# Отчет:

# Подход машинного обучения для анализа преступности. Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах

Мерембаев Тимур, Мұса Ақбота

**КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ имени К. И. Сатпаева**

*Институт информационных и телекоммуникационных технологий*

akbota.mussa.z@gmail.com

***Абстракт.*** *Для того чтобы лучше подготовиться к реагированию на преступную деятельность, важно понимать закономерности преступности. В нашем проекте мы анализируем данные о преступности из города Торонто, с использованием наборов данных полиции Торонто и наборов данных штата Канада. Целью данной статьи является использование методов машинного обучения для классификации криминального инцидента по типу в зависимости от его возникновения в заданное время и в определенном месте. Эксперимент проводится на базе данных, содержащих записи преступлений Торонто с 2014 по 2017 год. Для этой контролируемой задачи классификации использовались набор инструментов и методы такие как: Basemap, folium, geopandas, Анализ временных рядов (ARIMA), Random Forest Classifier, Gaussian Naive Bayes модели и тд. Была построена модель прогнозирования с использованием классификатора случайных лесов и модель смогла классифицировать преступления с точностью примерно* 65%.

*В этой работе мы также пытаемся решить проблему интеллектуального анализа текста и классификации его на позитивные, негативные и нейтральные для выявления преступлений, применяя анализ настроений в отношении названий и содержания статей. Анализ преступности является одним из наиболее важных видов деятельности большинства интеллектуальных правоохранительных организаций во всем мире. Анализ данных - это мощный инструмент, который можно эффективно использовать для анализа больших баз данных и получения важных аналитических результатов. В этой работе будут рассмотрены различные распространенные теории, лежащие в основе НЛП, и способы их использования для сбора мнений пользователей в газетных статьях. Результаты экспериментов с классификатором показывают, что классификатор способен выполнить классификации с точностью 77,8% ,*

*полученных из статей. Результаты этого исследования будут иметь большое значение для исследователей, добавив еще одну перспективу наивного Байеса в мнении майнинга, а также для правоохранительных органов выявление криминальной информации в статьях.*

***Ключевые слова:*** *анализ временных рядов, Arima, Random Forest, GaussianNB, анализ преступности, очистка веб-страниц, обработка естественного языка(NLP), машинное обучение, NLTK,*

## 1. Введение

Безопасность людей одна из самых важных обязанностей правительств во всем мире. Таким образом, главная цель здесь является снижение уровня преступности. Есть много преступлений типы инцидентов, таких как кражи со взломом, кражи, грабежи, транспортное средство преступления, убийства, вооруженная торговля, сексуальные преступления и международные преступления и др. Преступная деятельность присутствует во всех регионах мира, что сказывается на качестве жизни и социально-экономическом развитии. Таким образом, это вызывает серьезную озабоченность многих правительств, которые используют различные передовые технологии для решения таких проблем. Криминологический анализ, подотрасль криминологии, изучает поведенческую модель преступной деятельности и пытается выявить признаки такого события.

Несмотря на то, что ценная информация доступна в удобочитаемой форме в онлайн-газетах и электронных архивах, программные системы, которые могут извлекать соответствующую информацию и представлять эту информацию, являются недостаточными, и это представляет значительный интерес для исследователей в области извлечения информации [1]. Даже если поисковые системы могут быть использованы для запроса конкретной информации (например, культурные события в Окленде), эти результаты запроса не обеспечивают историческую перспективу (т. е. если есть 100 статей о культурных событиях, пользователю, возможно, придется прочитать все из них, чтобы получить некоторые идеи, такие как увеличение количества опер в городе). Хотя можно было бы вручную прочитать результаты и извлечь ценную информацию, этот процесс утомителен и подвержен ошибкам. Таким образом, эта работа направлена на "добывание" информации, доступной в онлайн-газетных статьях.

В этом исследовании используется набор данных из Торонто Police Service Data, который содержит данные о преступной деятельности в окрестностях города Торонто в течение 3 лет. Я использовал различные методы классификации, такие как случайный лес , наивный байесовский, линейная регрессия, Arima модель и тд. н.

