УДК 004.912
DOI:10.38028/ESI.2023.29.1.015

Обработка слабоструктурированных текстовых данных для использования в моделях анализа

Макарова Елена Андреевна
Брянский государственный технический университет,
Россия, Брянск, *m4karova.e@yandex.ru*

Аннотация. При создании моделей анализа данных часто целесообразно использовать в них данные различной формы и структуры – числовые, категориальные, текстовые, видео и т.д. В статье выполнено исследование влияния текстовых данных без чёткой структуры на качество моделей анализа, выявлена зависимость точности моделей анализа от используемых способов обработки слабоструктурированных текстовых данных. Описана модель интеллектуальной обработки слабоструктурированных текстовых данных, включающая в себя методы визуализации и алгоритмы трансформации данных, предложенные автором в предыдущих работах. Предложена модификация алгоритма трансформации ошибочных написаний, построенная на использовании моделей векторного представления слов. Проведен эксперимент по использованию данных разной структуры в рамках решения задачи классификации резюме соискателей. Приведен пример обработки слабоструктурированных текстовых данных для решения задачи классификации резюме соискателей по подходящим им профессиям. Описаны этапы построения модели интеллектуальной обработки данных, включая разведочный анализ, извлечение и трансформацию данных. Описаны проблемы, свойственные данным, использующимся в эксперименте, таким как: орфографические ошибки, использование разной терминологии для описания одних и тех же понятий и т.д. Приведены примеры объединения словосочетаний с высокой степенью семантической близости и поиска ошибочных написаний распространенных в выборке терминов. Рассчитана точность применения моделей классификации, построенных на данных, обработанных различными способами. Эксперименты показали, что использование слабоструктурированных данных для этой задачи почти не даёт прироста точности модели в случае использования их без предварительной обработки и повышает точность классификации на несколько процентов в случае их корректной обработки.

Ключевые слова: слабоструктурированные текстовые данные, анализ данных, классификация данных, анализ резюме соискателей

Цитирование: Макарова Е.А. Обработка слабоструктурированных текстовых данных для использования в моделях анализа / Е.А. Макарова // Информационные и математические технологии в науке и управлении. – 2023. – № 1(29). – С. 178-189. – DOI:10.38028/ESI.2023.29.1.015.

Введение. Анализ данных широко применяется в различных предметных областях для быстрого и эффективного нахождения различных закономерностей в данных и извлечения информации, необходимой для принятия решений. При этом в качестве источника для анализа могут выступать данные различной формы и структуры: от табличных данных числового типа до неструктурированных данных видеопотоков. Во многих предметных областях важным источником являются текстовые данные разной степени структурированности – информация из различных Интернет-ресурсов, социальных сетей, внутренних баз данных. В статье рассматривается процесс обработки слабоструктурированных текстовых данных (ССТД). Под слабоструктурированными данными понимаются все промежуточные формы между строгой структурой и её полным отсутствием [1]. В ряде случаев использование ССТД позволяет улучшить качество моделей анализа. Например, добавление информации, полученной из новостей и социальных медиа, в модель оценки риска банкротства юридических лиц даёт прирост точности этой модели [2]. Однако, в некоторых случаях прироста качества может не быть, но увеличивается трудоемкость процесса анализа. В работе [3], связанной с анализом текстовых данных для прогнозирования финансовых рисков, обсуждается, что улучшение качества моделей возможно лишь в случаях, когда данные предварительно корректно обработаны, в ином случае качество построенной модели, наоборот, ухудшится.

1. Степень разработанности темы обработки ССТД. Существуют различные подходы, методы и алгоритмы обработки ССТД для использования в моделях интеллектуального анализа данных (ИАД). В обзоре актуальных тенденций в области предварительной обработки текстовых данных [4] подчеркивается, что при проведении анализа текстовых данных необходимо уделять внимание параметрам их обработки, приводя характеристики используемых данных и методов их преобразования. Это необходимо, так как выбор параметров очистки, трансформации, векторизации будет влиять на итоговый результат исследования данных.

Проблема автоматизированной обработки ССТД с целью дальнейшего использования в моделях анализа данных состоит в том, что особенности структуры представления информации в данных зависят от их источника, и даже в одном источнике информация, введенная разными людьми, может быть структурирована по-разному. В итоге, при разработке программного обеспечения или проведении исследований, использующих подобные данные, специалистам в области анализа данных, совместно со специалистами в предметной области, необходимо настроить конвейер их обработки. Этот этап, по некоторым оценкам, может занимать до 70% трудозатрат проекта, использующего текстовые данные [5]. Ручная обработка подобных данных не является оптимальным подходом в большинстве случаев из-за высоких трудозатрат экспертов и высокой скорости накопления новых данных.

В статье предложена модель интеллектуальной обработки ССТД для дальнейшего использования их в моделях анализа данных. Материалы исследования базируются на результатах, полученных автором ранее [6-9].

2. Модель интеллектуальной обработки ССТД. В моделях, описывающих жизненный цикл анализа данных, помимо этапов постановки задачи и, непосредственно, проведения анализа и интерпретации его результатов, отдельно выделяют этапы разведочного анализа и подготовки данных к анализу. Например, это предусматривает модель CRISP-DM [10]. Для решения описанной во введении проблемы была разработана модель интеллектуальной обработки ССТД, опирающаяся на распространенные модели исследования данных. В рамках предложенной модели не рассматриваются построение и оценка качества готовых моделей, а только этапы изучения и подготовки данных. Схема применения модели для этих этапов представлена на рисунке 1.

