Impact Factor: 9.2 ISSN-L: 2544-980X

МЕХАНИЗМ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ПРИ АНАЛИЗЕ ТЕКСТОВЫХ ДАННЫХ

О.Бабомурадов[1], О.Туракулов[2]

[1]*Исполнительный директор филиала Казанского федерального университета в Джизаке, Джизак, Узбекистан. E-mail: bobomuradov@gmail.com*
[2]*Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан*. E-mail: *o_xolmirzayevich@mail.ru*

Аннотация. Данная статья посвящена описанию современного состояния средств организации механизма предварительной обработки при анализе текстовых данных в социальных сетях, направлениям исследовательской перспективы и предлагаемым подходам. В статье представлен теоретический анализ подходов к обработке данных, предлагаемое решение и результаты.

Введение. В настоящее время в океане данных быстро увеличивается количество информации, относящейся к разным категориям и типам. Из-за большого объема данных пользователю становится сложнее извлечь из них нужную ему информацию. Чтобы человечество могло искать и извлекать необходимую ему информацию, необходимо обрабатывать данные, анализировать их, а точнее извлекать из данных нужные части. Такая постановка проблемы предполагает, что вместо традиционных методов анализа данных, которые в основном направлены на проверку уже существующих гипотез о данных, было бы уместно использовать интеллектуальный анализ для выявления структуры данных, ранее неизвестных связей и закономерностей между ними.

Сбор данных варьируется в зависимости от цели их использования и типа хранения. Разные типы данных требуют разных подходов. Гомогенный подход может давать разные результаты обработки для одного вида и другого для другой категории. Особенно в наши дни наличие очень больших объемов данных вызывает трудности в их обработке.

Повышение эффективности решения задач классификации, анализа или прогнозирования электронных текстовых документов, совершенствование самого механизма классификации, анализа или прогнозирования оправдывается различными исследователями. Сосредоточение внимания на повышении качества анализа неструктурированного текста может повлиять на точность классификации документов. Успешная реализация начального этапа обработки документов позволяет сократить время, затрачиваемое на анализ или классификацию (на основе заранее заданных классов), и повысить качество. Предварительная обработка текста делится на 2 основных метода: извлечение признаков (ИП) и выбор признаков (ВП).

Теоретические основы первичной обработки. Направление разделения на признаки в основном делится на морфологический, синтаксический и семантический анализ. Морфологический анализ работает с отдельными словами, присутствующими в текстовом документе, путем токенизации, флексий и словосочетаний. При токенизации слова разделяются путем удаления знаков препинания, которые рассматриваются как последовательность слов в текстовом документе. Также из текста удаляются дефисы и ключевые слова. Использование этого процесса приводит к уменьшению количества слов в документе и повышению эффективности обработки текста. Он предполагает реализацию «основы слова» путем лингвистической нормализации ряда слов, то есть путем извлечения корня слова. Подход решает

задачу выделения корневых слов из основного содержания при анализе слов. За счет этого можно сократить портфель слов и облегчить процесс обучения. Для выполнения этой операции можно использовать различные алгоритмы, такие как перебор $suf - x - striping$, $afx\ x - remove$, n-gramm [1-4].

Чтобы из предложения в тексте получить логическое значение, то оно должно подчиняться правилам грамматики. В зависимости от языка текстовой части в синтаксическом анализе приводятся грамматические характеристики. При морфологическом анализе осуществляется анализ морфологических характеристик элементов текста на основе частей речи, а при синтаксическом анализе определяются связи между элементами текста в виде смысловых частей, где отражаются аргумент, признак, союз, дизъюнкция и т. д. Синтаксический анализ используется для определения грамматической структуры текста, состоящего из частей речи, таких как существительные, глаголы, наречия и прилагательные. Синтаксический анализ будет состоять из определения части речи (POS-связывание), а также частей анализа, которые будут использоваться для извлечения логического смысла из текста [6].

Процесс определения POS используется для выражения контекстной связи данного слова на основе грамматических правил. Если в процессе синтаксического анализа класс лексической принадлежности слов ясен, то процесс синтаксического анализа облегчается [Yoshida 2007]. Во многих источниках определение ПОС проводилось разными подходами. [6,7]. МСМ (марка скрытой модели) — один из наиболее многофункциональных таких подходов. Поскольку эту модель легко понять и применить, сформулированы правила для генерации входных последовательностей [8]. Изучение грамматических особенностей-это вопрос принятия решения о том, что такое родословная модели. Этот подход использует грамматический анализ сверху вниз или снизу вверх [9,10].

Распределенный текст на родном языке принимается как понятный, что требует разработки механизма, автоматически генерирующего его [11]. WordNet-Affect или SentiWordNet имеет механизм, ориентированный на выявление настроений (эмоций) в текстах для классификации ключевых слов [10,11]. Однако такие системы требуют больших МБ.