В этой работе мы также пытаемся решить проблему интеллектуального анализа текста и классификации его на позитивные, негативные и нейтральные для выявления преступлений, применяя анализ настроений в отношении названий и содержания статей. Анализ данных - это мощный инструмент, который можно эффективно использовать для анализа больших баз данных и получения важных аналитических результатов. Например, новый иммигрант может захотеть сравнить различные города на основе уровня преступности или сравнить различные районы конкретного города, чтобы выбрать более безопасный. Путешественник может захотеть узнать, какие части конкретного города следует избегать. В настоящее

время эта информация не является легкодоступной для пользователей, но они могут быть получены из газетной статьи конкретного региона, как правило, сообщают о важных преступлениях.

## 2. Соответствующие работы

Многие исследователи уделяют большое внимание криминальной сети анализ, но некоторые из них предложили системы для преступника сетевой анализ для обработки арабского языка. В данном разделе количество научно-исследовательских работ, посвященных криминальной сети анализ и выявление преступлений рассматриваются. Эта литература обзор разделен на два раздела: Литература по преступности обнаружение и анализ преступной сети. В этом разделе представлен обзор соответствующей работы в области выявления преступлений. Он также предоставляет краткую информацию о двух методах, названных Распознавание сущностей (NER) и условные случайные поля (CRF), используемые в данной работе. Исследователи, работающие в области извлечения информации о преступлениях, использовали несколько методов. В частности, для этой цели исследователи использовали такие методы, как краудсорсинг, интеллектуальный анализ данных и машинное обучение. Проект WikiCrimes [2]. использует силу толпы, где люди сообщают подробности преступления в интернете, и другие пользователи могут использовать эту информацию для принятия решений. Однако ограничением такого подхода является сложность проверки подлинности опубликованных преступлений.

Исследователи изучили методы извлечения соответствующей информации из неструктурированных документов. Процесс извлечения информации из неструктурированных документов затруднен, поскольку он написан на естественном языке и структура документа заранее не известна (по сравнению со структурированными файлами, такими как базы данных). Однако была проделана большая работа по идентификации объектов (например лицо, место, организация) из неструктурированных документов в области обработки естественного языка. Часто вызывается как распознавание именованных сущностей (NER), это метод был показан, используемый во многих областях (для обзора см. [3]. Например, проект Coplink [4] исследователей из Университета Аризоны направлена на выявление информации о преступниках из полицейских отчетов. Он использует систему извлечения сущности, которая основана на методах ИИ, для автоматического выявления личностей преступников, а также для анализа преступных сетей с использованием кластеризации и блочного моделирования.

## 3. Анализ Набора Данных

Предоставленный набор данных имеет различные "функции", каждая из которых имеет различную релевантность. В этой главе мы перейдем к анализу этой базы данных и извлечению из нее полезной информации.
Каждая запись содержит следующую информацию:

- Даты: временная метка момента совершения преступления. Он находится в следующем формате: Y-m-d H: i: s. E. g.: 2015-05-13 23: 53: 00
- Категория: категория преступления. Например: ордера
- Описание: краткое описание преступления. Например: ордер на арест
- День недели: день недели, в который произошло преступление. Например: среда
- Район: район города, в котором было совершено преступление. Например: Северный
- Резолюция: краткое описание резолюции о преступлении. Например: "арест, арестован”
- X: широта места преступления. Например: -122.425891675136
- Y: долгота места преступления. Е. Г.: 37.7745985956747

Если мы нарисуем другую графику, мы увидим более четко, как распределяются данные