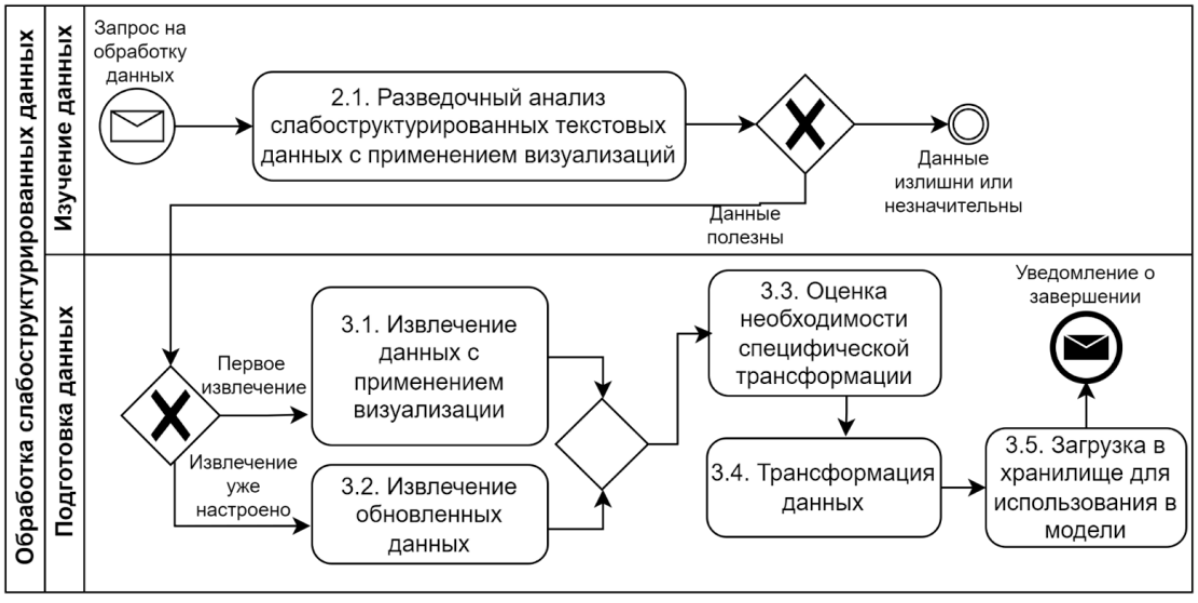

Рис. 1. Модель интеллектуальной обработки ССТД

Основные свойства модели обработки ССТД формально задаются следующим образом:

$$A = < P, S_a, S, O, Pr; T >$$

где P – параметры текстовых данных;

S_a – общее количество текстовых данных в источнике;

S – количество экземпляров текстовых данных, необходимых для обработки в рамках текущей задачи анализа;

O – накопленная информация о прошлых обработках;

Pr – ряд преобразований, которые необходимо осуществить над данными;

T – время на обработку массива данных: показатель, который стремится к уменьшению в данной модели.

В свою очередь, параметры текстовых данных (P) в этой модели описываются такими свойствами, как размер, процент несловарных значений, степень дублирования текстов.

От описанных параметров зависит, какие дополнительные способы обработки необходимо использовать на этапах извлечения и трансформации, чтобы добиться высокого качества данных с минимальным привлечением специалиста в предметной области. Кроме того, от ряда свойств данных зависит, будет ли внедрение модели интеллектуальной обработки в конвейер полезно для увеличения скорости обработки данных, или же наоборот.

Использование предложенной модели интеллектуальной обработки ССТД оказывает положительное влияние на процесс создания модели анализа данных при соблюдении следующих условий:

- данные для анализа не могут быть обработаны полностью автоматически с достаточным уровнем точности и, соответственно, требуют привлечения специалиста в предметной области;
- процесс анализа и, соответственно, обработки, повторяется с появлением новых данных;
- большое количество уникальных значений данных, которые необходимо обрабатывать.

При выполнении всех вышеописанных условий время, затраченное на подготовительный этап обработки, будет оправданно. Формально, данные условия можно описать следующим образом:

$$T_a = P + \sum_{j=1}^{c} S_j * T_E * V,$$

$$T_m = \sum_{j=1}^{c} S_j * T_E,$$

$$T_a < T_m ,$$

где T_a – общее время обработки при полностью ручном подходе; P – время, необходимое на первичную настройку; c – количество циклов обработки; j – номер цикла предобработки; T_E – среднее время работы специалиста в предметной области над одной единицей данных (строкой или документом), включая время простоя; S_j – множество уникальных значений текстовых единиц в цикле обработки j; V – доля данных для ручной валидации, зависит от структурированности и качества исходных текстов, выбирается пользователем на этапе настройки работы системы, по умолчанию взято за 0,05; T_m – общее время обработки при полностью ручном подходе.

Рассмотрим отдельные этапы использования модели интеллектуальной обработки ССТД.

2.1. Этап разведочного анализа ССТД. Разведочный анализ данных важен при работе с новыми данными или с данными сложной структуры. На этапе разведочного анализа ис-

следователь может не только лучше изучить характеристики данных для правильной трансформации, но и сгенерировать новые гипотезы. Для лучшего погружения в сложные данные эффективно используются различные методы визуализации [11]. Сложность заключается в том, что большая часть методов визуализации работает с числовыми или категориальными данными. Адаптация методов визуализации для работы со ССТД описана автором в предыдущей статье [7]. Использование разведочного анализа ССТД необходимо из-за высокой трудоемкости их обработки для дальнейшего использования в моделях анализа. По его результатам часть данных может быть отклонена для использования в модели или, наоборот, добавлена. Пример использования визуализации с целью разведочного анализа будет продемонстрирован далее, в экспериментальной части статьи. Схема применения визуализации в процессе разведочного анализа представлена на рисунке 2. Разработанные автором интерактивные визуализации также используются для настройки извлечения данных из источника по запросу специалиста в предметной области (рисунок 3).

