит компьютеру не делать различий между одним и тем же словом в разных склонениях, а “понимать” его как одно слово, это позволит обезопасить точность дальнейшей работы. На рисунке 4 приведен пример того, как слова приводятся к словарной/начальной форме.

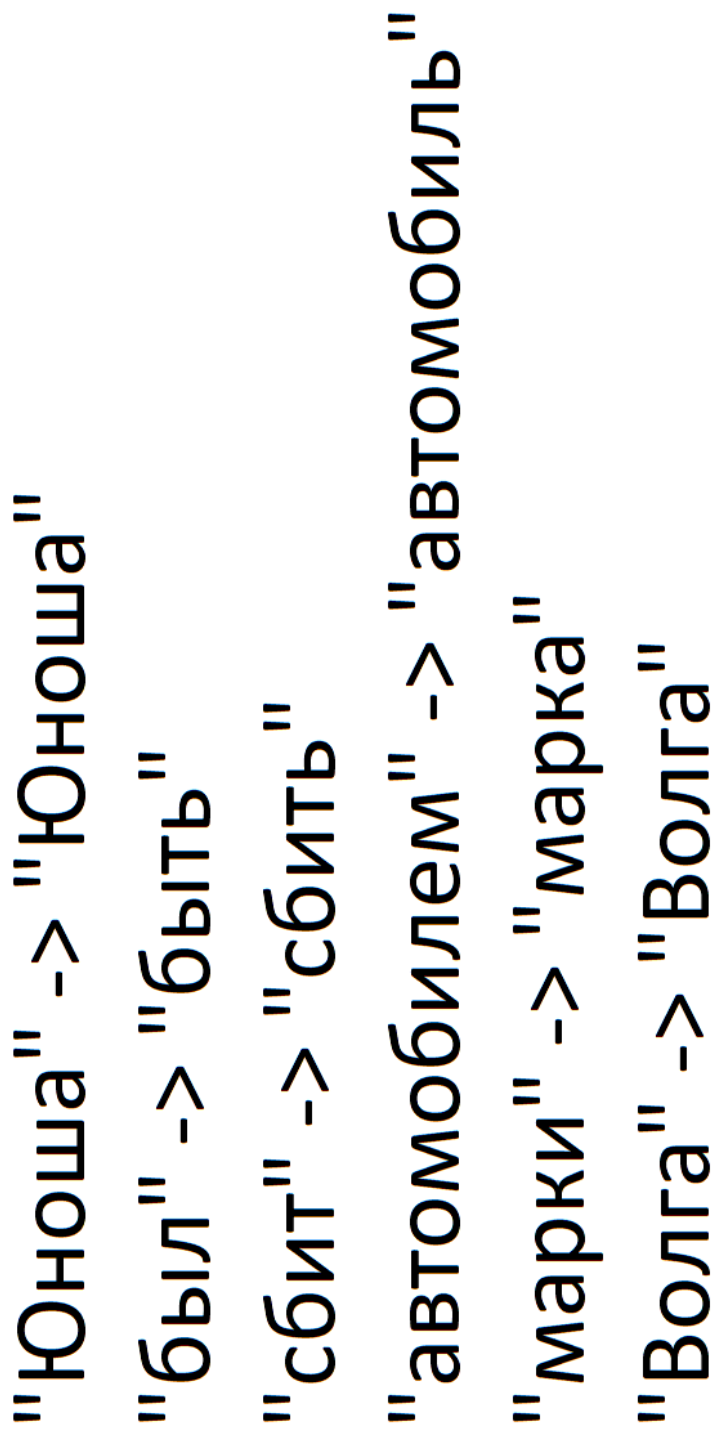

Рисунок 4. Пример проведенной лемматизации

Морфологическая разметка представляет собой присвоение каждому слову его части речи, например “Стекло_существительное” и “Стекло_глагол”. Из примера со словом “Стекло” видно, что первое слово является существительным, тогда как второе глаголом, это говорит о том, что первое слово употреблялось в отношении изделия, сделанного из стекла, а второе слово употреблялось по отношению к жидкости, которая стекала откуда-то. Правильно выполненная морфологическая разметка вкупе с лемматизацией способны решить так называемую задачу снятия омонимии, которая является открытой задачей в обработке естественного языка и по сей день. На рисунке 5 приведен результат проведенной морфологической разметки.

Для того, чтобы определить несет ли в себе текст криминогенное содержание, предлагается воспользоваться словарным методом, это значит, что должен быть создан словарь, содержащий специфичные слова, принадлежащие определенному классу новостных текстов, в нашем случае криминогенному. Затем необходимо измерить близость новостных текстов к криминогенному словарю, что в свою очередь позволит определить принадлежность текста к криминогенному классу. Использование словарного метода позволит создать модель, которая будет являться белым ящиком, то есть будет легко интерпретируемой от начала и до конца работы модели в отличие от черных ящиков, которые довольно сложны в интерпретации [7, 8]. На рисунке 6 можно увидеть пример криминогенного и не криминогенного текста.

"Юноша" -> "Юноша_существительное"
"был" -> "быть_глагол"
"сбит" -> "сбить_глагол"
"автомобилем" -> "автомобиль_существительное"
"марки" -> "марка_существительное"
"Волга" -> "Волга_существительное"

Рисунок 5. Пример морфологической разметки

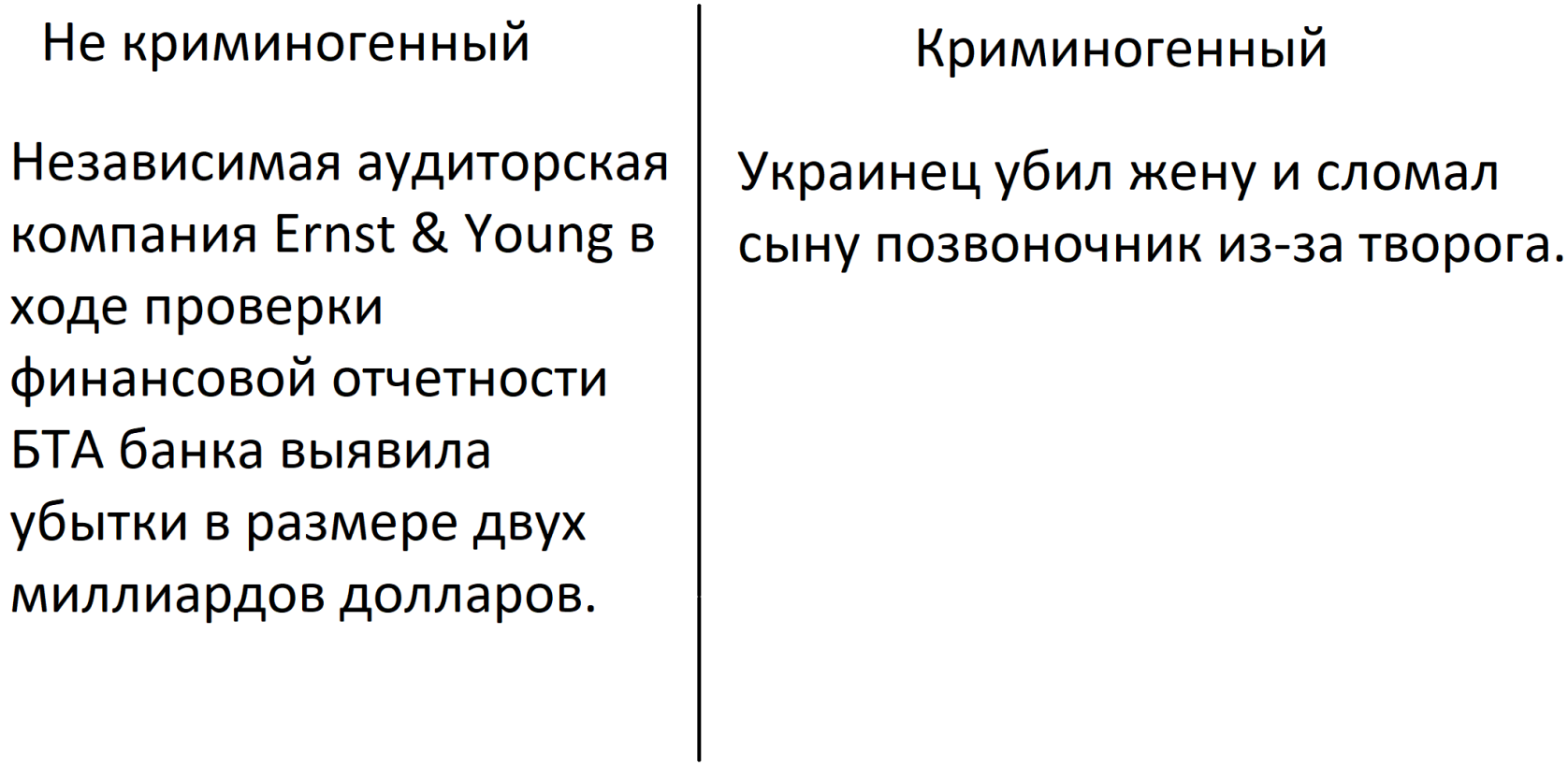

Рисунок 6. Пример криминогенного и не криминогенного текста

Так же одним из плюсов словарного метода, по сравнению с некоторыми другими методами машинного обучения является отсутствие нужды в создании датасета, то есть размеченного набора данных, используя который обычно обучается модель классификации. Создание датасета достаточно хорошего качества для точной классификации требует значительных человеческих ресурсов, т.к. разметку отдельных текстов на криминальный / не криминальный, необходимо будет делать вручную, заранее прочитав текст и определив его принадлежность к криминогенному классу, также количество текстов, которые нужно разметить должно быть довольно большим и измеряться тысячами или даже десятками тысяч текстов.

Также состав такого размеченного датасета должен быть достаточно разносторонним, то есть, иметь в себе не только тексты, связанные с судебными разбирательствами в русле криминогенности, но и криминогенные тексты другого рода, для того, чтобы модель смогла обучиться на более полных данных, в противном случае есть риск натренировать модель на поиск только криминогенных текстов в русле судебных тяжб, что она конечно сможет делать, но уловить криминогенные тексты в которых нет упоминания судебных дел, с большой вероятностью, не сможет, потому что ранее она никогда не видела таких или похожих текстов.

Словарь, в свою очередь тоже требует усилий для его сбора, однако для того, чтобы собрать подходящий словарь требуется намного меньше усилий, так как собрать слова, относящиеся к какой либо теме легче, чем читать текст и решать относится ли прочитанный текст к нужному классу или нет, когда количество текстов может быть более десяти тысяч. Количество слов исходного словаря для построения модели классификации может не превышать и трёх сотен, а после процедуры автоматического увеличения исходного словаря можно получить криминогенный словарь размером до трёх тысяч слов.