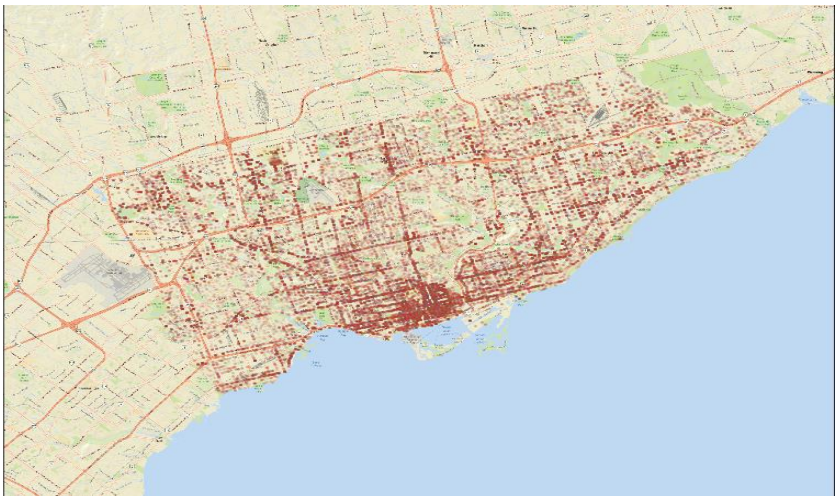


Рисунок 1

В наборе данных, мы анализируем 5 различных категорий преступлений, которые являются: нападение, взлом и проникновение, грабеж, мошенничество и угон автомобилей. На рисунке 2 видно, что в последние годы количество преступлений стало больше и это показывает растущую тенденцию.

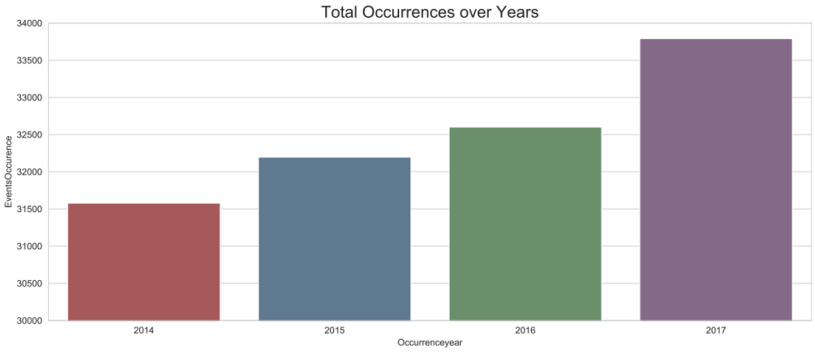



Рисунок 2 - Число крупных преступлений за последние годы

Еще одним интересным анализом было бы подсчет количества преступлений, которые происходят в каждый день недели, таким образом, мы могли бы увидеть, является ли это соответствующей информацией или нет. Рисунок 3 показывает, что пятница и суббота являются наиболее выбранным днем преступниками, в то время как остальная часть недели распределена довольно равномерно.

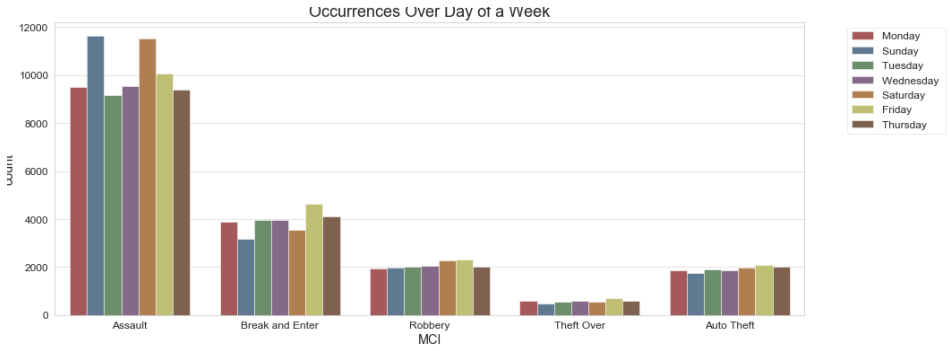
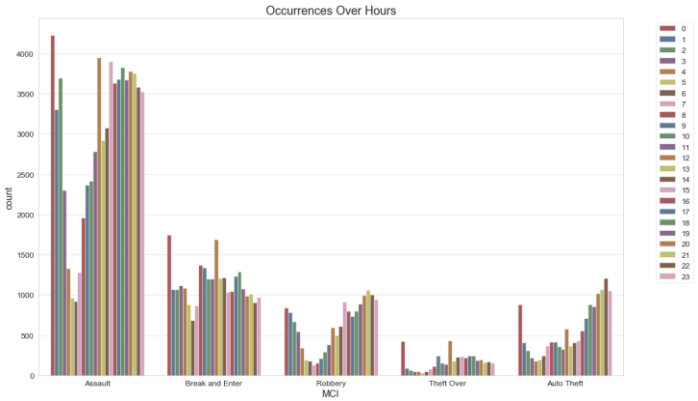



Рисунок 3 — Происшествия в течение дня недели

Мы также можем видеть распределение преступности по часам дня (Рисунок 4). Это свидетельствует о том, что большинство преступлений было совершено между полуднем и полуночью (включая и то, и другое).