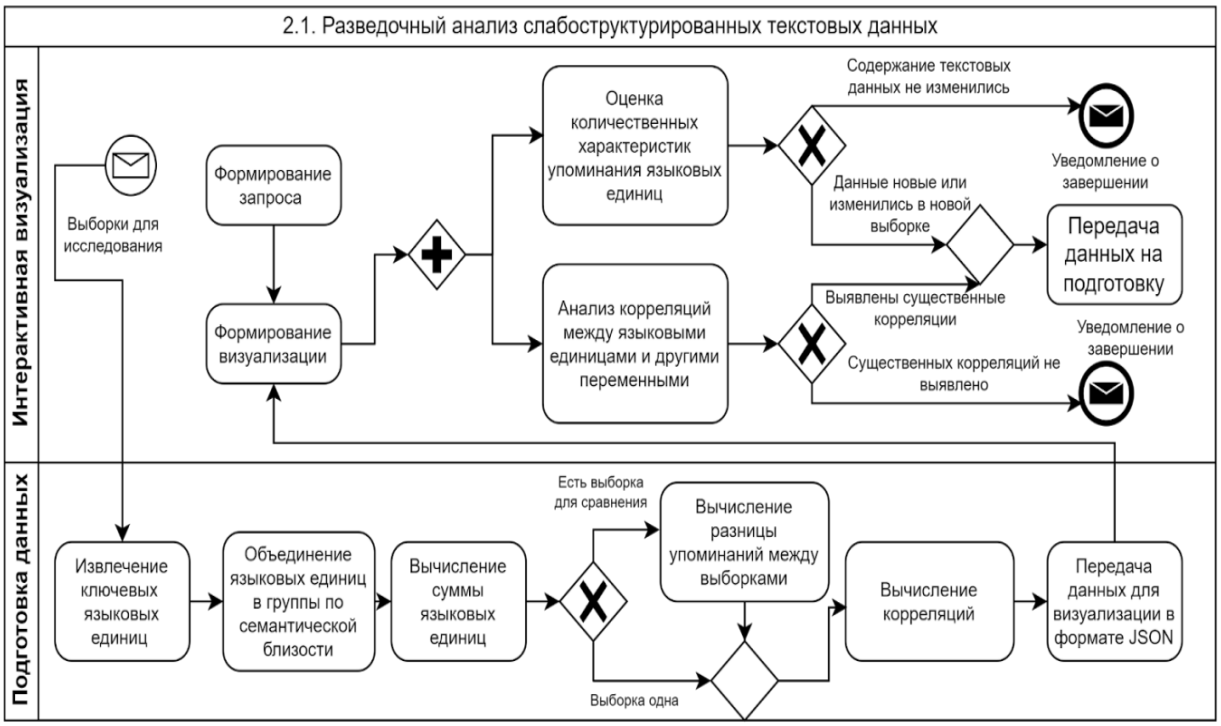

Рис. 2. Использование визуализации для разведочного анализа ССТД

2.2. Особенности трансформации ССТД. Трансформация включает в себя следующие этапы:

1) очистка данных, с учетом особенностей их структуры;
2) оценка количества несловарных значений;
3) оценка количества данных с высокой степенью семантической близости;
4) трансформация специфических сокращений и распространенных в наборе данных орфографических ошибок;
5) вычисление семантической близости между единицами данных и группировка их по этому признаку;
6) извлечение признаков из данных, используя результаты предыдущих этапов и с привлечением специалиста в предметной области.

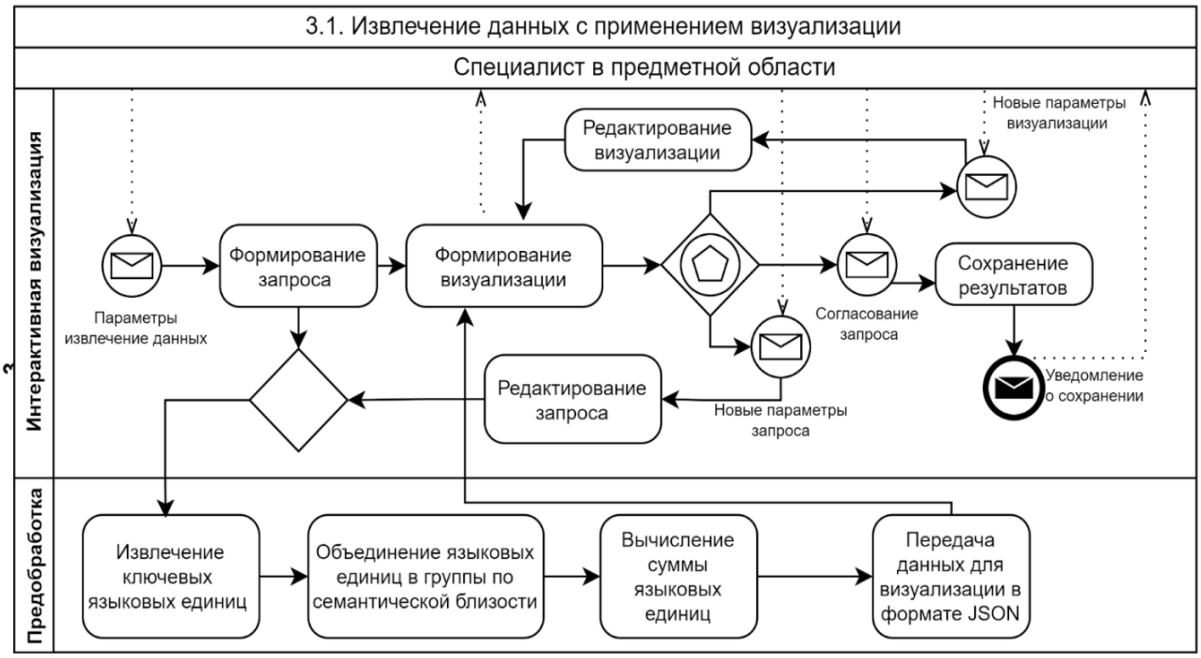

Рис. 3. Извлечение ССТД с применением визуализации

Трансформация специфических сокращений была подробно рассмотрена автором в работе [8]. В алгоритм трансформации были внесены некоторые корректировки, с целью возможности использования его для исправления распространенных в выборке ошибок. Для этого на предварительно обученной модели Word2Vec вычисляются ближайшие, по семантической близости, соседи распространенных в выборке слов. Для автоматической трансформации слов, являющихся ближайшими соседями, как ошибочных написаний исходного слова, должны быть выполнены ряд условий:

1) значение семантической близости находится в диапазоне от 0,7 до 0,99;
2) первый и последний символ сравниваемых слов совпадают;
3) средняя часть слов отличается не более, чем на 20% или 2 символа для слов, короче 8 символов;
4) слово не содержит букв другого алфавита, цифр и иных символов.

Существуют другие методы и инструменты исправления ошибок [12], выбор инструмента будет зависеть от задачи и используемых в проекте технологий.

Экспериментально было выявлено, что для расчёта семантической близости между ошибочными написаниями следует использовать модель, обученную на униграммах [13].

Алгоритм группировки по семантической близости подробно рассматривался в предыдущих работах [9], где на различных данных было опробовано использование классических метрик семантической близости (коэффициент Жаккара, косинусное расстояние) и применение метрики Word Mover Distance, использующей векторное представление слов для оценки близости текстовых единиц со схожим смыслом, но записанных разными словами. Так как метрика Word Mover Distance опирается на значения, которые могут давать нестабильный результат в зависимости от выборки, на которых обучалась модель Word2Vec, она используется совместно с классическими метриками и через обязательную верификацию процесса специалистом в предметной области. Кроме того, алгоритм предусматривает сохранение и повторное использование результатов оценки специалиста, что даёт выигрыш во времени при обработке новых данных.

В качестве этапов трансформации, в зависимости от решаемой задачи, могут выступать извлечение данных, обогащение данных и т.д. Перед применением моделей анализа данных,

построенных на машинном обучении, могут быть использованы различные классические методы обработки естественного языка, как, например, выделение именованных сущностей, извлечение связей, тематическое моделирование и т.д. [14].

Полученные в результате этих процессов данные могут быть использованы в модели анализа данных как отдельные переменные, описывающие ту же сущность, что и исходный текст. Например, переменная может быть дихотомической, отражающей факт упоминания конкретного слова или темы в тексте, или категориальной, описывающей результат извлечения какой-то характеристики исследуемой сущности.

Для использования самих ССТД в моделях анализа данных, необходимо провести их векторизацию. Корректно выбранные параметры векторизации влияют на качество итоговых моделей [15], выбор параметров зависит от параметров ССТД.

3. Описание эксперимента. С целью оценки влияния способов обработки ССТД на качество моделей анализа был проведен эксперимент. В начале эксперимента модель анализа данных была построена на числовых и категориальных данных, а далее в неё были добавлены ССТД, обработанные различными способами. В качестве задачи анализа данных выбрана классификация резюме соискателей по подходящих им профессиям. Для эксперимента взяты данные с сайта «Работа в России» [16]. Так как данный ресурс используется государственными органами по содействию труду и занятости населения, в нём присутствует много резюме людей, которые ищут первую работу или меняют сферу деятельности в связи с потерей работы. С помощью моделей классификации данных, учитывающих навыки, которыми владеет соискатель, и положительный опыт (приглашения на собеседования, полученные другими соискателями), возможна разработка рекомендательной системы, которая будет предлагать соискателю релевантные его образованию, навыкам и опыту позиции. Кроме того, данные из обученных моделей могут использоваться для составления наиболее эффективного с точки зрения привлечения внимания потенциальных работодателей резюме.

Из имеющегося набора данных были собраны 195 272 резюме соискателей, которые подробно описали свои навыки. Помимо навыков, в обезличенных резюме присутствует такая информация, как: возраст, пол, опыт работы, населенный пункт, полученные сертификаты, «гибкие навыки» (личностные качества, как, например, ответственность, исполнительность и т.д.). Первая задача, которая стоит перед исследователем: изучить влияние тех или иных данных на успешность резюме, а также выявить, существует ли значимая разница между упоминанием навыков от людей, претендующих на разные профессии. В таблице 1 приведены примеры описания навыков, которые по смыслу подходят под желаемые позиции.

Таблица 1. Примеры записей поля «навыки»

Описание навыков	Позиция
Умение работать с людьми, ПК (word, excel, power point)	Секретарь
Autocad, компас 3d, Excel, Delphi	Инженер
Вождение более 15 лет: категории b, c, d	Водитель
Знание программы 1с бухгалтерия 8.3, опытный пользователь ПК: MSWord, Excel, Internet, 1с.	Бухгалтер
Кмс по боксу, бывший сотрудник МВД	Охранник

Однако, исходя из визуализации, построенной на выборке 1000 случайных резюме, самыми упоминаемыми навыками являются «ответственность, исполнительность, уверенный пользователь ПК, пунктуальность, коммуникабельность». Модель представлена на рисунке 4, два цвета использованы с целью визуального отделения одних языковых единиц от других.