Одной из процедур предварительной обработки при классификации текстовых документов является выделение признаков, с помощью которых можно удалить из текста неактуальную и избыточную информацию. Для этого создается набор слов, и значение слова в документе определяется в определенной единице [xua 2009]. Предложены методы в виде признака векторов интервальной частоты (ВИЧ) [5].

TF определяет частоту повторения слов в текстах из сборника документов. Определить соответствующую группу (класс) текстовых документов можно по частоте встречающейся в них группы слов. В текстовом документе ИЧ считаются несовпадающие части слов. С другой стороны ВИЧ рассматривает появление редких слов в текстовом документе. Механизм IFIDF предназначен для определения частоты и релевантности данного слова путем рассмотрения комбинации двух вышеупомянутых инструментов [12].

В ряде источников исследователи использовали методы учета меры соответствий (сходств) для предварительной обработки текстов [10,13]. Сходство измеряется путем взятия слов (терминов), близких друг к другу грамматически (по структуре) и близких по содержанию, и определяется их количеством. Учитывается степень релевантности слов назначению (количество ключевых слов, представляющих предметную область). Например, такие слова (термины), как «ученик», «преподаватель», «обучение» относятся к числу основных идентификационных слов образовательного процесса, и эти термины тесно связаны друг с другом. Однако это может неточно выражать содержание, созданное в тексте. Потому что некоторые предложения близки, но могут означать противоположные мысли. «Ученик» и «учитель» — субъекты, занимающие две стороны в системе образования. Например, «Я думаю, что этот год не принесет нам неудач», это предложение указывает на то, что оно принесет успех. Однако слово «неудача» является

выражением негативной эмоции (около 75%). Необходимо, чтобы слова были снабжены достаточным лингвистическим (естественным) корпусом, чтобы можно было отделить их от такого содержания. В то же время при использовании методов измерения сходства необходимо апробировать алгоритм, проведя экспериментальное исследование на большом объеме данных [14,15].

В исследованиях используются такие методы, как латентно-семантическое индексирование (ЛСИ) и случайное отображение (СО) ФС, ЛСИ стремится обеспечить лексическое соответствие с помощью семантического разделения, а СО разрабатывает карту близости текстов из содержания большого набора документов [16].

Изучили наиболее распространенные подходы классификации и метаинформации (из текста) (кластеризации) для анализа текста в социальных сетях.

Если внимательно присмотреться, то вопрос классификации посредством предварительной обработки текстовых документов показывает актуальность вопроса анализа текстовых документов в веб-среде. Благодаря проведенной работе, наш анализ следующего этапа показывает актуальность анализа текстовых документов в веб-среде. Благодаря разработкам наших анализов следующий этап будет направлен на решение задачи анализирования текстовых документов в веб-среде. Потому что развитие веб 2.0 позволило пользователям не только ограничиваться статусом использования, но и внести свой вклад в развитие этой сети [1,2]. В веб 2.0 сегмент данных, генерируемый пользователями и различными специальными организациями, считается основным фактором развития рекламных услуг и маркетингана рынке [3,4]. Однако многообразие этой информации требует дополнительной обработки [5,6]. Для этого важными задачами являются обработка, сортировка и извлечение необходимых данных из больших массивов [7]. Этот вид полуавтоматической сортировки данных, сортировки и (значимого) обнаружения закономерностей является важным механизмом изучения (в социальных сетях) данных [3]. Ряд проблем может возникнуть при обработке (анализе) «сырых» данных, не прошедших специальную подготовку [8]. Ряд трудностей, возникающих при изучении мнений социальных групп, требуют решения. Они сгруппированы следующим образом:

1. Реальные рабочие системы имеют больше естественных языков, чем экспериментальные [9,10]. Чем больше используется естественных языков, тем больше ошибок в определении выражаемых обществом мыслей и чувств, поскольку необходимо внести определенные изменения или параметризации в методы и модели относительно структуры естественного языка, в противном случае интерпретация будет неверной.
2. Если рассматривать ряд социальных сетей, то сама программная платформа или интерфейс содержит помимо основной информации постороннюю информацию, такую как реклама, дополнительная информация. Эти посторонние элементы отвлекают от основного контекста и могут снизить точность классификации [11].
3. Контент, создаваемый пользователями, может включать в себя различные стикеры или предложенные системой символы (смайлики, эмодзи, гифки). Даже в этом случае ключевая информация, отражающая чувства владельца или владельцев контента, может быть потеряна [12]. Сосредоточение внимания на решении этих проблем признано приоритетным, и исследования методов, моделей и алгоритмов решения должны быть активизированы. В целом извлечение содержания (идеи) из текстов является одной из основных задач и делится на следующие части:
    1) Классификация сущности (эмоции), выраженной в тексте, является вопросом классификации психических состояний представителей общества [14,15];
    2) Поиск основных признаков мысли – это вопрос выявления чувств по предложениям [16];
    3) Вопрос проведения сравнений на основе уже существующих шаблонов [17].

