DOI 10.37882/2223-2966.2025.04.10

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ ПОПУЛЯРНЫХ АЛГОРИТМОВ ПОИСКА, АНАЛИЗА И ФИЛЬТРАЦИИ ИНФОРМАЦИИ С ПРИМЕНЕНИЕМ СИСТЕМ ИНТЕЛЛЕКТУАЛИЗАЦИИ

COMPARATIVE ANALYSIS OF THE EFFICIENCY OF POPULAR ALGORITHMS FOR SEARCHING, ANALYZING, AND FILTERING INFORMATION USING INTELLIGENCE SYSTEMS

K. Gorbunov
S. Ivanov

Summary. Modern companies face the challenge of finding relevant instructions in their knowledge bases, resulting in the need to process many incoming customer requests. This issue demands optimal solutions under limited IT resources and regulatory constraints. This article presents a comparative analysis of the performance of search and information filtering algorithms on small datasets within business constraints. The study compares algorithms such as TF-IDF, Bag of Words, Word2Vec, and FastText. The results of the conducted experiment demonstrated that the most efficient algorithm for small data samples is an improved TF-IDF algorithm, enhanced with text preprocessing functionality, hyperparameter optimization, and a hybrid approach incorporating KNN. The obtained results allowed for an increase in the accuracy of information retrieval without significant time loss. Thus, the proposed approach can be adapted to address a wide range of tasks in the field of text information processing.

Keywords: text processing, TF-IDF, Bag of Words, Word2Vec, FastText, search algorithms, machine learning, information filtering, small datasets.

Горбунов Константин Дмитриевич
аспирант, университет ИТМО
gorbunov.kostya@bk.ru
Иванов Сергей Евгеньевич
кандидат физико-математических наук, доцент, университет ИТМО
serg_ivanov@itmo.ru

Аннотация. Современные компании сталкиваются с проблемой поиска релевантных инструкций в базе знаний, вследствие чего возникает необходимость обработки большого количества входящих клиентских обращений. Данная проблема требует оптимальных решений в условиях ограниченных IT-ресурсов и законодательных ограничений. В данной статье представлен сравнительный анализ работы алгоритмов поиска и фильтрации информации на малых датасетах в условиях бизнес-ограничений. Исследование построено на сравнении таких алгоритмов, как TF-IDF, Bag of Words, Word2Vec и FastText. Результаты проведенного эксперимента показали, что наиболее эффективным алгоритмом для применения на малых выборках данных стал доработанный алгоритм TF-IDF, дополненный функционалом по предобработке текста, оптимизации гиперпараметров и гибридным подходом с использованием KNN. Полученные результаты позволили увеличить точность поиска информации без существенной потери времени. Таким образом, предложенный подход может быть адаптирован для решения широкого круга задач в сфере обработки текстовой информации.

Ключевые слова: обработка текста, TF-IDF, Bag of Words, Word2Vec, FastText, алгоритмы поиска, машинное обучение, фильтрация информации, малые датасеты.

Введение

Обработка и классификация обращений клиентов является достаточно трудоемкой задачей для организаций, особенно если имеется большой поток входящих обращений. Несмотря на то, что у компаний может быть создана собственная база знаний по наиболее распространенным вопросам, стандартных алгоритмов поиска по ней может быть недостаточно для эффективного самостоятельного поиска [1].

Данную проблему подтвердили и проведенные глубинные интервью среди компаний, у которых есть собственная база знаний, а также отдел абонентской поддержки. Респонденты ответили, что в среднем на обработку 1 обращения у специалистов уходит 15 минут, из которых около 7 минут уходит на классификацию его в нужный отдел.

Цель и задачи исследования

Целью исследования является определение оптимального алгоритма поиска, анализа и фильтрации информации для малых датасетов (до 1000 записей, в каждой из которых содержится до 5 признаков) с учетом бизнес-ограничений, а также ограничений ИТ-ресурсов.