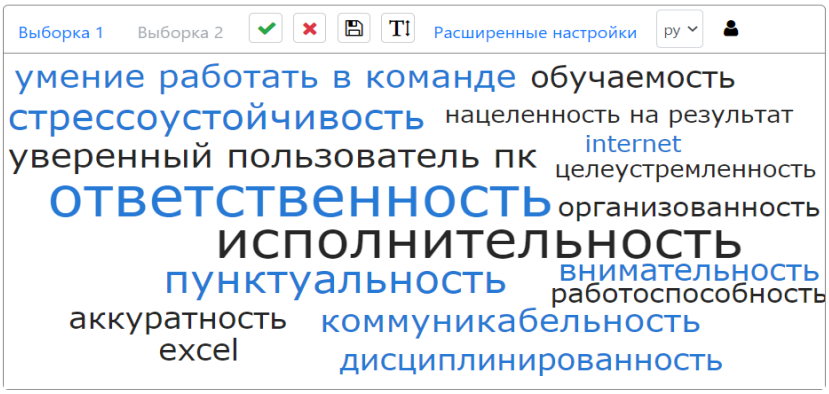

Рис. 4. Визуализация данных о навыках из случайных резюме

Подобные качества присущи представителям многих профессий и, более того, учитывая форму резюме источника данных, большинство из них должны относиться к полю «гибкие навыки». В случае, если подобным образом описывали навыки большинство соискателей, использовать их для модели классификации малоэффективно.

Построим визуализацию, взяв за основу профессию «бухгалтер» и выбрав в качестве выборки для сравнения профессию «водитель», и оценим различающиеся по профессиям упоминания навыков в специально разработанном интерфейсе визуализации. В качестве порога упоминаемости, при котором упоминание какой-либо языковой единицы будет считаться достаточно возросшим или уменьшившимся, выбран коэффициент 1.5. Визуализация представлена на рисунке 5.

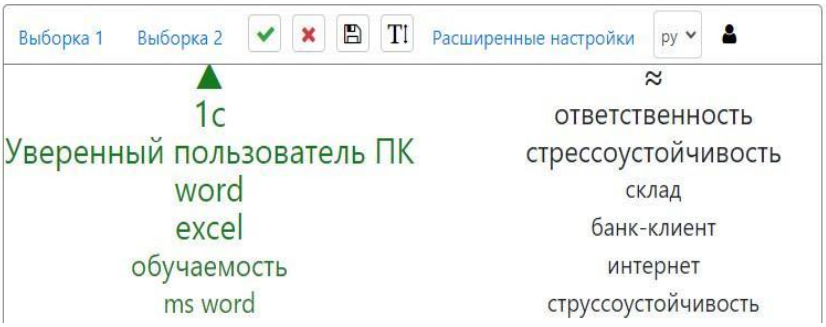

Рис. 5. Визуализация, демонстрирующая разницу упоминаний навыков между разными профессиями

Как видно из приведенной визуализации, ключевые упоминаемые навыки различаются. Следующий шаг разведочного анализа – оценить различие упоминания навыков между соискателями, получившими приглашения на собеседования, и случайными соискателями по профессии «Бухгалтер». Визуализация представлена на рисунке 6.

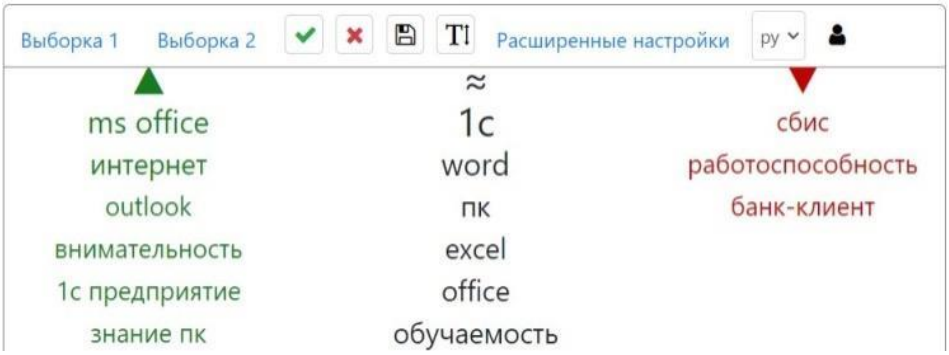

Рис. 6. Визуализация, демонстрирующая разницу упоминаний навыков между выборками резюме

Изучая визуализации, можно сделать вывод, что в качестве ключевых слов выступают не только отдельные слова, но и словосочетания, что может влиять на выбираемые параметры векторизации перед загрузкой данных в модель анализа. Также при изучении визуализаций наблюдается проблема, свойственная данным, заполняемым вручную разными людьми, а именно: обозначение одних и тех же языковых единиц разными словами или сочетаниями слов. Автоматическое объединение подобных слов и словосочетаний возможно при сочетании высокого уровня семантической близости и совпадения лемм исходного и объединяемого слова (словосочетания). Во избежание ошибок, вызванных одинаковым контекстным употреблением разных слов с совпадающими леммами в составе, часть выборки проходит валидацию специалистом в предметной области. Пример семантически близких сочетаний приведён в таблице 2.

Таблица 2. Выражения, близкие к словосочетанию «быстро обучаюсь»

Выражение	Мера семантической близости (от 0 до 1)	Кол-во упоминаний в изучаемой выборке
легко обучаюсь	0,8762	840
быстро обучаема	0,7863	693
быстро учусь	0,7838	269
готова обучаться	0,7747	74
легкообучаема	0,7693	1157
готова учиться	0,7627	56
быстро осваиваю	0,7514	191
легко усваиваю	0,7218	79
легкообучаемая	0,7188	680
быстрообучаема	0,7184	395

Это относится и к различным способам записи названий программного обеспечения или инструментов, которыми владеет соискатель: например, указание сокращенного или полного наименования инструмента, допущение ошибок. Пример представлен в таблице 3.