Дальнейшие исследования могут быть направлены на апробацию традиционных методов машинного обучения и дальнейшее сравнение словарного метода классификации и традиционных методов классификации текстов. Под традиционными методами понимается решение задачи бинарной классификации с помощью различных алгоритмов машинного обучения, например логистической регрессии или регрессии с градиентным бустингом. Для обучения таких моделей набор текстов будет превращен в так называемый мешок слов, который представляет собой матрицу, где строки — это тексты, а столбцы — это все слова, встречающиеся во всем наборе текстов. Заполнение матрицы в мешке слов происходит путем подсчета слов в конкретном тексте, количество конкретного слова в тексте будет записано в соответствующую ячейку матрицы, таким образом модель машинного обучения получит возможность выявить некие шаблоны, присущие текстам разных тематик. Так же датасет должен содержать в себе целевой столбец, содержащий информацию о том, является текст криминогенным или нет. На основании выявленных шаблонов, присущих криминогенным текстам, а так же шаблонов, которые

присущи всем остальным текстам, модель станет способна отделять криминогенные тексты от не криминогенных, что будет являться разновидностью бинарной классификации.

Пример датасета, состоящего из мешка слов можно увидеть в таблице 0.

Наличие качественных данных зачастую является решающим условием для достижения хорошего качества моделей машинного обучения. Шаги сбора, очистки и подготовки данных должны быть выполнены безошибочно, что позволит добиться чистоты исходных данных.

Таблица 0

Пример датасета из мешка слов

id	Мама	я	мыть	окно	рама	зверски	убивать	грабить	…	Криминоген?
0	1	0	1	0	1	0	0	0	…	0
1	0	1	1	1	0	0	0	0	…	0
2	0	0	0	0	0	1	1	1	…	1
3	0	1	1	1	0	1	0	0	…	0
4	1	1	1	0	1	0	0	0	…	0

1 Подход

1.1 Набор данных

Сбор необходимых данных является важной частью работы по анализу данных. Компании, которые так или иначе занимаются анализом данных вынуждены нанимать сотрудников, которые способны производить сбор данных из баз данных или из каких-либо открытых источников, очистку собранных данных и приведение их к единому формату, тем самым подготавливая данные к самому анализу. Данные и их чистота – это одни из самых важных вещей в анализе данных, ведь если производить анализ неточных и/или не приведенных к единому формату данных, то получить хоть какой-нибудь внятный результат из такого анализа будет невозможно.

Для решения поставленной задачи необходимо собрать новостные тексты, принадлежность которых к классу «криминогенный» будет определена. В результате опроса населения на предмет «какой Казахстанский новостной сайт самый популярный» был определен новостной портал Tengrinews [9]. Таким образом с упомянутого новостного портала был взят набор текстовых новостных материалов, размером в пять тысяч текстов за период с начала 2014 года по середину 2019 года. Набор данных, то есть датасет был получен с помощью использования технологии т.н. парсинга интернет-сайтов. Парсинг – это довольно популярная технология в области обработки данных, потому что, зачастую, получить нужные для анализа данные по запросу – невозможно.

Парсинг интернет-сайтов осуществляется с помощью различных библиотек языков программирования, которые способны получать HTML код интернет страницы, а также необходимую информацию из полученного кода, при указании на то, в каком именно параграфе или блоке согласно разметке HTML находится нужная информация.

Пример HTML кода, из которого можно получить заголовок новостной статьи приведен на рисунке 1, на котором голубыми областями отмечен заголовок новостной статьи и соответствующий ему раздел внутри HTML кода.

Рисунок 1. Пример получения заголовка новостной статьи из HTML кода

Парсинг – это довольно ресурсоемкое занятие в связи с тем, что во время парсинга происходит постоянное получение все новых и новых интернет страниц путем HTTP get запроса и получения соответствующей HTML страницы с сервера интернет-сайта, каждая такая итерация нагружает не только компьютер с помощью которого производится парсинг, но и сервер интернет-сайта, который парсят, а таких итераций нужно проделать тысячи или сотни тысяч. Например, основываясь на опыте автора данной работы, для получения 162 тысяч новостных текстов с довольно отказоустойчивого сервера интернет-сайта, имея в распоряжении около 32-ух ядер процессора потребовалось около 10 часов. Таким образом за 20 минут, имея в распоряжении 32 ядра процессора можно получить около 5400 новостных текстов, но среднестатистические компьютеры в наше время обладают только 8-ю ядрами, а это значит, что на парсинг 5400 текстов, имея 8 ядер уйдет около 80 минут. Это значит, что компьютер, с помощью которого производится парсинг будет недоступен для других задач все то время, что происходит парсинг, так как все ресурсы процессора будут заняты парсингом. Конечно же можно задать ограничение на использование ресурсов во время парсинга, уменьшив количество задействованных ядер, но в таком случае времени на парсинг потребуется еще больше. Обычно для такого рода задач используются серверы, внутри которых находится 64 и более ядер процессора, а также графические ускорители (видеокарты), использование которых для различных расчетов позволяет достигать высочайших скоростей, как парсинга, так и любых других компьютерных расчетов.

Часть собранного датасета представлена для ознакомления на рисунке 2.

	title	text	date	views	tags
0	Главный недостаток Путина	Премьер-министр России Владимир Путин дал инте...	2019-06-02	13.0	['Кризис, Путин Владимир']
1	Недропользователи попросили снизить налоги на ...	Казахстанская горнорудная корпорация ENRC обра...	2019-06-02	12.0	['ENRC, Кризис, Налоги']
2	Казахстанский предприниматель арестован за уст...	В Костанае частный предприниматель привлечен к...	2019-06-02	8.0	[]
3	Мажилис разрешил "Самрук-Казына" покупать обан...	Мажилис одобрил закон "О Фонде национального б...	2019-06-02	6.0	['Мажилис, Правительство Казахстана, Самрук-Ка...

Рисунок 2. Датасет Tengrinews

Как видно на рисунке 2, датасет состоит из столбцов: title – заголовок, text – текст, date – дата, views – количество просмотров и tags – теги новостных текстов, а также строк, в которых находится соответствующая столбцам информация. Основная работа будет проводиться со столбцом text.

Датасет необходимо подготовить к дальнейшей работе, а именно: разбить тексты на отдельные слова для того, чтобы из цельного объекта “Текст” выделить подобъекты, из которых состоит текст, то есть объекты “Слово”. Далее необходимо поставить каждое слово в каждом тексте в начальную форму, например Автомобилем – Автомобиль, это позволит объединить слова в разных склонениях в одно и то же слово и избежать искажения результатов исследования из-за отдельной обработки одинаковых слов в разных склонениях. Поставить слово в начальную форму, сохранив его смысл позволит морфологическая разметка, то есть определение частей речи слов, например “стекло – часть речи: существительное, начальная форма: стекло” и “стекло – часть речи: глагол, начальная форма: стекать (стечь). Правильная постановка слов в начальную форму может помочь избежать омонимии, то есть, случаев, когда слова схожи по написанию, но различны по смыслу, например в предложениях «стекло было разбито» и «молоко стекло по его усам» для полноценного понимания смысла предложений необходимо точно понять, какое значение слово «стекло» имеет в каждом из предложений. Это относит нас к понятию лексической неоднозначности. В работе [10] авторы предлагают использовать контекст, в котором находится неоднозначное слово для решения задачи разрешения омонимии. Таким образом понимание значения слов позволит производить более точную классификацию текстов.

Постановка слов в начальную форму и их морфологическая разметка были проделаны с помощью библиотеки UDpipe [11], также в работе [12] авторы этой библиотеки описывают процесс морфологической разметки, говоря о том, что обученная нейронная сеть имеет на входе слово, а на выходе несколько вариантов морфологической разметки для заданного слова, выбирается же тот вариант, который является наиболее частым правильным вариантом в соответствии с данными обучения нейронной сети. На рисунке 3 приведен пример того, как выглядят слова из текстов в датасете после постановки их в начальную форму и морфологической разметки. Таблица с расшифровкой морфологических тэгов представлена в таблице 1.