Данные параметры датасета были выбраны, поскольку такие характеристики были наиболее популярны у опрошенных организаций.

Задачи исследования определены следующие:

1. Определение бизнес-ограничений
2. Обзор литературы
3. Обзор методов поиска без применения систем интеллектуализации
4. Обзор методов поиска с применением систем интеллектуализации
5. Определение методологии исследования
6. Проверка работы методов на выбранном датасете
7. Доработка наиболее эффективного алгоритма поиска под выбранный датасет
8. Валидация полученных результатов

Бизнес-ограничения

Рассматривая бизнес-ограничения, следует отметить следующие пункты:

1. Федеральный закон «О персональных данных» №152-ФЗ. Данный законопроект включает в себя необходимость соблюдения требований законодательства в отношении обработки, хранения и передачи персональных данных. В случае его нарушения компании грозят штрафы, а также репутационные потери. В связи с этим, решения, использующие программный интерфейс API многих популярных LLM-моделей и, в особенности, хранящие пользовательские данные за рубежом, могут не подойти.
2. Ограниченность ресурсов ИТ-инфраструктуры. Включает в себя ограниченные возможности по обновлению аппаратного обеспечения и программного обеспечения.
3. Необходимость оперативной установки и интеграции с существующими системами. Длительные сроки установки новых решений могут блокировать развитие бизнеса. Помимо этого, у многих компаний есть потребность в совместимости между новыми и текущими или устаревшими системами.
4. Возможность масштабирования. Потребность в гибких решениях, которые могут адаптироваться к изменяющимся условиям рынка.

Обзор литературы

В современных алгоритмах, не использующих искусственный интеллект, поиск осуществляется преимущественно за счет использования SQL-оператора LIKE или же регулярных выражений. SQL-оператор LIKE — это наиболее простой способ фильтрации и поиска по подстроке. С помощью регулярных выражений можно очистить окончания слов и убрать пунктуацию, но поиск также будет осуществляться по подстроке.

Среди преимуществ использования таких классических алгоритмов можно выделить высокую скорость, а также отсутствие необходимости дополнительных интеграций. Однако, данные алгоритмы требуют точного совпадения слов, учитывают стоп-слова, пунктуацию и другие синтаксические конструкции при поиске и, наоборот, не учитывают контекст.

В исследовании [2] рассматриваются популярные алгоритмы обработки естественного языка с применением искусственного интеллекта, среди которых можно выделить следующие:

1. TF-IDF и косинусное сходство.
2. Bag of Words
3. Word2Vec
4. FastText

В статье [3] подробно рассмотрен принцип работы алгоритма TF-IDF и косинусное сходство. TfidfVectorizer создает векторное представление текстов с помощью корпуса слов, при этом учитывая как частоту термина в конкретном документе, так и его распространенность среди всех документов.

Работа алгоритма Bag of Words подробно изложена в статье [4]. На основе обучающего текстового корпуса BoW создает словарь, который содержит уникальные слова (или токены) из всех документов. Каждому документу соответствует вектор фиксированной длины, в котором элементы вектора представляют бинарное присутствие каждого слова из словаря в этом документе.

Основная идея Word2Vec, описанная в статье [5], заключается в том, что слова, имеющие схожие значения, будут близки друг к другу в векторном пространстве. Это позволяет моделям более эффективно обрабатывать тексты, учитывая семантические связи между словами.

Рассматривая модель FastText, стоит сказать, что в отличие от Word2Vec, который рассматривает каждое слово как единичный токен, FastText разбивает слова на n-граммы (последовательности символов фиксированной длины), FastText может учитывать частичные формы слов, что особенно полезно для языков с множеством склонений и спряжений [6].

Методология исследования

Оборудование, используемое для проведения исследования имеет следующие технические характеристики:

1. Данные процессора: 4 ядра с тактовой частотой 3 ГГц
2. Объем оперативной памяти: 4 ГБ

Датасет, используемый в исследовании, содержит 350 инструкций из базы знаний ИТ-компании, структура которой представлена в таблице 1.