Во многих случаях для обеспечения высокой точности используются языковые корпуса, но это создает сложность [18-20]. Решение задач на основе словаря может увеличить количество сравнений и повысить вероятность достижения NP-полной задачи [14,16,21,22].

Существует несколько алгоритмов для обнаружения контента (настроений) в документах, включая SVM, линейную регрессию, искусственные нейронные сети, LDA-дезагрегацию, генетические алгоритмы [23-27,29-32].

Текстовый интеллект использует базу данных неструктурированного текста на естественном языке предприятия или сети [17]. Существует механизм, основанный на двух методах анализа, из которых закрытая словарная база слов и словосочетаний решает задачи, охватывающие определенную предметную область [18]. Второй механизм — агрегирование открытых словарей, для которого синтаксический анализ осуществляется независимо от типа естественного языка, но этот подход также неприменим к задаче НЛП (понимания текста на естественном языке) [19]. Исследования, связанные с НЛП, описаны в [20]. Он широко используется при практическом применении синтаксического анализа открытых словарей [21] и при отборе необходимых кандидатов, что считается социальной проблемой [22,23,24,25].

Оба типа задач направлены на решение критической проблемы анализа текста на естественном языке.

Основываясь на приведенном выше анализе, мы видим, что роль механизма предварительной обработки неоценима при анализе или классификации текста на естественном языке. Данная исследовательская работа посвящена решению проблемы предварительной обработки классификации текстовых документов.

Подход к уменьшению размерности символов и анализу текста на естественном языке. Уменьшение пространства символов важно при классификации текста [26], [27]. В связи с этим можно рассмотреть подходы, создающие разные ситуации [28], [29]. Пространство символов текстовых последовательностей в векторных моделях на основе терминов неодинаковы. В этом случае наблюдается чрезмерное потребление ресурсов памяти и времени. Эту проблему можно решить, используя несколько подходов. Выбор подходов зависит от сложности ситуации. Первым из них является анализ главных компонентов, целью которого является выявление некоррелированных частей, поиск дисперсии новой независимой переменной и стремление сохранить ее, что приводит к разделению. Это можно сделать следующим образом. Его можно формализовать в следующем виде: набор данных $x^{(i)}; i = 1,\ldots,m$ и каждый $i(n\ \ m)$ для $x^{(i)} \in \ ^{n}$ будет данным. Матрица $X \quad j$ - будет вектором, рассчитывается j из показания переменной x_j [29], [30].

Анализ главных компонентов можно использовать в качестве инструмента предварительной обработки для уменьшения размера набора данных перед запуском алгоритма обучения, который предоставляется в качестве входных данных $x^{(i)}s$. Анализ главных компонентов можно использовать в качестве алгоритма уменьшения помех, чтобы избежать проблемы избыточности. Анализ главных компонентов — это еще один метод уменьшения размерности, который обобщает линейный анализ главных компонентов на нелинейный случай с использованием метода ядра [31].

Анализ самостоятельных компонентов в зависимости от уровня сложности строящейся лексики представлен Дж. Хераултом. Позже после Дж. Хераулта и С. Юттена велись дальнейшие разработки этого метода. Анализ независимых компонентов — это метод статистического моделирования, при котором наблюдаемые данные выражаются в виде линейного преобразования.

Линейный дискриминантный анализ - широко используемый метод классификации данных и уменьшения размерности. Линейный дискриминантный анализ особенно полезен, когда частоты внутри классов неравны и их производительность оценивается на случайно сгенерированных тестовых данных. Класс-зависимые и класс-независимые преобразования - это два подхода к линейному дискриминантному анализу, которые применяются между дисперсией класса и отношением дисперсии класса, а также между общей дисперсией и отношением дисперсии класса соответственно.

Было показано, что факторизация неотрицательной матрицы или аппроксимация неотрицательной матрицы является очень мощным методом для очень многомерных данных, таких как анализ текста и последовательностей. Этот метод является перспективным методом уменьшения размеров.

Уменьшение размера на основе этих методов включает в себя следующие 5 этапов:

После первоначальной обработки, извлечения символов, такого как извлечение индекса m термина и очистка текста, доступны документы с символами n;

N создать документ $\left(d \in \{d_1, d_2, \ldots, d_n\}\right)$, где $a_{ij} = L_{ij} \times G_i$ вектор L_{ij} j обозначает локальный объем i термина — в документе, и G_i - это i глобальные объемы документа;

Применение метода один за другим ко всем терминам во всех документах;

Проекция полученного вектора r документа в многомерное пространство;

Используя то же преобразование, сопоставьте набор тестов с r -мерным пространством.

В тех случаях, когда классификация текстовых документов связана с естественным языком, существует ряд подходов разработки семантического анализатора посредством семантического моделирования при реализации синтаксического анализа текста на естественном языке.

Анализ текста на естественном языке зависит от языковой группы и структуры. Важным фактором при этом является создание семантики путем изучения синтаксической структуры текста, то есть анализа логических связей.