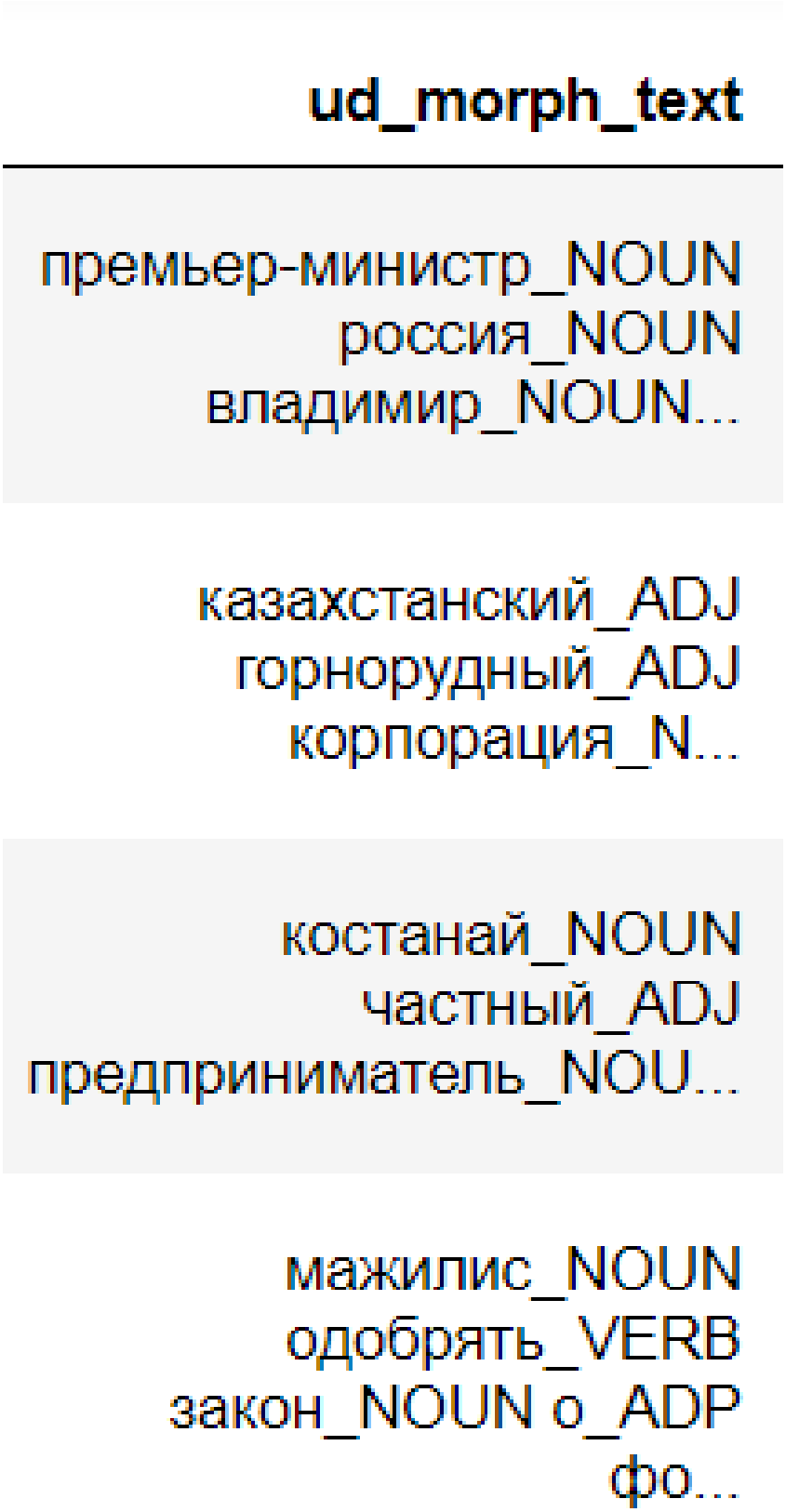

Рисунок 3. Морфологически размеченные слова в начальной форме

Таблица 1

Расшифровка морфологических тэгов

Тэг	На английском	На русском
ADJ	Adjective	Прилагательное
ADP	Adposition	Союз
ADV	Adverb	Наречие
AUX	Auxiliary	Вспомогательный (глагол)
CCONJ	Coordinating conjunction	Координационное соединение
DET	Determiner	Определитель
INTJ	Interjection	Междометие
NOUN	Noun	Существительное
NUM	Numeral	Цифра
PART	Particle	Частица
PRON	Pronoun	Местоимение
PROPN	Proper noun	Имя собственное
PUNCT	Punctuation	Пунктуация
SCONJ	Subordinating conjunction	Подчиненное соединение
SYM	Symbol	Символ
VERB	Verb	Глагол
X	Other	Другое

1.2 Криминогенный словарь

Криминогенный словарь был собран вручную на основании экспертного (субъективного) мнения автора данной работы о том, какие слова имеют отношение к нарушениям закона, например: подраться, кровопотеря, наручник, арестант, конфликт и т.п. На рисунке 4 представлен собранный криминогенный словарь, состоящий из 274 слов, который уже прошел ту же подготовку к работе, что и тексты из датасета.

Качество словаря играет важную роль в работе моделей, основанных на словарном подходе, если словарь будет включать в себя двусмысленные слова, то есть будет содержать слова, которые могут использоваться в разных тематиках с одинаковым успехом, то в конечном итоге такой словарь будет собирать довольно «грязный» класс текстов, то есть такой класс, в котором будут не только тексты искомой тематики, но и тексты, принадлежащие к другим тематикам. Для того, чтобы собирать «чистый» класс текстов, используя словарный метод – словарь должен быть тоже «чистым» - содержать в себе слова, которые принадлежат только искомой тематике.

	POS_tagged
0	подраться_VERB
1	хулиган_NOUN
2	штрафовать_VERB
3	садистский_ADJ
4	арестант_NOUN
...	...
269	кровопотеря_ADV
270	наручник_NOUN
271	душить_VERB
272	конфликт_NOUN
273	издевательство_NOUN

Рисунок 4. Криминогенный словарь, собранный вручную

1.3 Близость текстов

Для того, чтобы определить, насколько анализируемый текст криминогенный или является ли текст криминогенным вообще, необходимо получить возможность автоматически определять сходство между анализируемым текстом и криминогенным словарем, таким образом, что чем больше будет сходство текста и словаря, тем больше он будет относиться к криминогенному классу.

Сходство между текстом и словарем предлагается определять, используя меру сходства Жаккара. Мера Жаккара успешно применяется в различных исследованиях по обработке естественного языка, например в работе [13], где

мера использовалась для того, чтобы определить уровень сходства между тематическими словарями и новостными текстами, а также в оригинальной статье автора данной меры, Поля Жаккара по биологии [14] и представляет собой меру совпадения множеств, в нашем случае множеств слов из текстов и множества слов из криминогенного словаря. Формула меры Жаккара приводится ниже (формула 1).

$$Sim = \frac{|A \cap B|}{|A \cup B| - |A \cap B|} \quad (1)$$

где A и B множества слов. Мера принимает значения от 0 до 1, где 0 – это полная несхожесть и, соответственно 1 – это полное сходство множеств. Например, сходство двух предложений *"Владимир Путин посетил больницу в Коммунарке, куда направляют и лечат больных коронавирусом и поблагодарил главврача за работу."* и *"Президент РФ Владимир Путин посетил президента Казахстана Нурсултана Назарбаева и поблагодарил за укрепление двухсторонних отношений."* будет равна 0,28, тогда как для предложений *"Олег Саленко рассказал свою версию ситуации с составлением протокола за нарушение условий карантина."* и *"Такины относятся к семейству полорогих их ближайшие родственники – козлы, бараны, овцебыки."* мера будет равна 0.

Таким образом видя разницу в мерах сходства между двумя парами предложений можно сделать вывод о том, что первая пара состоит из более схожих предложений, а вторая состоит из абсолютно непохожих друг на друга предложений.

Одной из особенностей меры Жаккара является задача определения порога сходства множеств, то есть, согласно какому порогу меры, множества будут действительно схожи. Предложенный автором данной работы метод для определения рабочего порога сходства описан в разделе 2.

2 Основные эксперименты

2.1 Сходство новостных текстов и исходного криминогенного словаря

После определения мер близости между текстами датасета и исходным криминогенным словарем, состоящим из 274 слов, была построена гистограмма распределения мер близости в датасете (рисунок 1).

На рисунке 1, по оси X располагаются значения мер Жаккара, а по оси Y количество новостных текстов. Судя по первому прямоугольнику, который почти достигает отметки в 4 тысячи текстов и охватывает значения мер Жаккара от 0 до 0.00415 можно сделать вывод о том, что основная доля текстов находится в упомянутых пределах. Основываясь на мнении о том, что датасет не может состоять только из криминогенных текстов, а так же на экспертном оценивании содержания текстов, попавших в вышеупомянутый промежуток мер – можно с уверенностью заявить, что данный промежуток является неким хранилищем для тех текстов, которые не относятся к криминогенному классу. Таким образом, в пределах мер Жаккара от 0.00415 и до конца по оси X находятся тексты, на которые стоит обратить внимание.

После проделанной классификации новостных текстов по порогу сходства более или равному 0.00415 в криминогенный класс попали тексты, которые затем были проанализированы путем экспертного оценивания их содержания. Таким образом, было выявлено, что наряду с искомыми текстами, в класс попали тексты, которые были несколько далеки от того, чтобы иметь отношение к криминогенному содержанию, например, текст, в котором говорится о плохой работе аэропорта Алматы [15], в следствие чего было принято решение о том, чтобы сместить порог сходства на одну ступень вправо согласно рисунку 1 и установить порог близости равным 0.00830, в результате чего удалось получить достаточно чистый криминогенный класс новостных текстов размером 779 из 5000 текстов.

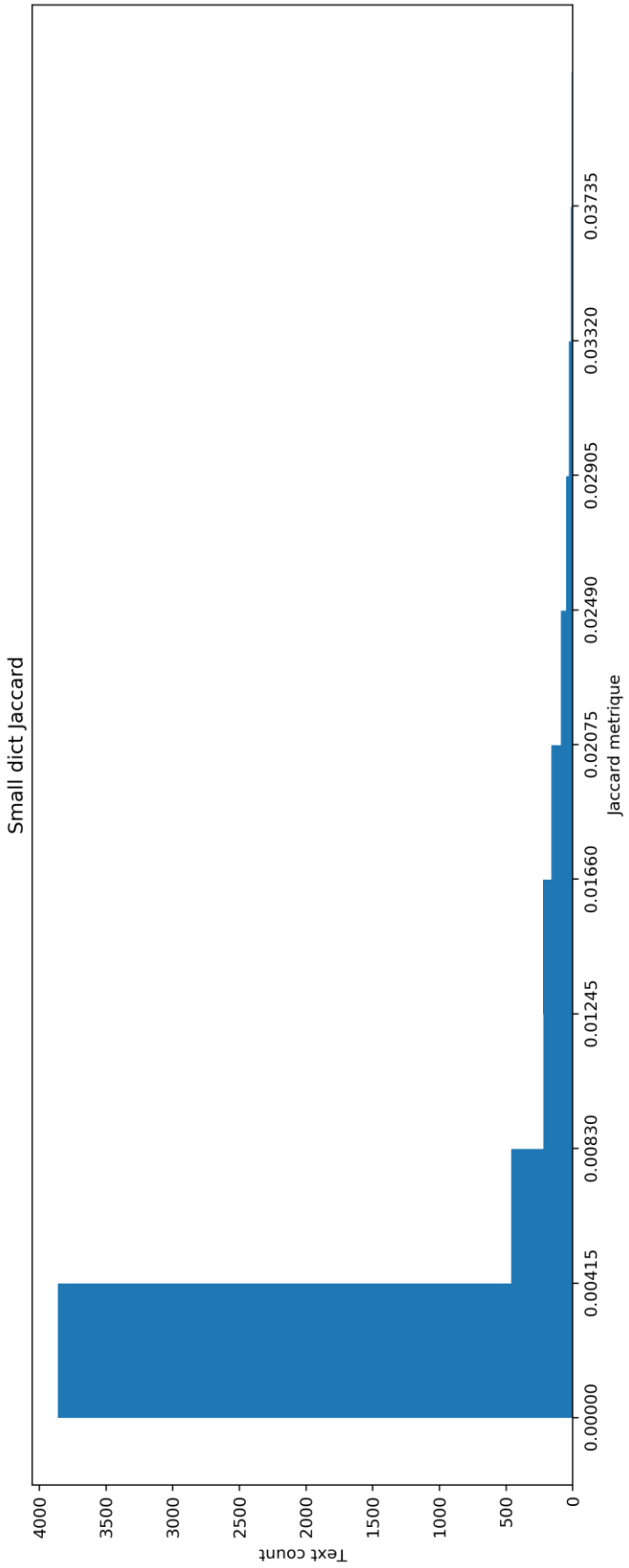

Рисунок 1. Распределение мер Жаккара между новостными текстами и исходным криминогенным словарем

2.2 Сходство новостных текстов и увеличенного криминогенного словаря

С целью проверки теории о том, что используя криминогенный словарь большего размера удастся найти большее количество соответствующих словарю текстов, был проведен эксперимент по автоматическому увеличению криминогенного словаря, с помощью подхода Word2vec [16]. Подход представляет собой вычисление векторного представления слов, то есть координат слов в векторном пространстве, основываясь на присутствии слова в контексте других слов. Слова, контекст которых схож, будут иметь близкие векторы и наоборот. Таким образом превращение слов в их векторные представления позволяют производить со словами различные операции (рисунок 2), а также вычислять близость между ними.