Таблица 1.

Структура таблицы базы знаний ИТ-компании

Название поля	Тип данных	Описание
ID	integer	Первичный ключ
name	varchar	Наименование инструкции
description	varchar	Детальное описание инструкции

Для валидации работы алгоритмов используется история клиентских запросов в поддержку, структура которых представлена в таблице 2. Поле knowledge_id служит для связи запроса с релевантной инструкцией из базы знаний. Всего в таблице содержится 300 строк.

Таблица 2.

Структура таблицы базы знаний ИТ-компании

Название поля	Тип данных	Описание
ID	integer	Первичный ключ
request_text	varchar	Текст запроса
knowledge_id	varchar	ID инструкции в базе знаний

Для сравнительного анализы был определен следующий перечень алгоритмов:

1. TF-IDF и косинусное сходство.
2. Bag of Words
3. Word2Vec
4. FastText

Для определения наиболее эффективного алгоритма поиска среди алгоритмов были составлены следующие критерии сравнения:

1. Accuracy
2. Precision
3. Recall
4. F1 Score
5. Среднее время обработки 1 запроса, секунд

На рисунке 1 приведены результаты метрик Accuracy, Precision и Recall для рассмотренных алгоритмов. На рисунке 2 приведены результаты метрик F1, а также среднее время обработки запроса.

Проанализировав результаты работы алгоритмов, был сделан вывод, что оптимальный алгоритм поиска для малых датасетов — TF-IDF и косинусное сходство.

Для достижения наилучших показателей выбранный алгоритм TF-IDF был доработан и дополнен следующим функционалом:

1. Предобработка текста
2. Лемматизация или стемминг
3. Удаление редко встречающихся и слишком распространенных слов (стоп-слов)
4. Корректировка гиперпараметров TF-IDF, базовый метод был переопределен для использования в нем логарифмических функций
5. На уровне частоты терминов (TF): использование логарифмов для уменьшения влияния длинных документов с большим количеством повторяющихся слов по формуле, представленной на рисунке 3.
6. На уровне IDF: введение сглаживания по формуле, представленной на рисунке 4. N — общее количе-