Большинство теорий (примененных в 80-е годы) перешли к описанию грамматики (wellformedness) с точки зрения правил лицензирования (licensing ruls), которые отражают правильность сложных слов. При таком способе описания языка синтаксис языка не дается, различные ограничения не связаны друг с другом. В этой линии путем анализа делается попытка найти выражение, удовлетворяющее всем ограничениям, где параллельно строится возможный вариант построения. В этом направлении в качестве правил, выражающих специфику языковых конструкций, уместно использовать принципы, более общие для ограничений в описании этих конструкций. В частности, такой подход позволяет выразить грамматические особенности лексических единиц без привязки установленных правил к конкретной конструкции.

Существует два способа применения правила синтаксиса: снизу вверх и сверху вниз. В первом случае используется правило, заменяющее структуру, описывающую правую сторону, на символ, обозначающий левую сторону. Во втором случае доказывается, что данное предложение начинается с начального символа S. При восходящем анализе часто можно применить правило более чем одним способом.

При разборе можно использовать два стандартных правила выбора альтернатив: «Широкий» поиск и «Расширенный» поиск. В первом случае запоминаются все возможные альтернативы и каждая из них открывается параллельно. Если альтернативы терпят неудачу, эти альтернативы удаляются из набора возможностей. При «расширенном» поиске в качестве сравнения берется одна из альтернатив, и в случае неудачи анализ начинается заново с начальной

точки. Использование нисходящего синтаксического анализа позволяет генерировать неграмматические альтернативы. С другой стороны, нисходящий анализ предотвращает выработку дизъюнктивной гипотезы, которая неверна для этого утверждения.

Анализировать результаты частичного распределения приоритетов этих альтернатив можно с помощью таблицы. Если по какой-либо причине анализ заходит в тупик, он возвращается к последней использованной точке правила и пробуется другое правило. Однако таблица, заполненная на основе предыдущего правила, будет сохранена и может быть использована по мере необходимости на текущем этапе распределения. При этом используются приоритетные аспекты (элементы) таблицы (или альтернативы), сформированные с использованием каждого используемого правила. Это оправдывает тот факт, что альтернатива, которая работает плохо при одном подходе, может работать лучше при другом. Для этого запоминаются как гипотезы, так и результаты испытаний. Этот подход называется схематическим анализом. Впервые он был предложен Мартином Кей в системе Powerful Parser [32].

Использование семантических моделей В настоящее время разработаны модели лингвистических парсеров, которые могут анализировать текст на естественном языке, в некотором смысле генерировать его на основе содержания предложений. Кроме того, существуют разные подходы к моделированию процесса, в основе которых лежат используемые при анализе инструменты, то есть элементы, определяющие качество модели лингвистического анализатора, такие как уровень «понимания», объем и методы представления знаний. Некоторые из них сформировали специфические системы, сформировавшие экспрессию на основе анализа текстов [33].

Синтаксический анализ включает в себя анализ исходного текста на его содержимое и выражение этого содержимого на внутреннем языке системы. Под переводом на системный язык понимается преобразование исходного текста в системные знания. Это знание воплощает в себе как содержание, так и форму частей и отношения между ними. Этот параметр анализа приводит к более глубокому «пониманию» смысла текста.

В существующих моделях лингвистического анализа можно выделить следующие методы разделения и описания содержания: компонентный анализ; концептуальная сеть; идентификация контента по изображению; Комплексный подход.

Первые попытки формализовать исходный текст были предприняты в рамках подхода компонентного анализа. Он утверждает, что семантика естественного языка представлена конечными терминами в наборе неструктурированных семантических понятий. При поиске слов результаты получаются путем анализа слов в отдельных группах. Позднее этот подход нашел отражение в модели «Семантических ролей» Ч. Филлмора [34].

К моделям второго класса относятся модели, в которых текстовое содержание представлено в виде концептуальной сети. С помощью этих моделей находятся концептуальные связи текста [36]. Концептуальная сеть — это квазиграф, который рассматривает как бинарные, так и троичные и четверичные отношения, а края соединяют не только вершины, но и другие края.

Следующий тип модели — «Семантическое предпочтение», в котором идентификация контента осуществляется по шаблонам. Отличительной особенностью моделей является отсутствие в них морфологического блока и синтаксического анализа, что является их недостатком. Недостатком является то, что он не обеспечивает глубокого анализа значения слов, необходимого для четкого определения смысловых отношений в тексте.

В этой модели (Уилкса) текст характеризуется следующими элементами: содержанием слов, сообщениями, фрагментами текста и семантической связностью [35].

В другом подходе для анализа используется метод таблицы образцов. Он основан на анализе ключевых слов, которые встречаются в предложении.

К моделям, учитывающим процесс морфологического, синтаксического и проблемного анализа, относятся модели, основанные на комплексном подходе к описанию языка. Примерами таких моделей являются модели «Контент-текст» и модели фрагментации контекста.