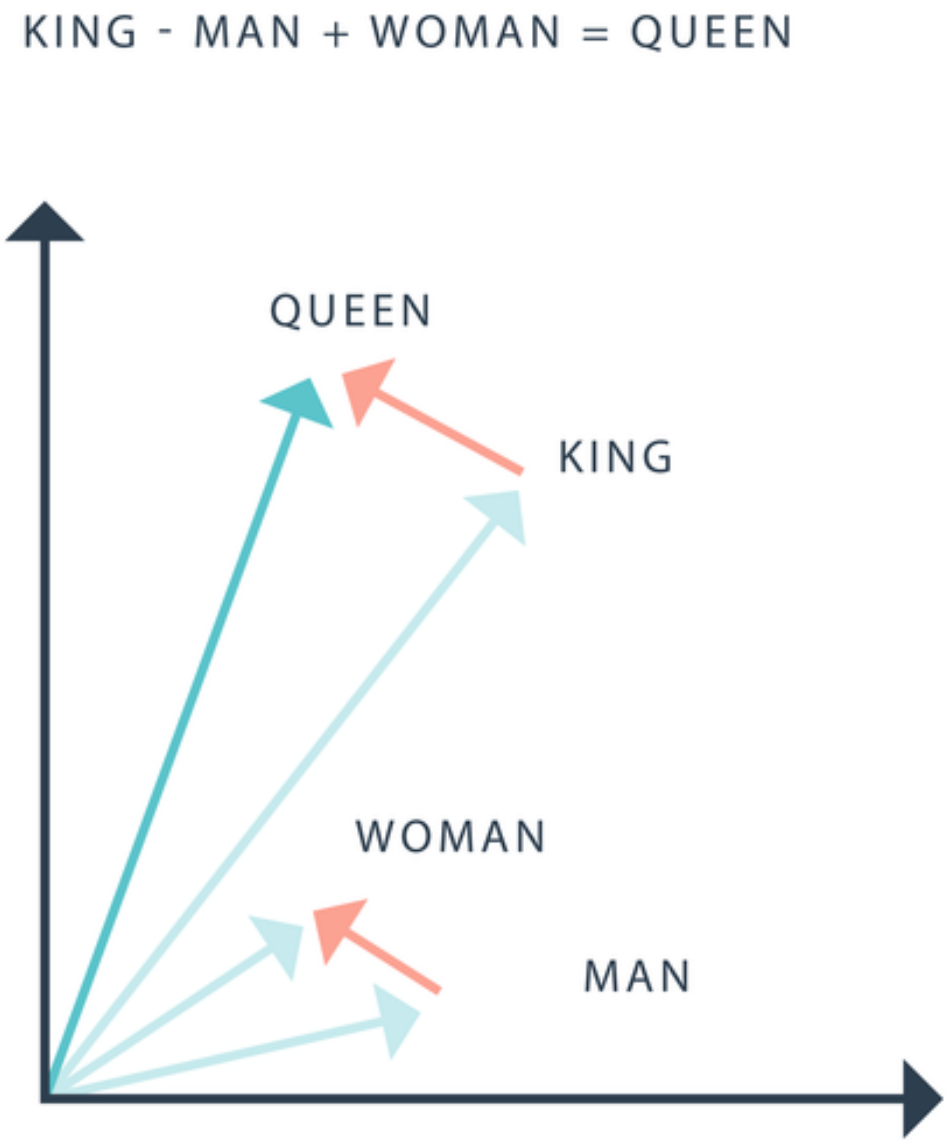

Рисунок 2. Пример операций между векторами [17]

На рисунке 2 можно увидеть довольно популярный пример того, как отняв от вектора KING вектор MAN и прибавив к тому, что получилось вектор WOMAN можно получить вектор QUEEN.

Модель Word2vec является нейронной сетью и перед началом работы её необходимо обучить. Алгоритм обучения модели Word2vec состоит в том, чтобы на основании контекста конкретного слова определить его векторное представление, такой алгоритм называется Continuous bag of words (CBOW). Контекстом является некоторое одинаковое количество слов до и после конкретного слова, на рисунке 3 представлен алгоритм обучения, контекстное окно которого равно двум, то есть два слова до и два слова после конкретного слова создают его контекст.

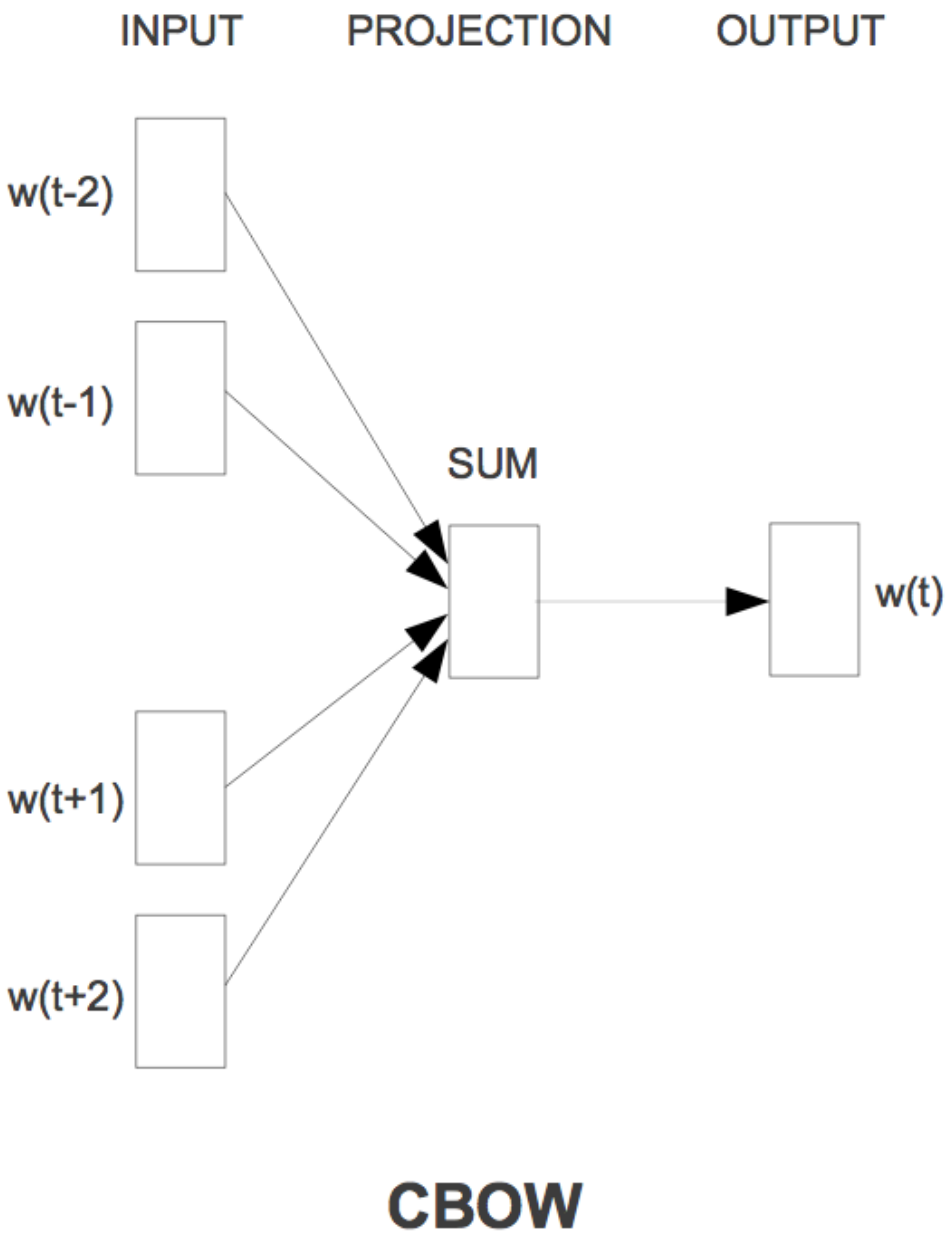

Рисунок 3. Алгоритм обучения CBOW

После обучения на огромном наборе текстовых документов (около 800 миллионов слов [18]), модель становится способной выдавать схожие по контексту слова к заданному слову, при условии, что ранее модель «видела» заданное слово во время обучения. Такой подход позволяет «машине» улавливать некоторый смысл слов, объединяя схожие по контексту слова в соответствующие группы.

Схожесть слов определяется вычислением косинуса угла между двумя векторами слов, чем больше косинус угла между двумя векторами, тем они ближе друг к другу, такая близость называется косинусной мерой или косинусной близостью. Таким образом, схожими словами со словом “Преступление” будут слова, представленные на рисунке 4, где также можно увидеть числовые значения косинусной близости.

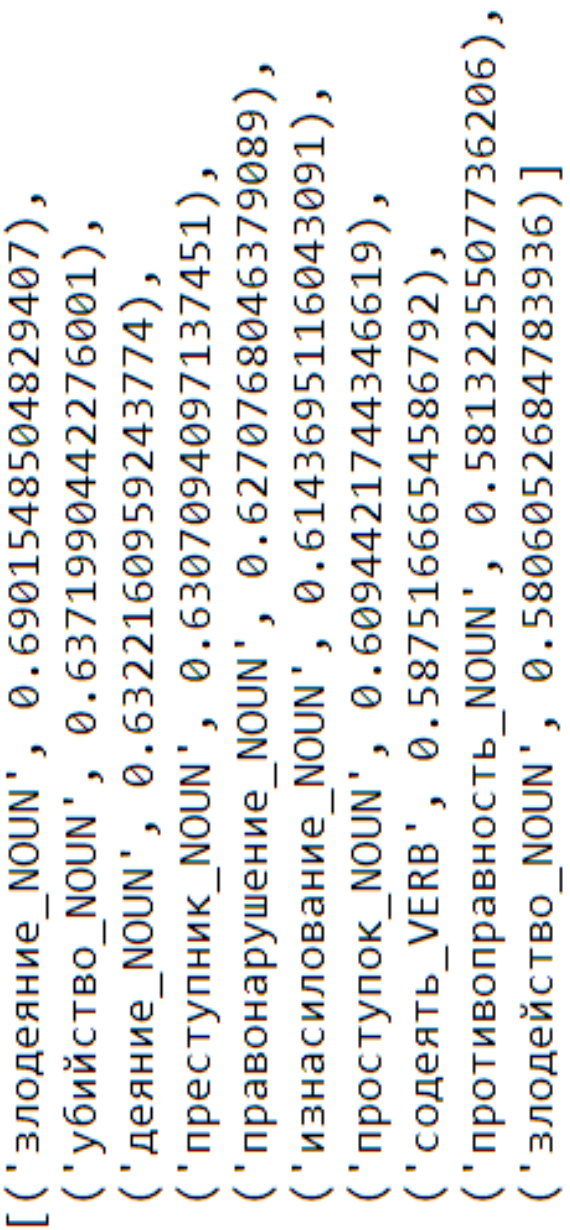

Рисунок 4. Десять схожих по контексту слов к слову “Преступление”

Как видно на рисунке 4, представленные моделью слова действительно находятся в одном контексте со словом “Преступление”.