## 4. Модели

**RF (random forest) -** это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

- Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления — свои случайные признаки).
- Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

**Гауссовский наивный Байес** - это контролируемый классификатор, который использует наивное предположение о том, что между двумя объектами нет зависимости. Этот классификатор реализуется путем применения Байеса

Теорема. Согласно теореме, класс и зависимый вектор признаков, состоящий из x0, x1, ..., xn, имеет следующее отношение:

P (y|x, x, ..., xn) = P (x0, x1,..., xn|y) ... (iv)

0 1 P (x0, x1,..., xn)

Эта вероятностная модель наряду с решающим правилом строит наивный байесовский классификатор. Существуют различные типы наивных байесовских алгоритмов классификации, основанных на распределении данных. В этой статье используется гауссовский наивный Байес, где предполагается, что данные распределяются в соответствии с гауссовским распределением.

## Performance Metrics

### Accuracy

В простейшем случае такой метрикой может быть доля документов по которым классификатор принял правильное решение.

$$Accuracy=P/N$$

где, P– количество документов по которым классификатор принял правильное решение, а N– размер обучающей выборки. Очевидное решение, на котором для начала можно остановиться.

**Confusion Matrix**

На практике значения точности и полноты гораздо более удобней рассчитывать с использованием матрицы неточностей (confusion matrix). В случае если количество классов относительно невелико (не более 100-150 классов), этот подход позволяет довольно наглядно представить результаты работы классификатора.

Матрица неточностей – это матрица размера N на N, где N — это количество классов. Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора. Когда мы классифицируем документ из тестовой выборки мы инкрементируем число стоящее на пересечении строки класса который вернул классификатор и столбца класса к которому действительно относится документ.

### 5.1 Gaussian Naive Bayes

Sklearn.naive_bayes provides GaussianNB class.

Хотя этот классификатор дает низкую точность, измерение потери журнала относительно лучше в этой модели. Анализ главных компонент дает плохой результат и в случае наивного Байеса.

### 5.2 Random Forest

Используются следующие параметры класса классификатора случайного леса: n_estimator, min_samples_split, criterion, где min_samples_split и criterion-параметры дерева решений случайного леса. N_estimator указывает количество деревьев для построения.

| **number_of_trees** | **Accuracy** | **Log loss** |
|---|---|---|
| 10 | 59.86% | 0.996 |
| 50 | 60.79% | 0.98 |