Таблица 3. Выражения, близкие к словосочетанию «word»

Выражение	Мера семантической близости (от 0 до 1)	Кол-во упоминаний в изучаемой выборке
ms word	0,9011	3841
office word	0,7971	711

Кроме того, статистика употребления несловарных слов в выборке указывает на то, что в ней присутствуют ошибки. Использование моделей векторного представления слов, построенных на контекстном употреблении, позволяет найти самые распространенные ошибки быстро, чтобы трансформировать их автоматически или с привлечением специалиста в предметной области. Пример результатов автоматического поиска ошибочных написаний распространённого в выборке слова представлен в таблице 4. Выделены слова, которые при использовании вышеописанного алгоритма трансформации ошибок могут быть исправлены автоматически.

Решение о необходимости исправления ошибок перед загрузкой данных в модель классификации следует принимать, исходя из потребностей задачи. Например, при анализе приглашений на собеседования, исходя из упоминания тех или иных навыков, исправление ошибок может ухудшить качество моделей, так как наличие ошибок в резюме может повлиять на решение потенциального работодателя не приглашать соискателя на собеседование.

Таблица 4. Слова, близкие к «excel»

Слово	Мера семантической близости (от 0 до 1)	Кол-во упоминаний в изучаемой выборке
exel	0,8737	1688
exsel	0,7910	368
excell	0,7285	130
exell	0,7036	71
ecxel	0,6900	123

3.1. Результаты эксперимента. Для экспериментальной проверки возможности использования обработанных данных в решении задачи классификации исходная выборка была разделена на обучающую (80%) и тестовую (20%) для контроля результатов классификации данных.

Для программной реализации использовалась библиотека Scikit-learn для языка программирования Python. Выбранная библиотека поддерживает различные алгоритмы машинного обучения, в том числе для решения задачи классификации по заранее заданным классам [17]. Из имеющейся выборки были выбраны 20 самых часто встречаемых названий позиций, на которые претендует соискатель. В число выбранных позиций не вошли 2 самых распространенных, но при этом слишком «общих» названий желаемых позиций: «Специалист» и «Менеджер» без указания дополнительных подробностей. Для классификации был выбран метод опорных векторов, который широко используется для решения задач классификации с заранее размеченными классами [18]. Были подсчитаны метрики top-1 и top-3 accuracy [19] на классификации соискателей по позициям, опираясь на имеющиеся категориальные поля и на данные об их навыках. Метрика по трём наиболее подходящим классам была выбрана из-за того, что одно и то же образование может подходить под несколько профессий, как и наборы навыков смежных профессий часто пересекаются. Другие алгоритмы классификации для решения этой задачи не использовались, хотя, в теории, могли показать лучшее качество для решения исходной задачи. Целью эксперимента являлось оценить изменение качества моделей анализа данных при применении одного и того же алгоритма классификации и разными наборами, или же способами обработки данных. Далее представлены 3 способа обработки данных, использованных в ходе эксперимента.

Способ 1 – очистка текста от лишних символов, векторизация по униграммам.

Способ 2 – очистка текста от лишних символов, векторизация по униграммам, биграммам и триграммам.

Способ 3 – очистка текста от лишних символов, объединение самых частоупотребимых семантически близких сочетаний в группы и замена на самое распространенное из них, векторизация по униграммам, биграммам и триграммам. Результаты представлены в таблице 5.

Таблица 5. Результаты классификации

Входы модели	Точность модели классификации	
	top-1 accuracy	top-3 accuracy
Без данных о навыках (только категориальные поля)	61,9%	75,1%
Содержит необработанные ССТД	62,9%	76,5%
Содержит обработанные ССТД (способ 1)	63,1%	77,6%
Содержит обработанные ССТД (способ 2)	64,9%	80,2%
Содержит обработанные ССТД (способ 3)	65,3%	80,9%

Заключение. Использование ССТД в моделях анализа данных может улучшить их качество. Однако, способы их обработки будут иметь заметное влияние на качество полученной модели анализа данных. В статье предложена модель интеллектуальной обработки ССТД, упор в которой сделан на этапы изучения и подготовки данных.

Был проведён эксперимент построения модели классификации соискателей по подходящим должностям с использованием данных различной структуры из резюме. В ходе обработки данных для построения модели классификации была отмечена важность использования разведочного анализа данных, который позволяет выявить различные особенности данных до их загрузки в модель.

Результаты эксперимента подтверждают, что для данной задачи способ обработки данных перед их использованием в модели классификации имел влияние на качество полученной модели. Таким образом, использование предварительно обработанных данных повысило точность модели на 1,2-3,4%, в зависимости от способа обработки, в то время как загрузка в модель ССТД без предварительной обработки повысило её качество на 1%.

Направлениями дальнейших исследований являются:

- исследование методов оценки результатов обработки ССТД с точки зрения эффективности их использования в моделях анализа данных;
- разработка способов улучшения ССТД, таких как: автоматическое извлечение комплексных признаков, устранение неоднозначностей и т.д.