В результате исходный криминогенный словарь был увеличен моделью Word2vec, которая обучалась на Википедии и национальном корпусе русского языка, получая по 30 дополнительных слов к каждому слову исходного словаря, затем словарь был очищен от повторяющихся слов. В итоге размер криминогенного словаря удалось увеличить с 274 до 3764 слов.

Повторив эксперимент из раздела 2.1, используя увеличенный словарь и порог равный 0.0025, в криминогенный класс попало 1223 текста, что заметно больше, по сравнению с 779 текстами из исходного словаря. Как видно на рисунке 5, на котором представлены распределения мер сходства между новостными текстами из датасета и увеличенным криминогенным словарем, если взять порог начала второй ступеньки на гистограмме, то есть взять порог равным чуть менее 0.0050, то в криминогенный класс попадет около 250 текстов, что намного меньше 779 текстов, полученных благодаря исходному словарю – соответственно порог второй ступеньки брать не стоит. В результате очередного экспертного оценивания содержания класса выявлено, что, в основном, в класс попали неподходящие новостные тексты. В следствии анализа причин такого поведения модели выявлено, что в словарь попали неподходящие к тематике криминогенного класса слова, например: сенситивный, тоскливость, пеня, квартплата и т.п. в связи с чем было решено вручную очистить увеличенный словарь от неподходящих слов.

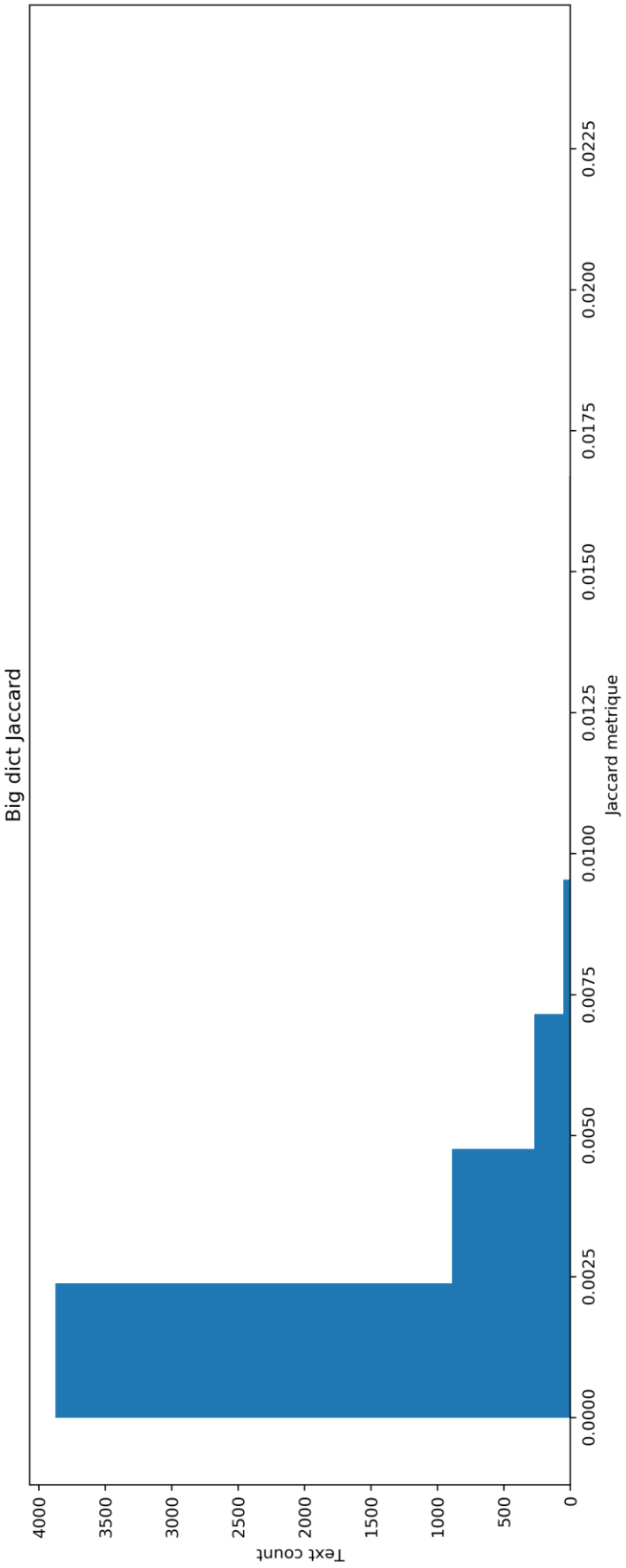

Рисунок 5. Распределение мер Жаккара между новостными текстами и увеличенным криминогенным словарем

2.3 Сходство новостных текстов и увеличенного-очищенного криминогенного словаря

В результате проведения процедуры очистки увеличенного словаря от неподходящих слов размер увеличенного-очищенного словаря стал равен 2286 словам.

После повтора эксперимента из пункта 2.1 с увеличенным-очищенным словарем и порогом меры Жаккара равным 0.0026 (рисунок 6), в криминогенный класс попали 1089 текстов. Заметно, что текстов стало меньше, чем при использовании не очищенного увеличенного словаря, однако в результате экспертного оценивания содержания криминогенного класса выявлено, что уточнение порога Жаккара до второй ступеньки, то есть до порога 0.0050 не требуется, так как криминогенный класс получился достаточно чистым.

Таким образом, можно сделать предварительные выводы о том, что увеличенный-очищенный криминогенный словарь способен наполнять криминогенный класс 1089 текстами по сравнению с исходным словарем, который наполнял криминогенный класс 779 текстами из набора в 5000 текстов. Пример экспертного оценивания содержания криминогенных классов приведен в разделе 2.4 данной работы.

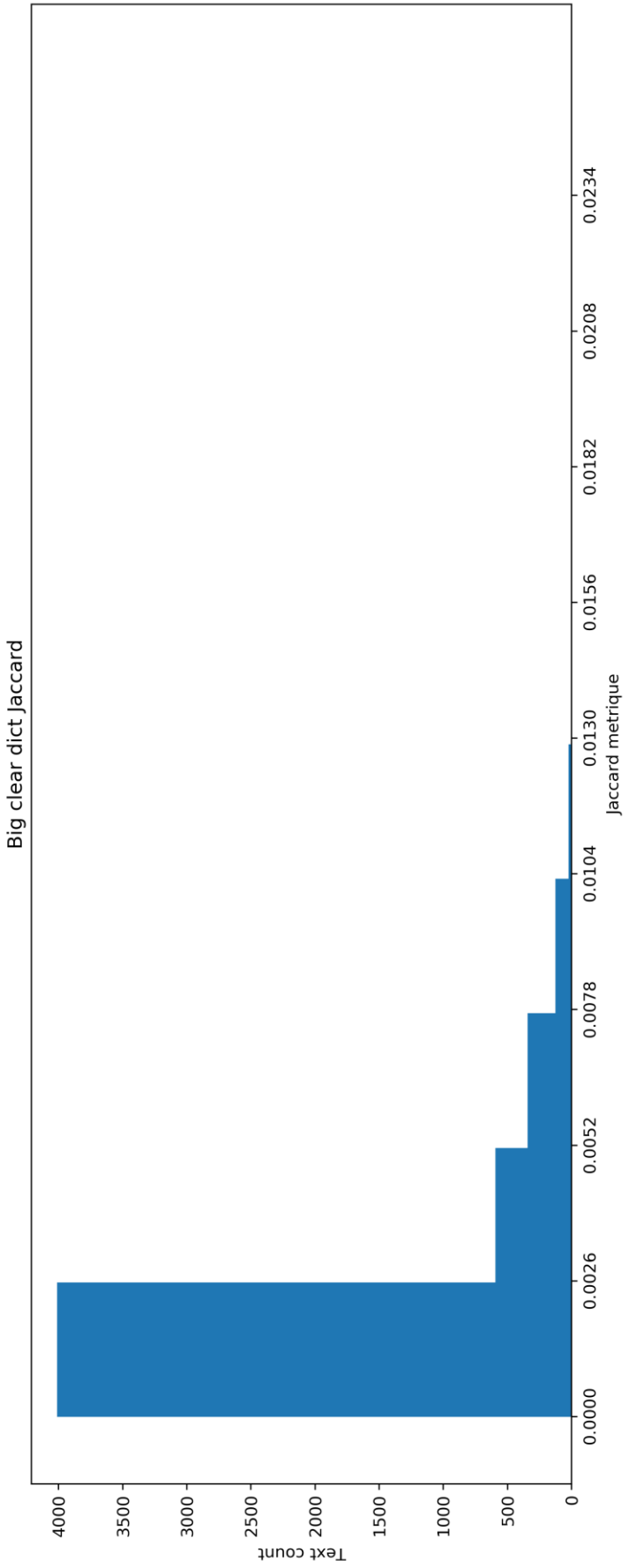

Рисунок 6. Распределение мер Жаккара между новостными текстами и увеличенным-очищенным криминогенным словарем

2.4 Эксперименты по оценке криминогенных классов

Для того, чтобы оценить чистоту полученных криминогенных классов было решено провести эксперимент. Суть его заключалась в том, чтобы оценить содержание и сравнить размеры криминогенных классов, полученных из трех видов криминогенного словаря: исходного, увеличенного грязного и увеличенного-очищенного. В таблице 2 приводится размер криминогенных классов, полученный благодаря использованию соответствующего словаря.

Таблица 2

Размер криминогенных классов

Метка класса	**Словарь**	**Размер криминогенного класса (кол-во текстов)**
A	Исходный	779
B	Увеличенный грязный	1223
C	Увеличенный-очищенный	1089

В дальнейшем классы, основанные на трех разных словарях, будут упоминаться по их метке класса, согласно таблице 2. В таблице 3 приведено количество совпадающих текстов между тремя парами сравнений криминогенных классов: A (исходный) против B (увеличенного грязного), A (исходный) против C (увеличенного-очищенного) и B (увеличенный грязный) против C (увеличенного-очищенного).