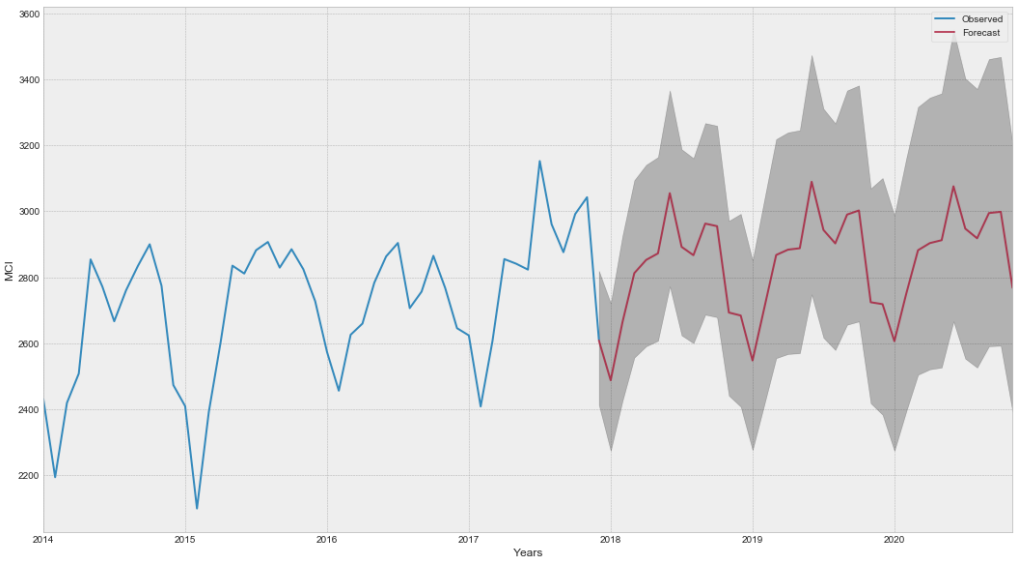
Table-9: Random Forest result

И точность и потеря улучшаются, в то время как количество деревьев в случайном лесу увеличивается. С большим количеством деревьев случайный лес имеет образцы для усреднения,что может уменьшить дисперсию.

## 6. Результаты

| **Algorithm** | **Accuracy** | **Log-loss** |
|---|---|---|
| Decision Tree Classifier | 56.79% | 1.054 |
| Gaussian Naive Bayes | 55.24% | 1.107 |
| K-Nearest Neighbor | 58.64% | 1.023 |
| Logistic Regression | 58.98% | 1.007 |
| Random Forest Classifier | 60.79% | 0.98 |

## Модель ARIMA




Анализ временных рядов используется для прогнозирования общего числа крупных преступлений в последующие годы. Серая заштрихованная область-это допуск для прогнозирования, синяя линия-это данные, которые мы использовали для прогнозирования, а красная линия-это прогноз.

## Заключение

Результат работы всех классификаторов увеличивается после использования меньшего количества классов. Значения потерь журнала улучшаются с большим запасом после уменьшения классов.

В ходе экспериментов можно увидеть, что несбалансированный набор данных был улучшен с помощью недостаточной выборки. Используя неполные данные, Random Forest успешно классифицировала преступную деятельность на основе времени и местоположения. С точностью 60,59% он смог превзойти другие алгоритмы машинного обучения. ARIMA - Прогнозирование преступности: преступность, скорее всего, будет расти, если не изменится политика полиции.

## 1. Веб скреперы (Web Scrapers)

Поскольку Интернет и Интернет продолжают расширяться, может стать затруднительным доступ к веб-страницам, не зная заранее, каков адрес страницы. Именно здесь появляются поисковые системы. Поисковые системы используют процесс, называемый веб-сканированием, который представляет собой алгоритм, предназначенный для сканирования или обхода коллекции веб-сайтов, который индексируется и ищется. Этот алгоритм имеет три основных компонента; веб-страница извлекается, она анализируется для извлечения всех связанных URL-адресов (Uniform Resource Locator), а для ранее невидимых URL-адресов повторите первые два шага. Веб-скребок отличается от веб-сканера, так как веб-сканер просто сканирует и индексирует, в то время как веб-скребок представляет собой автоматизированный инструмент, который запрашивает веб-сервер для получения веб-страницы и анализирует веб-страницу для извлечения информации. Пока веб-сканер сканирует множество веб-страниц, чтобы найти ссылки, Формат этих ссылок остается тем же, однако при использовании веб-скреперы для извлечения данных с веб-страниц формат разметки изменяется между различными веб-сайтами. Несмотря на это, веб-скребки часто используются для доступа к тем областям веб-сайта, которые не доступны поисковым системам.

## 2. Обработка естественного языка

Область НЛП восходит к нескольким десятилетиям и значительно выросла за эти годы. Изначально, благодаря появлению Всемирной паутины, собирающей данные из ограниченного набора оцифрованных документов, произошел взрыв информации на многих языках. Значительный объем работы был проделан в области поиска информации (IR), которая считается областью применения обработки естественного языка. Прежде чем обсуждать методы IR, давайте немного углубимся в теоретические и практические аспекты НЛП.

### 3.1 Традиционный подход - ключевые понятия

Первоначально подход НЛП следовал следующим дискретным шагам.