Список источников

1. Климанская Е.В. Методы обработки слабоструктурированных данных в автоматизированных системах на железнодорожном транспорте / Е.В. Климанская, А.В. Чернов, В.И Янц.// Известия вузов. Северо-Кавказский регион. Серия: Технические науки, 2013. – №1 (170). – С. 18-23.
2. Guo L., Shi F., Tu J. Textual analysis and machine leaning: crack unstructured data in finance and accounting. The journal of finance and Data Science, 2016, vol. 2, pp. 163-170, DOI:10.1016/j.jfds.2017.02.001.
3. Dorfleitner G., Priberny Ch., Schuster S. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. Journal of Banking & Finance, 2016, no. 64, pp. 169-187. DOI:10.1016/j.jbankfin.2015.11.009.
4. Hickman L., Thapa St., Tay L. Text preprocessing for text mining in organizational research: review and recommendations. Organizational research methods, 2020, no. 25, DOI:10.1177/1094428120971683.
5. Perez J., Iturbide E., Olivares V. A data preparation methodology in data mining applied to mortality population databases. J Med Syst, 2015, pp. 39-152, DOI:10.1007/978-3-319-16486-1_116.
6. Макарова Е.А. Модель обработки слабоструктурированных текстовых данных на русском языке для интеллектуальной поддержки информационного управления в динамических организационных системах / Е.А. Макарова, Д.Г. Лагерев // Модели, системы, сети в экономике, технике, природе и обществе, 2022. – № 3. – С. 104-125.
7. Makarova E.A., Lagerev D.G., Lozbinev F.Y. Approaches to visualizing big text data at the stage of collection and pre-processing. Scientific Visualization, 2019, no. 11(4), pp. 13-26, DOI: 10.26583/sv.11.4.02.
8. Лагерев Д.Г. Поиск и раскрытие сокращений в русскоязычных данных медицинских информационных систем / Д.Г. Лагерев, Е.А. Макарова // Вестник компьютерных и информационных технологий. 2020. – № 7. – С. 44-54.
9. Макарова Е.А. Оценка семантической близости новостных сообщений на основе анализа заголовков / Е.А. Макарова, Д.Г. Лагерев // Вестник компьютерных и информационных технологий, 2021. – № 7. – С. 46-56.
10. Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing, 2000, vol. 5, pp.13-22.
11. Захарова А.А. Визуальная аналитика и когнитивные методы для обработки и анализа гетерогенных данных мультисенсорных систем: проблемы и тенденции / А.А. Захарова, А.Г. Подвесовский, Д.Г. Лагерев // Информационные и математические технологии в науке и управлении, 2019. – №4 (16), – С. 60-74. DOI:10.25729/2413-0133-2019-4-05.
12. Wang Y., Wang Y., Dang K. A comprehensive survey of grammatical error correction. ACM Trans. Intell. Syst. Technol, 2021, no. 12, vol. 5. pp. 1-51, DOI:10.1145/3474840.

13. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. In proceedings of workshop at ICLR, 2013, DOI:10.48550/arXiv.1301.3781.
14. Большакова Е.И. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Е.И. Большакова, К.В Воронцов, Н.Э. Ефремова [и др]. –М.: Изд-во НИУ ВШЭ, 2017, – 269 с.
15. Solomonova Y., Khlopotov M. Russian text vectorization: an approach based on SRSTI classifier. In: digital transformation and gobal society. DTGS 2019. Communications in computer and information Science. Springer. Cham, 2019, vol. 1038, DOI:10.1007/978-3-030-37858-5_64.
16. Работа в России: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru // Роструд; обработка: Бабушкина В.О., Тимошенко А.Ш., Инфраструктура научно-исследовательских данных. АНО «ЦПУР», 2021. Лицензия CC BY-SA. – URL: http://data-in.ru/data-catalog/datasets/186/.
17. Гребнев К. Н. Машинное обучение с помощью библиотеки Scikit-learn языка Python / К. Н. Гребнев // Математический вестник педвузов и университетов Волго-Вятского региона, 2017. – № 19. – С. 277-281.
18. Zhang Y. Support vector machine classification algorithm and its application. Information computing and applications. ICICA 2012, vol. 308, pp 179-186. DOI:10.1007/978-3-642-34041-3_27.
19. Hossin M, Sulaiman M.N. A review on evaluation metrics for data classification evaluations. International journal of data mining & Knowledge management process, 2015, no. 5, pp. 1-11. DOI:10.5121/ijdkp.2015.5201.

Макарова Елена Андреевна. *Аспирант факультета информационных технологий Брянского государственного технического университета. Направления исследований: обработка текстовых данных, интеллектуальный анализ данных. AuthorID: 959125, ORCID: 0000-0002-5410-5890, m4karova.e@yandex.ru. Россия, 241035, г. Брянск, бул. 50 лет Октября, д. 7.*

UDC 004.912
DOI:10.38028/ESI.2023.29.1.015

Processing of semi-structured text data for use in data analysis models

Elena A. Makarova

Bryansk State Technical University, Russia, Bryansk, *m4karova.e@yandex.ru*

Abstract. In creating data analysis models, it is often advisable to use data of various forms and structures in them - numerical, categorical, textual, video, etc. The article studies the influence of text data without a clear structure on the quality of analysis models, reveals the dependence of the accuracy of analysis models on the methods used for processing semi-structured text data. A model for intelligent processing of semi-structured text data is described, which includes visualization methods and data transformation algorithms proposed by the author in previous works. A modification of the algorithm for the transformation of erroneous spellings, based on the use of vector word representation models, is proposed. An experiment was conducted on the use of data of different structures in the framework of solving the problem of classifying resumes of applicants. An example of processing semi-structured text data for solving the problem of classifying resumes of applicants according to their professions is given. The stages of building a data mining model are described, including exploratory analysis, data extraction and transformation. Problems inherent in the data used in the experiment are described, such as: spelling errors, the use of different terminology to describe the same concepts, etc. The accuracy of applying classification models based on data processed in various ways is calculated. Experiments have shown that the use of semi-structured data for this task almost does not increase the accuracy of the model if they are used without preliminary processing and increases the classification accuracy by several percent if they are correctly processed.

Keywords: semi-structured text data, data analysis, data classification, CV analysis

References

1. Klimanskaya E.V., Chernov A.V., Yants V.I. Metody obrabotki slabostrukturirovannyh dannyh v avtomatizirovannyh sistemah na zheleznodorozhnom transporte [Methods of semi-structured data processing in automated systems on railway transport]. Izvestiya vuzov. Severo-Kavkazskiy region. Seriya: tekhnicheskiye nauki [Scientific journal bulletin of higher educational institutions North Caucasus region], 2013, no. 1 (170), pp. 18-23.
2. Guo L., Shi F., Tu J. Textual analysis and machine leaning: crack unstructured data in finance and accounting. The journal of finance and Data Science, 2016, vol. 2, pp. 163-170, DOI:10.1016/j.jfds.2017.02.001.
3. Dorfleitner G., Priberny Ch., Schuster S. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. Journal of Banking & Finance, 2016, no. 64, pp. 169-187. DOI:10.1016/j.jbankfin.2015.11.009.