Таблица 3

Количество совпадающих текстов между криминогенными классами

Пара	**Кол-во совпадающих текстов**
A vs B	730
A vs C	738
B vs C	1017

Основываясь на данных из вышеприведенных таблиц, можно сделать следующие выводы:

1) Согласно таблице 3 количество совпадающих текстов в классах B и C равно 1017 текстам. Выходит, что класс B содержит еще 206 текстов, которых нет в классе C. Проведя анализ двадцати случайно взятых текстов из упомянутых 206 текстов, было выявлено, что только 5 из 20 текстов можно назвать

криминогенными. Исходя из вышесказанного, можно предположить, что около 75% текстов, благодаря которым класс B оказался больше класса C не несут в себе криминогенного содержания. На рисунке 7 представлены заголовки проанализированных двадцати текстов.

	dirty_vs_clear	ids
0	Главный недостаток Путина	0
1	Газпром подсчитал сумму ущерба	48
2	Нур-медиа сообщил подробности приобретения "Ли...	124
3	Палеонтолога-любителя обвинили в краже динозавра	170
4	Генпрокуратура потребовала вернуть казахстанск...	276
5	Бывшему спецназовцу присвоили пожизненную инва...	285
6	Тенге обесценится на 10 процентов	292
7	Новый сезон и старые амбиции	328
8	В СИЗО Тараза опровергают факт голодовки редак...	358
9	ОБСЕ призывает освободить Есергепова	383
10	Российский бизнесмен застрелился из-за финансо...	414
11	650 человек прокомментировали видеоблог Медведева	447
12	Безопасность на нигерийских дорогах обеспечива...	485
13	Биржевики подвели итоги 2008 года	564
14	С "Яблока" и "Патриотов России" взыщут деньги ...	716
15	Путин обошел Медведева по политической активности	761
16	Газпром оспорит действия Украины в международн...	786
17	В Узбекистане заблокировали Live Journal	833
18	В Алматы машинам запретят останавливаться ближ...	836
19	В Израиле штрафуют за перевозку непристегнутых...	878

Рисунок 7. Двадцать из 206 текстов, благодаря которым класс B больше класса C

2) Согласно таблице 3 количество одинаковых текстов в классах A и C равно 738 текстам, тем самым выходит, что класс C содержит в себе еще 351 текст, которых нет в классе A. Проанализировав двадцать случайно взятых текстов из 351, выяснилось, что в этот раз 15 из 20 текстов являются криминогенными. На основании этого можно предположить, что около 75% текстов, из тех, благодаря которым класс C оказался больше класса A несут в себе криминогенное содержание. В результате, благодаря автоматическому увеличению исходного словаря удалось достигнуть качественного и количественного увеличения криминогенного класса, основанного на увеличенном-очищенном словаре (C). На рисунке 8 представлены 20 случайно взятых текстов из 351.

	clear_vs_small	ids
0	Казахстанский предприниматель арестован за уст...	2
1	Великий и ужасный Федор Емельяненко	26
2	Журналисты Казахстана выступили против контрол...	41
3	Доходы американской секс-индустрии достигли 9,...	64
4	В Казахстане готовится соглашение по запрету о...	125
5	В Казахстане строительные фирмы воровали деньг...	134
6	В интернете откроется "школа блоггеров"	148
7	В Москве гражданин Армении бросился под поезд	198
8	В Павлодаре мужчина взорвал себя	237
9	Петербуржец подал в суд на организаторов конце...	267
10	Газета "Тасжарган" оспорит решение суда	280
11	Хакеры взломали сервер армии Великобритании	312
12	Журналист заплатит 1,5 миллиона тенге за оскор...	329
13	В Дагестане ввели режим контртеррористической ...	340
14	Семья Осборнов выиграла дело против таблоида T...	348
15	Мать отсудила у детского сада 80 тысяч рублей	393
16	Пассажиры разбившегося Ми-8 имели разрешение н...	395
17	Химическое оружие впервые применили древние персы	396
18	Турист из России избил итальянского чемпиона	416
19	Пассажиры упавшего МИ-8 охотились на краснокни...	432

Рисунок 8. Двадцать из 310 текстов, благодаря которым класс увеличенного чистого словаря оказался больше класса исходного словаря

2.5 Эксперимент с тысячей текстов

С целью выявления качественной разницы между криминогенными классами, полученными благодаря исходному (A) и увеличенному-очищенному (C) словарям, было решено запустить построенную модель классификации на одной тысяче новостных текстов, которые ранее не были использованы для экспериментов. Повторив шаги очистки и подготовки текстов к работе, были также повторены шаги, описанные в разделе 2.1 данной работы.

В результате использования словаря A удалось собрать в криминогенный класс 401 текст из тысячи, используя порог первой ступеньки гистограммы равный 0.0045. Использовав словарь C, по порогу первой ступеньки равному 0.00135, удалось собрать 505 текстов из тысячи. В таблице 4 приводится размер криминогенных классов, полученных благодаря словарям A и C. В таблице 5 приведено количество совпадающих текстов между криминогенными классами по словарям A и C.

Таблица 4

Размер криминогенных классов по словарям A и C

Метка класса	**Словарь**	**Размер криминогенного класса (кол-во текстов)**
A	Исходный	401
C	Увеличенный-очищенный	505

Таблица 5

Количество совпадающих текстов между криминогенными классами по словарям A и C

Пара	**Кол-во совпадающих текстов**
A vs C	390

Сначала могло показаться, что словарь C дал увеличение криминогенного класса на 104 текста, но после анализа части текстов, попавших в криминогенный класс благодаря словарю C, стало ясно, что большая их часть не совсем подходит под определение криминогенности.

На рисунке 9 можно ознакомиться с вышеупомянутыми текстами, которые были добавлены в криминогенный класс благодаря использованию словаря C, которых, в то же время не имеется в криминогенном классе, полученном благодаря словарю A, таким образом, получается, что это одни из тех текстов, благодаря которым криминогенный класс словаря C был больше криминогенного класса словаря A.

	1000_exp_clear_vs_small	ids
0	Россия защитит границы Абхазии и Южной Осетии ...	21596
1	В Киевском зоопарке зебра сломала шею о забор	21598
2	В Москве ликвидируют торговые точки на остановках	21620
3	В Москве безбилетник ранил из пистолета контро...	21623
4	Прокуратура "забраковала" движение "Сапсанов" ...	21626
5	Тайгер Вудс признался в любовных связях со 120...	21630
6	Владимир Кличко начал переговоры о проведении ...	21650
7	В Астане загорелся строящийся торгово-развлека...	21665
8	Выяснили причину смерти найденного в Якутии ма...	21670
9	В Великобритании пчеловода убили собственные п...	21690
10	В России продолжат рассекречивать документы о ...	21700
11	Совет Федерации одобрил уточнения в новом зако...	21702
12	Актеру из "Грозовых ворот" пересадят глаз	21719
13	По приказу Медведева рассекретили документы по...	21724
14	Пожар уничтожил крыши двух башен Псковского кр...	21727
15	Жертвами ДТП в Королеве оказались военнослужащие	21728
16	По законам плей-офф	21733
17	ФАС лишит владельцев участков на Рублевке прав...	21756
18	Российским банкам разрешили закрывать счета бе...	21762
19	Россия продаст Украине газ через посредника	21763

Рисунок 9. Двадцать текстов, которыми криминогенный класс словаря С больше криминогенного класса словаря А

Вышеупомянутые результаты породили теорию о том, что во-первых, выборка текстов довольно мала и словарь А уже справился с работой, выловив столько криминогенных текстов, сколько было в таких узких количественных рамках, а во-вторых, что порог близости для словаря С был выбран неверно, из-за чего в криминогенный класс словаря С и попали неподходящие тексты.

Таким образом, было решено сдвинуть порог близости для класса С на порог второй ступеньки гистограммы, равный 0.00265. С порогами близости и распределением мер Жаккара для криминогенных классов по словарям А и С можно ознакомиться на рисунках 10 и 11.

В итоге, используя порог второй ступеньки, в криминогенный класс по словарю С попало 367 текстов, а не 505, как было при пороге первой ступеньки.

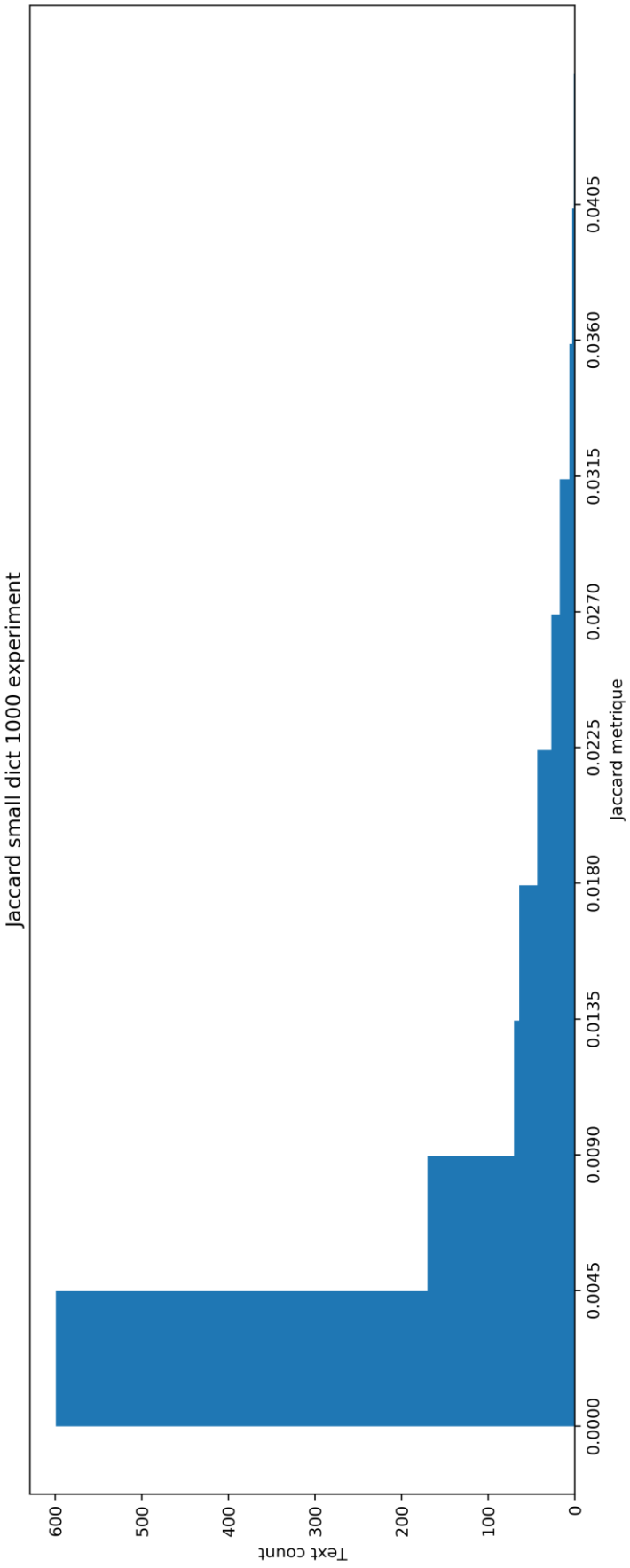

Рисунок 10. Распределение мер Жаккара между тысячей текстов и исходным словарем (A)