Модель «контент-текст» реализует многоуровневый контент-транслятор текстов и наоборот [37]. Выделяют четыре основных уровня: фонетический, морфологический, синтаксический и проблемный. Каждый из них разделен еще на два поверхностных и глубинных уровня, кроме проблемного. Эту модель можно использовать в системах, где необходимо понять полное содержание текста. Однако для полноценного применения модели « Контент - текст» необходимо учитывать индивидуальные особенности очень большого количества пар сотен тысяч словарных, морфологических и лексических единиц.

Модель фрагментации контекста основана на трехуровневой системе: лингвистическая модель, основные механизмы обработки предложений и фрагментация лингвистических знаний. В модели одновременно с преобразованием обнаруженных синтаксических отношений в смысловые осуществляется очень глубокий синтаксический анализ.

Таким образом, модели лингвистических процессоров с разными подходами и ориентациями реализуют следующие возможности:

формирование знаний по заданному тексту и формирование правильных предложений естественного языка на основе заданных значений содержания;

изменение сочетания этих предложений;

оценка их зависимости и решение других вопросов.

Алгоритм семантического анализатора. Основным инструментом решения этих задач является семантический язык написания высказывания и механизм определения совместимости предложений естественного и семантического языков. Процесс анализа показал, что разработанные модели имеют как преимущества, так и недостатки. На основе результатов анализа построены обобщенный алгоритм семантического анализатора и алгоритмы идентификации ролевой структуры и аргументации.

Обобщенный алгоритм семантического анализатора. Алгоритм семантического анализа осуществляется в пять этапов:

поиск слов-сказуемых (ПС) и соответствующих им словарных статей;

поиск аргументов ПС;

определение смысловой роли аргументов, признанных подходящими для каждой из статей словаря ПС, независимо от других аргументов;

выбор оптимального распределения смысловых ролей по аргументам для заданной части словаря путем оптимизации распределения ролей;

выбор оптимальной части словаря ПС и соответствующего ей набора смысловых ролей.

На первом этапе в предложении используются морфосинтаксические шаблоны для поиска ПС. В основном условия размещаются в части речи, а синтаксис - в сетевом месте. Использует лемму ПС из семантического словаря в текстовых предложениях для поиска слова-сказуемого – главного предложения. После этого каждому аргументу для данной части словаря присваивается набор семантических ролей и определяется их объем. Он распределяет семантические роли между аргументами таким образом, чтобы гарантировать, что каждый семантический аргумент для данного ПС не порождает более одной роли. Аргументу семантической роли эвристически присваивается весовое значение от 0 до 1 путем сравнения символов и оценивается. Таким образом, в результате для соответствующей части словарного запаса ПС формируется набор семантических ролей с лучшим охватывающим семантическим аргументом.

Алгоритм определения ролевой структуры предложения представлен на рисунке 1:

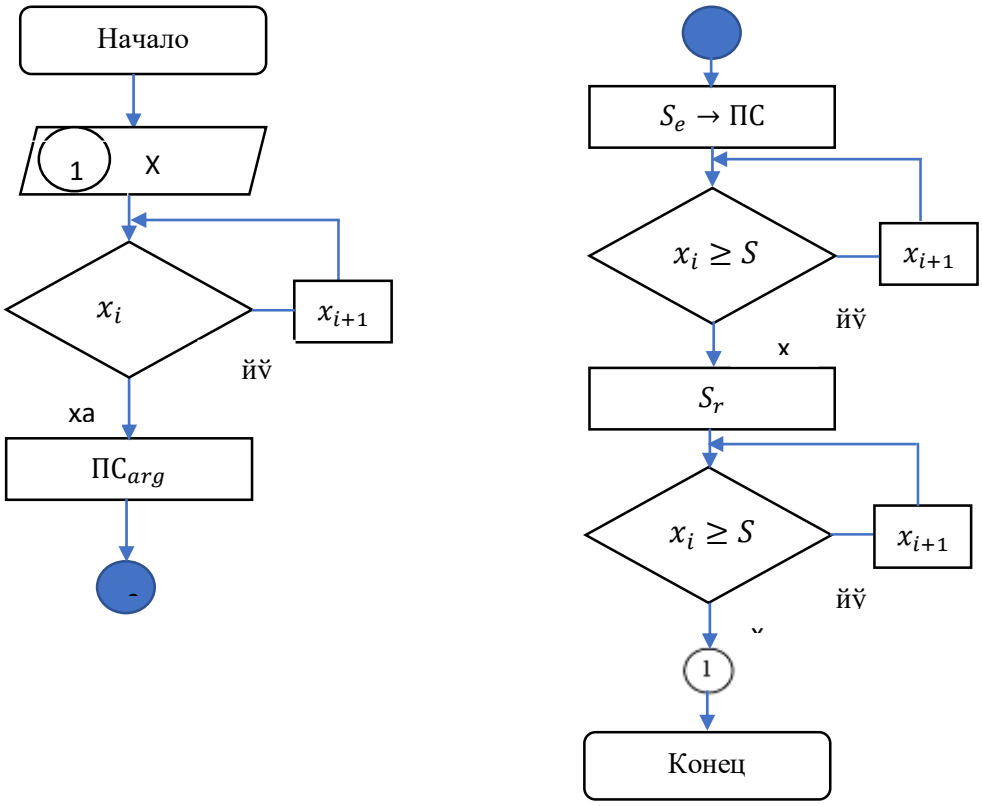

Рисунок 1. Алгоритм определения ролевой структуры предложения.