4. Hickman L., Thapa St., Tay L. Text preprocessing for text mining in organizational research: review and recommendations. Organizational research methods, 2020, no. 25, DOI:10.1177/1094428120971683.
5. Perez J., Iturbide E., Olivares V. A data preparation methodology in data mining applied to mortality population databases. J Med Syst, 2015, pp. 39-152, DOI:10.1007/978-3-319-16486-1_116.
6. Makarova E.A., Lagerev D.G. Model' obrabotki slabostrukturirovannyh tekstovyh dannyh na russkom jazyke dlja intellektual'noj podderzhki informacionnogo upravlenija v dinamicheskih organizacionnyh sistemah [Model of processing semi-structured text data in russian for intellectual support of information management in dynamic organizational systems]. Models, systems, networks in economics, technology, nature and society [Models, systems, networks in economics, technology, nature and society], 2022, vol. 3.
7. Makarova E.A., Lagerev D.G., Lozbinev F.Y. Approaches to visualizing big text data at the stage of collection and pre-processing. Scientific Visualization, 2019, no. 11(4), pp. 13-26, DOI: 10.26583/sv.11.4.02.
8. Lagerev D.G., Makarova E.A. Poisk i raskrytie sokrashhenij v russkojazychnyh dannyh medicinskih informacionnyh sistem [Features of preliminary processing of semi-structured medical data in russian for use in ensembles of data mining models]. Vestnik komp'iuternykh i informatsionnykh tekhnologii [Bulletin of computer and information technologies], 2020, no. 7(193), pp. 44-54.
9. Makarova E.A., Lagerev D.G. Ocenka semanticheskoj blizosti novostnyh soobshhenij na osnove analiza zagolovkov [Determining the semantic proximity of news messages based on titles analysis]. Vestnik komp'iuternykh i informatsionnykh tekhnologii [Bulletin of computer and information technologies], 2021, no. 7(205), vol. 18, pp. 46-56.
10. Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing, 2000, vol. 5, pp. 13-22.
11. Zakharova A.A., Podvesovskii A.G., Lagerev D.G. Vizual'naja analitika i kognitivnye metody dlja obrabotki i analiza geterogennyh dannyh mul'tisensornyh sistem: problemy i tendencii [Visual analytics and cognitive methods for processing and analysis of heterogeneous data in multi-sensor systems: issues and trends]. Informatsionnyye i matematicheskiye tekhnologii v nauke i upravlenii [Information and mathematical technologies in science and management], 2019, no. 4 (16), pp. 60-74, DOI: 10.25729/2413-0133-2019-4-05.
12. Wang Y., Wang Y., Dang K. A comprehensive survey of grammatical error correction. ACM Trans. Intell. Syst. Technol, 2021, no. 12, vol. 5. pp. 1-51, DOI:10.1145/3474840.
13. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. In proceedings of workshop at ICLR, 2013, DOI:10.48550/arXiv.1301.3781.
14. Bolshakova E.I., Vorontsov K.V., Efremova N.E. Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh [Automatic natural language processing and data analysis], M.: Publishing house of the national research university higher school of economics, 2017, 269 p.
15. Solomonova Y., Khlopotov M. Russian text vectorization: an approach based on SRSTI classifier. In: digital transformation and gobal society. DTGS 2019. Communications in computer and information Science. Springer. Cham, 2019, vol. 1038, DOI:10.1007/978-3-030-37858-5_64.
16. Rabota v Rossii: obrabotannye i obiedinennye svedenija o vakansijah, rezjume, otklikah i priglashenijah portala trudvsem.ru [Processed and combined information about vacancies, resumes, responses and invitations of the trudvsem.ru portal]. Rostrud; edited by Babushkina V.O., Timoshenko A.Sh., Research Data Infrastructure, ANO TsPUR, 2021, License CC BY-SA, available at: http://data-in.ru/data-catalog/datasets/186/.
17. Grebnev K. N. Mashinnoe obuchenie s pomoshh'ju biblioteki Scikit-learn jazyka Python [Machine learning using the Python Scikit-learn library]. Matematicheskij vestnik pedvuzov i universitetov Volgo-Vjatskogo regiona [Mathematical bulletin of pedagogical universities and universities of the Volga-Vyatka Region], 2017, vol. 19, pp. 277-281.
18. Zhang Y. Support vector machine classification algorithm and its application. Information computing and applications. ICICA 2012, vol. 308, pp 179-186, DOI:10.1007/978-3-642-34041-3_27.
19. Hossin M, Sulaiman M.N. A review on evaluation metrics for data classification evaluations. International journal of data mining & Knowledge management process, 2015, no.5, pp. 1-11. DOI:10.5121/ijdkp.2015.5201.

Makarova Elena Andreevna. *Postgraduate student of the Information Technologies Faculty at Bryansk State Technical University. Research interests: text data processing, data mining. AuthorID: 959125, ORCID: 0000-0002-5410-5890, m4karova.e@yandex.ru, 241035, Russia, Bryansk, 50 let Oktyabrya blvd., 7.*

Статья поступила в редакцию 23.12.2022; одобрена после рецензирования 11.01.2023; принята к публикации 11.01.2023.

The article was submitted 12/23/2022; approved after reviewing 01/11/2023; accepted for publication 01/11/2023.