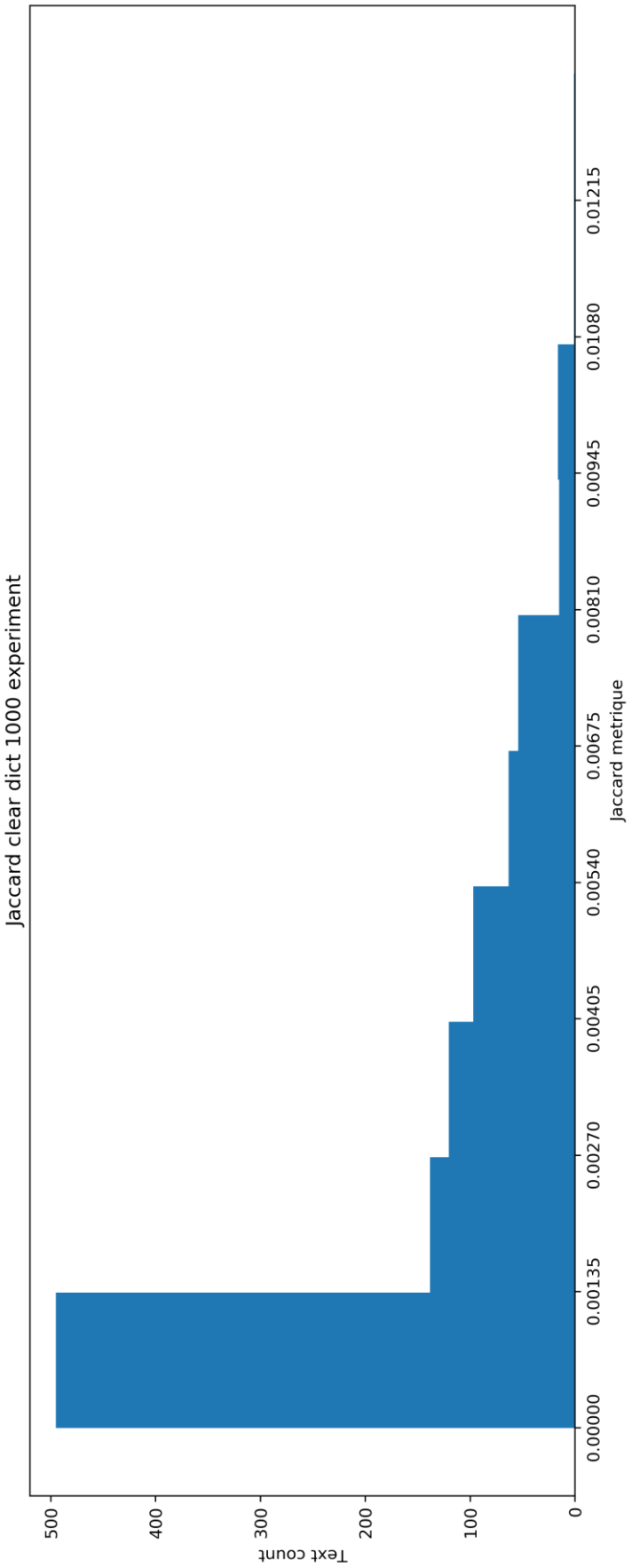

Рисунок 11. Распределение мер Жаккара между тысячей текстов и увеличенным-очищенным словарем (C)

Так как размер криминогенного класса по словарю C изменился, обновленная таблица 4 приведена ниже в таблице 6.

Проведя анализ совпадающих текстов между криминогенными классами по словарям A и C выяснилось, что теперь только 329 текстов вместо 390 имеются в обоих криминогенных классах (таблица 7), что порождает дополнительную теорию о том, что криминогенный класс по словарю A имеет около 72 неподходящих текстов. С целью подтверждения или опровержения этой теории был проведен анализ вышеупомянутых 72 текстов.

Таблица 6

Обновленная таблица 4 размер криминогенных классов по словарям A и C

Метка класса	**Словарь**	**Размер криминогенного класса (кол-во текстов)**
A	Исходный	401
C	Увеличенный-очищенный	367

Таблица 7

Количество совпадающих текстов между криминогенными классами по словарям A и C

Пара	**Кол-во совпадающих текстов**
A vs C	329

Анализ 72 текстов, которые предположительно являются неподходящими к криминогенному классу показал, что большая часть этих текстов является криминогенной (рисунок 12). Таким образом, сделан вывод в подтверждение теории о том, что выборка текстов слишком мала. Для определенной выборки текстов, размером в одну тысячу, достаточно использования исходного словаря (A) размером в 274 слова для того, чтобы выловить максимальное количество криминогенных текстов. В то время, как для выборки из 5 тысяч текстов использование увеличенного-очищенного словаря (C) было лучше (см. раздел 2.4 данной работы). Данные результаты говорят о том, что выборка из тысячи текстов оказалась довольно бедной на различные криминогенные слова, в результате чего, исходного криминогенного словаря (A) оказалось вполне достаточно. Естественно, что, чем больше выборка, тем более обширным и вариативным будет набор криминогенных слов, которые будут содержаться в текстах, а значит, что у исходного словаря (A) будет намного меньше шансов выловить все криминогенные тексты в отличие от увеличенного-очищенного словаря (C).

	1000_exp_small_vs_clear2nd	ids
0	В Иркутской области перевернулся пассажирский ...	21597
1	Мэрия Москвы попросит Daimler опровергнуть све...	21624
2	Банк ЦентрКредит указал на незаконность проверки	21638
3	Темиртауский суд не вернул банку ошибочно выда...	21639
4	Зенит обыграл ЦСКА и вышел на второе место в т...	21682
5	Google пригрозил засудить сайт Groggle из-за с...	21704
6	Финпол Казахстана начал внеплановую проверку "...	21707
7	В Алматы из-за сговора продавцов подорожали со...	21750
8	ФАС оштрафовала "ТНТ" за продакт-плейсмент	21758
9	Активистов "России молодой" задержали за пикет...	21775
10	Преподавателя МГУ поймали на взятке в 35 тысяч...	21816
11	На фальшивом сайте iTunes обнаружили вирус для...	21871
12	Амстердамский банк заберет имущество аэропорта...	21891
13	С 2011 года МВД РК перестанет заниматься техос...	21893
14	За гибель работника "СУАЛ-ПМ" оштрафовали на 5...	21895
15	Глава Goldman Sachs "слил" информацию о сделке...	21919
16	Шендерович закрыл свой блог в "Живом Журнале"	21937
17	В Новой Зеландии разбился военный вертолет	21945
18	В Санкт-Петербурге иномарка снесла остановку	21947
19	Грузия отпустила задержанный российский танкер	21949

Рисунок 12. 72 текста из криминогенного класса по словарю А, которых нет в криминогенном классе по словарю С

В дальнейшем, данный эксперимент был пересмотрен в отношении интерпретации результатов. Порог близости к увеличенному криминогенному словарю был взят по первой ступеньке, а не по второй, как было ранее, в результате было получено 505 текстов в криминогенный класс и 390 текстов схожих между двумя классами исходным и увеличенным. В итоге были проанализированы все 115 текстов, благодаря которым класс увеличенный оказался больше класса исходный, в результате анализа этих текстов выяснилось, что половина из них криминогенны, а вторая половина содержит тексты о футбольных и политических штрафах. Выводы эксперимента: увеличенный словарь все же дал больший криминогенный класс несмотря на то, что половина из 115 текстов оказалась не криминогенной, а так же исходный словарь должен содержать слова, которые принадлежат только одной тематике для того, чтобы после увеличения словарь так же содержал слова из одной тематики. Чистота исходного словаря является одной из тем дальнейших исследований.

3 Рассуждения, процент криминогенности и метрики оценки качества классификации

Стремительное развитие СМИ и их влияние на мысли общества, как отмечено во введении данной работы, подталкивает к тому, чтобы производить активный мониторинг СМИ стало необходимой частью работы по улучшению жизни людей. Если кто-либо говорит, что влияние СМИ на общество преувеличено, то автор данной работы считает, что влияние СМИ на общество недооценено и этому свидетельствует количество разных компаний, которые производят мониторинг СМИ (можно произвести поиск в google по предложению «мониторинг сми», чтобы подтвердить вышесказанное).

Известно, что людям свойственно сопереживание, таким образом, читая новости, где говорится о различных бедах, которые постигли других людей, человек, читающий это будет испытывать скорбь/сочувствие/печаль и различного рода переживания. На основании жизненного опыта автора данной работы переживания вредят здоровью человека, сначала психологическому, затем и физическому. Если человек окажется сломлен психологически, то вскоре он будет сломлен и физически.

В связи с вышесказанным создание фильтра “от всего негативного” будет являться чем-то вроде ограничительных мер по отношению к СМИ, которые, как известно, являются сильным инструментом воздействия на умы людей. С другой стороны, так как роль СМИ состоит в том, чтобы освещать события, происходящие в обществе – использование вышеупомянутого фильтра должно быть добровольным. Как и огонь, приносящий пользу людям, при правильном его использовании, а также приносящий людям погибель при неправильном его использовании.

Процент криминогенности.

Для того, чтобы сделать описанную в данной работе модель полностью автоматизированной понадобится автоматизировать вычисление верного порога близости между поданными модели текстами и криминогенным словарем. Это связано с тем, что мера Жаккара зависит от количества слов в поданных на ее вход текстах и количества слов в криминогенном словаре. Так как количество слов в поданных модели текстах будет постоянно меняться, то и результат близости по мере Жаккара будет меняться.

Идея вычисления постоянного порога близости текстов к криминогенному словарю заключается в том, чтобы вычислить, какой минимальный процент криминогенные слова составляют в текстах, которые были отнесены к классу криминогенных. На рисунке 1 представлен датасет, состоящий из 1089 криминогенных текстов, которые были получены из выборки размером в 5000 новостных текстов.

	text	ud_morph_text	ids	percents
0	В Костанае частный предприниматель привлечен к...	костанай_NOUN частный_ADJ предприниматель_NOU...	2	0.130
1	Потеряв работу, житель города Уилмингтон в Кал...	терять_VERB работа_NOUN житель_NOUN город_NOUN...	6	0.096
2	На Украине, в селе Селидово, мужчина зверски ...	украин_ADV село_NOUN селидово_ADV мужчина_NO...	12	0.091
3	В последнее время в Казахстане серьезно услож...	последний_ADJ время_NOUN казахстан_NOUN серь...	24	0.098
4	Офицер-контрактник из Костанайского гарнизона ...	офицер-контрактник_NOUN костанайский_ADJ гарн...	25	0.126
...	...	...	...	...
1084	Независимая аудиторская компания Ernst&Young в...	независимый_ADJ аудиторскать_NOUN компания_NOU...	4983	0.044
1085	К председателю КНБ Амангельды Шабадарбаеву и г...	председатель_NOUN кн_NOUN амангельды_NOUN шаб...	4984	0.088
1086	Глеба Агеева и его сестру Полину перевели в де...	глеб_NOUN агеев_NOUN сестра_NOUN полина_NOUN...	4986	0.110
1087	Автобус с детьми перевернулся на юго-западе Ки...	автобус_NOUN ребенок_NOUN переворачиваться_VE...	4992	0.082
1088	Пострадавшие во время стрельбы в супермаркете ...	пострадавший_NOUN время_NOUN стрельба_NOUN с...	4998	0.146

Рисунок 1. Проценты содержания криминогенных слов в текстах

В столбце percents на рисунке 1 представлены проценты содержания криминогенных слов для каждого текста из 1089. Текст под номером 2, где говорится про мужчину из села на Украине является очень криминогенным по мнению автора данной работы, в этом тексте описаны настоящие зверства, содержание этого текста не будет приведено в данной работе во избежание нанесения вреда читающему. Процентное содержание криминогенных слов для этого текста 0.091, что является 9% из 100%.