1.Текстовая предварительная обработка / токенизация
2.Лексический анализ
3.Синтаксический анализ
4.Семантический анализ

### 3.2 Методолгия исследования

Цель этой работы состоит в том, чтобы определить место кражи из корпуса и классифицировать каждое предложение в статье в приговоре о месте преступления (CLS), а не в приговоре о месте преступления (NO-CLS). В этом разделе описывается методология, используемая для классификации предложений в газетных статьях в

CLS и никаких-CLS предложений. На рис.1 показаны шаги, используемые в методологии, использованной в данной работе.

1. *Корпус building*- первым шагом является создание корпуса соответствующих газетных статей для нашего исследования. Для этого мы использовали BeautifulSoup Web Screen Scraper tool. BeautifulSoup - это веб-инструмент для извлечения конкретной информации с веб-сайтов. Если мы тренируем инструмент, указывая и нажимая на детали, которые должны быть извлеченный, инструмент может извлечь тот же набор информации автоматически из данного пула документов. Он также сохраняет извлеченную информацию в различных форматах для последнего использования. Мы создали корпус статей, связанных с воровством.

2. *Токенизация* предложений-при извлечении соответствующих газетных статей отдельные предложения должны быть извлечены (т. е. статья должна быть разделена на отдельные предложения). Мы использовали токенизатор Punkt из Nltk toolkit для этой цели.Токенизатор делит данную статью на список отправлений.

Методология, используемая для анализа настроений новостных статей в данной статье, основана на лексикон-ориентированном подходе. Анализ настроений обычно может проводиться с использованием контролируемых или неконтролируемых подходов. Контролируемый подход состоит из набора помеченных обучающих данных, которые используются для построения модели классификации с целью использования этой модели для классификации новых данных, для которых метки отсутствуют. Неконтролируемые или лексические подходы к анализу настроений не требуют каких-либо обучающих данных. В этом подходе чувства, передаваемые словом, выводятся на основании полярности слова. В случае предложения или документа полярности отдельных слов, составляющих документ, в совокупности передают смысл предложения или документа. Таким образом, полярность предложения-это совокупность (сумма) полярностей отдельных слов (или словосочетаний) .

Этот подход использует некоторые предопределенные списки слов, так что каждое слово в списке связано с определенным настроением. В дальнейшем этот подход может использовать следующие методы

1.Словарные методы: в этих методах словарь лексики используется для того, чтобы найти слова с положительным мнением и слова с отрицательным мнением.

2. Корпусные методы: в этих методах используется большой корпус слов и на основе синтаксических шаблонов другие слова мнения могут быть найдены в контексте.

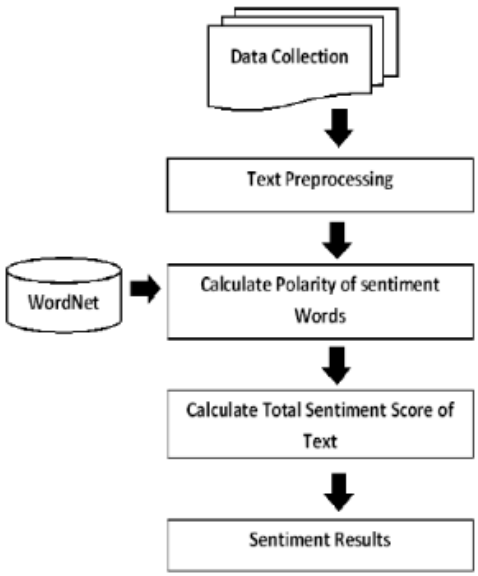



Рис 1. Методология Анализа Настроений

## 4. Заключение

Результат работы всех классификаторов увеличивается после использования меньшего количества классов. Значения потерь журнала улучшаются с большим запасом после уменьшения классов.

В ходе экспериментов можно увидеть, что несбалансированный набор данных был улучшен с помощью недостаточной выборки. Используя неполные данные, Random Forest успешно классифицировала преступную деятельность на основе времени и местоположения. С точностью 60,59% он смог превзойти другие алгоритмы машинного обучения.