При определении ролевой структуры предложения, прежде всего, на основе заданных исходных данных формируется морфосинтаксическая структура (X) предложения, производится сравнение всех ПС, а также аргумент ПС и его определенный объем. На основании заданного аргумента ПС выбирается семантическая часть словаря для ПС и формируется список ПС части словаря. Семантическая роль (СР) аргументов определяется на основе всех выбранных частей посредством следующего условия. При этом с помощью этих весов осуществляется распределение ролей по аргументам. После этого производится повторное сравнение по ПС, выбирается наиболее подходящая для ПС часть словарного запаса и формируется ролевая структура предложения.

Результаты экспериментальных исследований. Он внедрен в Центре по внедрению электронного образования в образовательных учреждениях при Министерстве высшего образования, науки и инноваций Республики Узбекистан (ранее Министерство высшего и среднего специального образования), формирующем комплекс информационных систем, включающих в себя несколько информационных ресурсов.

Данная система включает в себя более десяти систем, некоторые из которых использовались в качестве информационно-ресурсных источников и объектов анализа в рамках исследования:

www.edu.uz – главный сайт министерства, который выступает в качестве поставщика информации. При этом в нем хранятся ссылки, позволяющие обратиться ко всем необходимым ресурсам;

my.edu.uz – портал интерактивных услуг и информационных систем в системе высшего образования;

vazir.edu.uz – электронная приемная министра, связаться с министром можно с помощью этого инструмента;

основа для проведения тестового контроля по различным предметам с помощью тестовой системы проверки знаний;

taklif.edu.uz – портал предложений для системы высшего и профессионального образования Республики Узбекистан;

Система мониторинга зарубежных вузов и факультетов, открытых в Узбекистане;

- справочное обслуживание с рабочего места;
- справочная служба с места учебы;
- услуга регистрации на прием руководителя;
- сервис вакансий.

Программный инструмент на основе разработанных алгоритмов был представлен для использования в процессе организации дистанционного и электронного образования в Центре внедрения электронного образования в образовательных учреждениях Министерства высшего и среднего специального образования Республики Узбекистан. Путем предварительной обработки и нормализации текстовых документов программное обеспечение обеспечило высокоточную классификацию текстовых документов, сообщений и комментариев на основе их эмоциональной значимости. Эффективность анализа узбекских текстовых документов составила 10-12%.

«Голос Джизака» внедрен для решения вопросов классификации эмоциональной значимости постов, новостей и различных сообщений на электронных информационных ресурсах газеты Мэрии Джизака и Совета народных депутатов в социальной сети. Применение разработанного механизма классификации позволило сократить время, затрачиваемое на анализ корреспонденции, на 50 %, повысив эффективность работы на 12-15 %.

Оно было использовано при анализе электронного контента Махаллинско-семейного издательства НИИ семьи и женщин. В результате применения показатель эффективности механизма анализа по классификации узбекских текстов составил 15-17%.

Для экспериментального исследования эмоциональной значимости корреспонденции из официального источника информации газеты «голос Джизака» было проанализировано более 2000 постов на узбекском языке.

В проведенных экспериментальных исследованиях результаты, полученные путем предварительной обработки, были применены к вопросам вставки штифтов и их разделения. Результаты представлены ниже.

```
[244]  sia.polarity_scores('Bu eng yomon narsa')

...    {'neg': 0.451, 'neu': 0.549, 'pos': 0.0, 'compound': -0.6249}
```

Рис. 2. Анализ включенных незначительных комментариев.

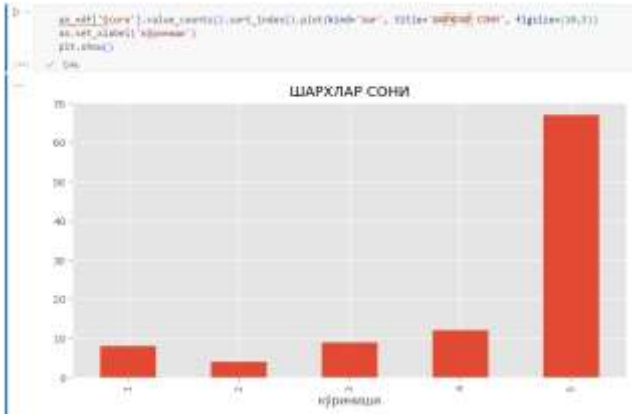

Рисунок 3. Оценка первых 100 отзывов по 5-бальной шкале.