Ознакомившись с рисунком 1, можно проследить некий шаблон, который показывает, что те тексты, которые имеют процент криминогенных слов менее 9, действительно являются не очень криминогенными. Например, текст под номером 1084, где говорится о махинациях в БТА банке, имеющий 4% криминогенных слов, не может сравниться по содержанию различных зверств с текстом под номером 2. Конечно, содержание информации о различных махинациях может делать текст криминогенным, но все же, если говорить о негативном влиянии на психологическое здоровье человека, то здоровье человека, прочитавшего текст под номером 1084 не должно пострадать, как оно скорее всего пострадало бы, прочитай он текст под номером 2.

Вышеописанные размышления приводят к тому, что называется тональностью в обработке текстов. Тональность, зачастую представляет собой эмоциональную окраску текста, то есть содержание эмоционально окрашенной лексики в тексте. Когда тональность высокая, это значит, что текст насыщен эмоциональной лексикой, когда тональность низкая, это значит, что текст вполне нейтрально повествует о чем-либо. Тональность также может быть направленной на какой-либо объект, упоминающийся в тексте, например в отзывах о ресторанах люди могут хорошо отзываться о каком-то конкретном ресторане и плохо отзываться о каком-то другом ресторане, такая задача очень полезна для бизнеса.

Наряду с вышеприведенным определением тональности может выступать и другое – тональность может определять эмоциональный отклик читателя текста. Если текст написан вполне нейтрально, но вызывает сильные негативные эмоции у большинства читателей этого текста, то такой текст будет обладать негативной тональностью и наоборот, если текст вызывает у большинства читателей сильные положительные эмоции, то такой текст будет обладать положительной тональностью. Таким образом автор данной работы делит понятие тональности на два вида: 1) высокая и низкая тональность (содержание эмоционально окрашенной лексики), 2) негативная и положительная тональность по отношению к читателю.

С помощью модели классификации текстов, которая описана в данной работе становится возможным создание модели, которая могла бы отлавливать тексты с негативной тональностью, описание которой приведено в абзацах выше. Таким образом конечный фильтр для читателей новостей стал бы полнее и качественнее, не концентрируясь только на криминогенных текстах.

Метрики оценки качества классификации.

В машинном обучении существуют определенные метрики, с помощью которых можно оценивать качество построенной модели классификации, но все они требуют точного знания об истинном количестве объектов каждого класса для того, чтобы оценить, насколько точно произведена классификация. Например, метрики Precision и Recall, показывающие точность и полноту классификации.

Точность определяется отношением правильно отмеченных объектов класса 1 к общему количеству отмеченных объектов, как принадлежащих к классу 1. Точность вычисляется по формуле 2:

$$precision = \frac{TP}{TP+FP} \quad (2)$$

где TP является количеством объектов, которые модель отметила, как принадлежащие к классу 1 и это оказалось верно; FP является количеством объектов, которые модель отметила, как принадлежащие к классу 1, но это оказалось неверно.

Полнота определяется отношением правильно отмеченных объектов класса 1 к общему количеству объектов класса 1 и вычисляется по формуле 3:

$$recall = \frac{TP}{TP+FN} \quad (3)$$

где FN является количеством объектов, которые модель отметила, как принадлежащие к классу 2 и это оказалось неверно, то есть эти объекты на самом деле принадлежали к классу 1.

В связи с вышесказанным, а также учитывая тот факт, что истинное количество криминогенных текстов неизвестно выводом будет являться невозможность использования такого рода метрик. Таким образом использование экспертного оценивания содержания криминогенных классов является единственным верным путем для оценки точности классификации модели, которая использует словарный подход в связи с отсутствием размеченного набора данных.

ЗАКЛЮЧЕНИЕ

В результате проделанной работы была проведена обработка (очистка, морфологическая разметка) текстовых данных из новостного потока Республики Казахстан. Вручную создан криминогенный словарь, состоящий из специфичных слов для криминогенного содержания, который затем был увеличен, с помощью автоматического-семантического расширения, используя алгоритмы машинного обучения на основании векторных представлений слов. Задействовалось параллельное программирование для ускорения расчетов. Предложен метод определения порога близости для меры близости Жаккара. Построена модель и предложен рабочий подход для классификации текстов для одного класса. Проведены эксперименты на выборках текстов разного размера с целью определения качества получаемых классов на основании исходного (A) и увеличенного-очищенного (C) словарей для каждой из выборок.

Используя полученную модель, при условии наличия дополнительных словарей тем, станет возможным производить мульти-классовую классификацию текстов.

Из пяти тысяч новостных текстов 1089 были классифицированы, как криминогенные, благодаря увеличенному-очищенному криминогенному словарю. Качество классификации было подтверждено экспериментами по оценке качества содержания криминогенных классов, полученных из трех криминогенных словарей.

Из одной тысячи новостных текстов, которые не были ранее использованы 401 текст был классифицирован, как криминогенный, благодаря исходному криминогенному словарю. Качество классификации было подтверждено экспериментами по оценке качества содержания двух криминогенных классов, полученных из двух криминогенных словарей (A и C).

Экспериментально доказано, что для большой выборки текстов (от пяти тысяч и более) больше подходит использование увеличенного-очищенного словаря, так как, чем больше набор текстов, тем больше различных криминогенных слов содержится в наборе данных. Для маленькой выборки текстов (тысяча) больше подойдет использование исходного словаря, так как количество различных криминогенных слов для небольшой выборки текстов будет не велико.

Дальнейшие исследования могут быть направлены решение следующих задач:

1. Создание достаточного чистого исходного словаря, который после увеличения не придется чистить от неподходящих слов. Это значит, что слова в исходном словаре должны быть односмысленны и достаточно специфичны для конкретной темы. К примеру, не использовать слово “яд”, так как близкими по контексту могут оказаться названия химических соединений.

2. Исследование качественной разницы классификации между подходом белого ящика, который описан в этой работе и подходом черного ящика, который

представляет собой обучение традиционных классификаторов на заранее размеченном наборе новостных текстов. Это значит, что будет необходимо собрать датасет, состоящий из криминогенных текстов, затем превратить такой набор данных в числовое представление каждого текста и подать полученный набор данных на обучение одного из классификаторов машинного обучения (например классификатор логистической регрессии), в результате произвести оценку качества классификации получившейся модели и сравнить ее с моделью, основанной на словарном методе. Это позволит более точно определить плюсы и минусы каждого из подходов классификации.

2.1 Разработанная в данной работе модель классификации может послужить инструментом для создания размеченного набора новостных текстов на криминогенный / не криминогенный классы, что значительно облегчит работу по созданию такого набора данных.

3. Если позиционировать модель, как средство фильтрования новостей на каждый день, например, если человек читает новости один раз в день утром, то все новости, собранные с утра прошлого дня до утра этого дня должны быть отфильтрованы так, чтобы криминогенные новости были удалены из списка этих новостей. Это значит, что необходимо сократить ручное вмешательство в работу модели, то есть найти верный и постоянный порог близости между собранными за 24 часа текстами и криминогенным словарем. Некоторые рассуждения и предложение нахождения процентного соотношения криминогенных слов в текстах приведены в разделе 3 данной работы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Кондратьев М. Е. Анализ методов кластеризации новостного потока //Тр. Восьмой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2006). —Ярославль. – 2006. – С. 108-114.
2. Полищук Ю. И. О негативном влиянии средств массовой информации на психическое здоровье //Независимый психиатрический журнал. – 2003. – Т. 1. – С. 62-65.
3. Phillips D. P. The impact of fictional television stories on US adult fatalities: New evidence on the effect of the mass media on violence //American journal of sociology. – 1982. – Т. 87. – №. 6. – С. 1340-1359.
4. Phillips D. P. The impact of mass media violence on US homicides //American Sociological Review. – 1983. – С. 560-568.
5. Padilla-Walker L. M., Coyne S. M., Collier K. M. Longitudinal relations between parental media monitoring and adolescent aggression, prosocial behavior, and externalizing problems //Journal of adolescence. – 2016. – Т. 46. – С. 86-97.
6. Часовских А. Обзор алгоритмов кластеризации данных //[Электронный ресурс]. URL: https://habrahabr.ru/post/101338. – 2010.
7. Tannam E. What are the benefits of white-box models in machine learning? //[Электронный ресурс]. URL: https://www.siliconrepublic.com/enterprise/white-box-machine-learning. – 2019.
8. Guidotti R. et al. A survey of methods for explaining black box models //ACM computing surveys (CSUR). – 2018. – Т. 51. – №. 5. – С. 1-42.
9. Tengrinews. Новостной портал Tengrinews. //[Электронный ресурс]. URL: https://tengrinews.kz/. – 2020.
10. Piedeleu R. et al. Open system categorical quantum semantics in natural language processing //arXiv preprint arXiv:1502.00831. – 2015.
11. Straka M. Ufal.UDpipe //[Электронный ресурс]. URL: https://pypi.org/project/ufal.udpipe/. – 2020.
12. Straka M., Hajic J., Straková J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing //Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). – 2016. – С. 4290-4297.
13. Мухамедиев Р. И. и др. Эксперименты по оценке средств массовой информации на основе тематической модели корпуса текстов //Cloud of Science. – 2020. – Т. 7. – №. 1. – С. 87-103.
14. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines //Bull Soc Vaudoise Sci Nat. – 1901. – Т. 37. – С. 241-272.
15. Tengrinews. Мажилисмен попросил Масимова навести порядок в Алматинском аэропорту. //[Электронный ресурс]. URL: https://tengrinews.kz/news/majilismen-poprosil-masimova-navesti-poryadok-almatinskom-2578/. – 2009.

16. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.
17. Shogry B. Making sense of messy bank data. //[Электронный ресурс]. URL: https://blog.plaid.com/making-sense-of-messy-data/. – 2017.
18. RusVectores. Обученные векторные модели. //[Электронный ресурс]. URL: https://rusvectores.org/ru/models/. – 2020.