Есть много направлений в анализе настроений, которые могут быть изучены. В этой статье был изучен анализ настроений новостей и блогов с использованием набора данных из zakon.kz состоит из новых статей в период с 2018 по 2019 год. Было отмечено, что категории бизнеса и спорта имеют больше положительных статей, в то время как развлечения и технологии имеют большинство отрицательных статей. Будущая работа в этом направлении будет основана на анализе настроений новостей с использованием различных подходов машинного обучения с разработкой онлайн-приложения, из которого пользователи смогут читать новости своих интересов. Кроме того, на основе методов анализа настроений, читатели могут настроить свою ленту новостей.

В этой исследовательской работе мы смогли построить алгоритм, который имеет возможность извлекать и хранить данные, искомые по различным темам с самого популярного сайта социальных сетей в Казахстане zakon.kz. в обзоре литературы был рассмотрен ряд классификационных моделей, из которых мы выбрали наивную модель классификатора Байеса из-за ее простоты в адаптации к

собранным данным. Мы разработали классификатор анализа настроений, который интегрирует эвристику получения информации с использованием набора инструментов естественного языка и обучил его на предварительно обработанном наборе данных из социальных сетей.

## 5. Будущие направления исследований

1. Для этого исследования использовались только данные о преступности, но, как показали многие исследования, социально-экономический стандарт конкретной области также является ключевым показателем возможной преступной деятельности. Этот агент машинного обучения может включать эти данные и может работать лучше.

Эта модель может также использоваться для других географических объектов. Это также поможет проанализировать преступления, совершаемые в разных местах, и лучше понять различные преступления и их связь с конкретной демографией.

Кроме того, существует множество передовых подходов к машинному обучению, которые можно изучить. Глубокое обучение и нейронные сети могут обеспечить более сбалансированное понимание преступной деятельности. Как было показано в этом исследовании, несбалансированные классы были основной проблемой при работе с конкретной базой данных. Продвинутые методы борьбы с несбалансированными классами также являются тем, что еще предстоит изучить.

2. Начнем с того, что социальные сети являются лучшим источником данных для криминальной сферы. Разработка системы выявления преступлений на казахском языке в области Преступности была основной целью настоящего документа. Основной вклад этого документа заключается в следующем: Автоматическое извлечение имен нарушителей из реальных неструктурированных текст преступления, в то время как традиционная система используется для структурирования базы данных системы. Алгоритм обнаружения информации о преступление может использоваться, как справочник или аптечка для сотрудников полиции. В системе анализатора настроений можно выполнить следующие улучшения:

1. Функциональные возможности для размещения других классификаторов, кроме наивного классификатора Байеса, могут быть разработаны в приложении. Эти классификаторы включают деревья решений и вспомогательные векторные машины. Результаты различных классификаторов можно сравнить в интерфейсе отчета для выбора наилучшего метода классификации.

2. Обучение на нескольких словах также могут быть изучены, чтобы решить ограничение.

3. Алгоритм также может быть улучшен, чтобы быть доступным на портативных устройствах, таких как мобильные телефоны.

4. Интеграция с мульти-и межъязыковыми словарями языков для удовлетворения динамической природы языка, используемого в социальных сетях.

**Литература**


1. Beckmann, M., Ebecken, N. F., & de Lima, B. S. P. (2015). A KNN undersampling approach for data balancing. *Journal of Intelligent Learning Systems and Applications*, *7*(4), 104.

2. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 427-434). ACM.

3. Braithwaite J. Crime, Shame and Reintegration. *Ambridge: Cambridge* University Press, 1989.

4. Cowie, J. & Lehnert, W. (1996), 'Information extraction', Communications of the ACM 39(1), 80–91.

5. Wiki Crimes (2013), 'Mapping crimes collaboratively', http://www.wikicrimes.org. Accessed: 15-08-2013.

6. Nadeau, D. & Sekine, S. (2007), 'A survey of named entity recognition and classification', Lingvisticae Investigationes 30(1), 3–26.

7. Finkel, J. R., Grenager, T. & Manning, C. (2005), Incorporating non-local information into information extraction systems by gibbs sampling, in 'Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, pp. 363–370.

8. Albert, R., Jeong, H. and Barabási, A. (1999) Internet: Diameter of the world-wide web, Nature, 401(6749), pp. 130-131. doi: 10.1038/43601.