На основе сгенерированных текстовых соответствий был создан массив слов разной полярности. Сентиментальный анализ взаимоотношений этой словесной основы дал следующий результат (4):

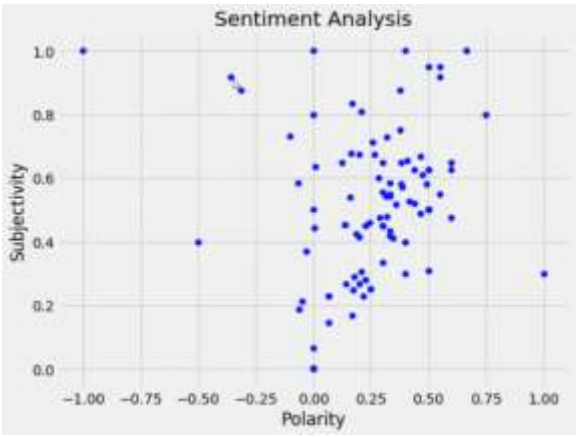

Рисунок 4. Зависимости субъективной поляризации в анализе настроений.

Здесь предел субъективности определяет эксперт ПС. Полярность больше 0 приводит к положительности. Позитивные эмоции выявляются в местах пересечения экспертной границы ПС и положительной полярной границы.

Заключение. Проведенные теоретические исследования нашли отражение в проделанной работе по разработке механизма обработки текстовых данных на узбекском языке на основе анализа подходов предварительной обработки при анализе текстовых данных и повышения их эффективности. При реализации механизма анализа данных разработан компонент предварительной обработки, направленный на определение структуры предложений и основанный на положительных результатах, полученных при анализе узбекских текстовых данных (документов), полученных для экспериментальных исследований.

Список использованной литературы.

1. Xin-She Yang Introduction to Algorithms for Data Mining and Machine Learning// Copyright © 2019 Elsevier Inc. All rights reserved. Academic Press, ISBN: 978-0-12-817216-2, 171p.
2. Hemlata Sahu, Shalini Shrma, Seema Gondhalakar A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2013, Volume 1, Issue 3; P. IndiraPriya, Dr. D.K. Ghosh A Survey on Different Clustering Algorithms in Data Mining Technique, International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.
3. M. A. Deshmukh, Prof. R. A. Gulhane Importance of Clustering in Data Mining, International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016
4. Jaro M. A. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida // Journal of the American Statistical Association.1989. | 84 (406). | Pp. 414{420. | DOI: 10.1080/01621459. 989.10478785.
5. Рассел С. Искусственный интеллект. Современный подход [Текст] / С. Рассел, П. Норвиг, 2-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1408 с.
6. Feldman R. The text mining handbook: advanced approaches in analyzing unstructured data [Текст] / R. Feldman, J. Sanger. – Cambridge University Press, 2007. – 410 p.
7. Moyotl-Hernandez E. An Analysis on Frequency of Terms for Text Categorization [Текст] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Procesamiento del lenguaje natural. – 2004. – Vol. 33. – P. 141-146.

8. Moyotl-Hernandez E. Some Tests in Text Categorization using Term Selection by DTP [Текст] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Proceedings of the Fifth Mexican International Conference on Computer Science ENC'04. – Colima. – 2004. – P. 161-167.
9. Большакова Е., Лукашевич Н., Нокель М. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. — 2013. — С. 31—37
10. Usama F., Smyth P., Piatetsky–Shapiro G. From Data Mining to Knowledge Discovery in Databases // Arti_cal intelligence Magazine. | 1996. |17(3). | Pp. 34-54.
11. Гмурман В. Е. Теория вероятностей и математическая статистика. — Москва : Высшая школа, 2013. — 479 с.
12. Вапник В. Н., Стерин А. М. Об упорядоченной минимизации суммарного риска в задаче распознавания образов // Автоматика и телемеханика. — 1978. — № 10. — С. 83—92.
13. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование / Под ред. А.К. Гуца. – Омск: "Омское книжное издательство", 2010. – Вып. 21. – С. 65-81. – [Электрон ресурс]. URL: http://msm.univer.omsk.su/sbornik/jrn21/sbornik_n21.pdf
14. Kunneman F., Bosch A. van den. Event detection in Twitter: A machine-learning approach based on term pivoting // Proceedings of the 26th Benelux Conference on Artificial Intelligence / Grootjen, F., Otworowska, M., Kwisthout, J. (ed.). – Nijmegen, 2014. – P. 65-72. – [Электрон ресурс]. URL: http://antalvandenbosch.ruhosting.nl/papers/event-detection-twitter.pdf
15. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys (CSUR). – New York, 2002. – Vol. 34, No. 1. – P. 1-47. – [Электрон ресурс]. URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf
16. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – Springer, 2009. – 745 p.
17. Rish I. An empirical study of the naive Bayes classifier // IJCAI 2001 workshop on empirical methods in artificial intelligence. – IBM New York, 2001. – Vol. 3, Issue 22. – P. 41-46. – [Электрон ресурс]. URL: http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf
18. Тузовский А. Формирование семантических метаданных для объектов управления знаниями // Известия Томского политехнического университета. — 2007. — Т. 310. — С. 108—112.
19. Amaravadi C. S. Knowledge Management for Administrative Knowledge //Expert Systems. | 2005. | 25(2). | Pp. 53{61.
20. Kuznetsov S., Poelmans J. Knowledge representation and processing with formal concept analysis // Wiley interdisciplinary reviews: Data mining and knowledge discovery. — 2013. — № 3. — С. 200—215.
21. Roussopoulos N. Conceptual Modeling: Past, Present and the Continuum of the Future // Conceptual Modeling: Foundations and Applications. 2009. | Pp. 139{152.
22. Hutchins J. ALPAC: The (In)Famous Report // Readings in machine translation. 2003. Vol. 14. P. 131–135.
23. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. : Пер. с англ. / Под ред. П. И. Браславского, Д. А. Клюшина, И. В. Сегаловича. М.: ООО «И.Д. Вильямс», 2011. 528 с.
24. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского университета, 2011. 512 с.
25. Deliyanni A., Kowalski R. A. Logic and Semantic Networks // Communications of the ACM. 1979. Vol. 22, no. 3. P. 184–192.
26. Корепанова А. А., Абрамов М. В., Тулупьева Т. В. Идентификация аккаунтов пользователей в социальных сетях «вконтакте» и «одноклассники» // Семнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2019, сборник научных трудов – 2019. – С. 153.]
27. О. Ж. Бабомурадов and Л. Б. Бобоев, “Ўзбек тилидаги матнли ҳужжатларни таснифлашнинг тасодифий ўрмон усули,” in *Олий таълим тизимида масофали таълимни жорий этишнинг*

техник-дастурий ва услубий таъминотини такомиллаштириш истиқболлари, Республика илмий-амалий конференсияси, Қарши, 28 май 2021, 2021, pp. 112–115.
28. О. Ж. Бабомурадов, Л. Б. Бобоев, "Таснифлашни баҳолаш ўлчовлари," in *Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, ЎзМУ Жиззах филиали, Республика миқиёсидаги илмий-техник анжумани*, 2020, pp. 146–153.
29. О.Ж.Бабомурадов, Л. Б. Бобоев, Х. Т. Дусанов, "Матнннинг кетма-кетлик модели," in *Инновацион ёндашувлар илм-фан тараққиёти калити сифатида: ечимлар ва истиқболлар, ЎзМУ Жиззах филиали, Республика миқиёсидаги илмий-техник анжумани*, 2020, pp. 192–197.
30. О. Ж. Бабомурадов, Н. С. Маматов, Л. Б. Бобоев, Б. И. Отахонова, "Text documents classification in Uzbek language," *International journal of recent technology and engineering*, vol. 8, no. 2, pp. 3787–3789, 2019.
31. Y. Du, J. Liu, W. Ke, and X. Gong, "Hierarchy construction and text classification based on the relaxation strategy and least information model," *Expert Systems with Applications*, vol. 100, pp. 157–164, 2018.
32. Гришеленок Д. А., Ковель А. А. Использование результатов математического планирования эксперимента при формировании обучающей выборки нейросети //Известия высших учебных заведений. Приборостроение. – 2011. – Т. 54. – №. 4. – С. 51-54].
33. [Sabuj M.S., Afrin Z., Hasan K.M.A. (2017) Opinion Mining Using Support Vector Machine with Web Based Diverse Data. / Pattern Recognition and Machine Intelligence. PReMI 2017. Lecture Notes in Computer Science, vol 10597. Springer, pp 673-678.
34. Filippov A., Moshkin V., Yarushkina N. (2019) Development of a Software for the Semantic Analysis of Social Media Content. // Recent Research in Control Engineering and Decision Making. ICIT 2019. Studies in Systems, Decision and Control, vol 199. Springer, Cham pp 421-432;
35. Хомский Н. Три модели описании языка//Кибернитеческий сборник.-1961.-Вын.2.-с.81-92.
36. Филмор Ч. Дело о падеже// Новое в зарубежной лингвистике. Вып. X.-М.: Лингвистическая семантика, Прогресс, 1981,-с.369-495.
37. Мельчух И.А. Опыт теорий лингвистических моделей «смысл-текст».-М.: Наука, 1974,-314с.
38. Yarushkina N. G., Moshkin V. S., Andreev I. A. The sentimentanalysis algorithm of social networks text resources based on ontology //Информационные технологии и нанотехнологии (ИТНТ-2020). – 2020. – pp. 226-232.