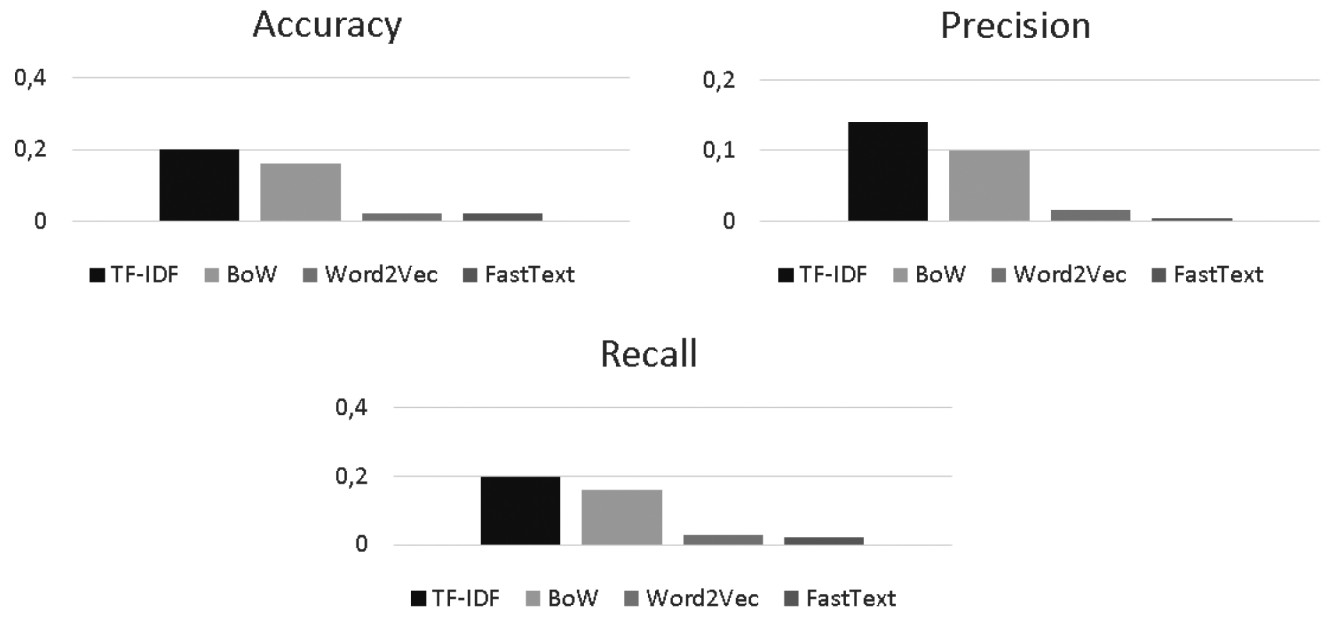

Рис. 1. Результаты метрик Accuracy, Precision и Recall

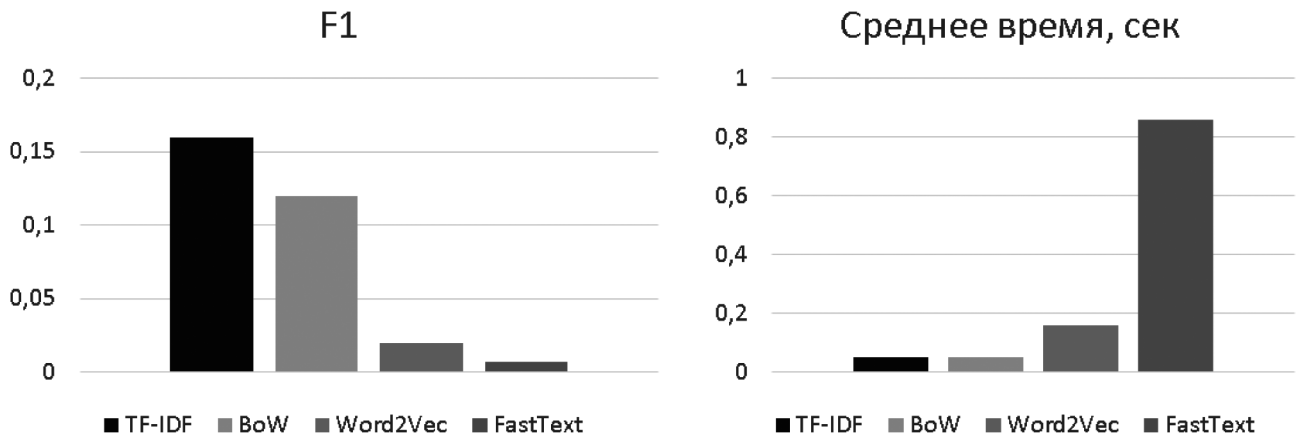

Рис. 2. Результаты метрик F1, Среднее время

ство документов, df(t)— количество документов, содержащих термин t.

7. Учет контекста с гибридными методами. Использование TF-IDF в сочетании с LSA (Latent Semantic Analysis), чтобы дополнить признаки скрытыми тематиками [7].
8. Использование TF-IDF в паре с KNN (Метод k-ближайших соседей) вместо косинусного сходства. [8]
9. Оптимизация гиперпараметров с помощью фреймворка Optuna [9].

$$\log(1 + tf)$$

Рис. 3. Формула расчета TF

$$idf(t) = \log(N + 1df(t) + 1) + 1$$

Рис. 4. Формула для расчета IDF

Оптимальные гиперпараметры, полученные с помощью фреймворка Optuna, для алгоритма knn:

1. radius: 6.02 (расстояние вокруг точки, в пределах которого будут рассматриваться все ближайшие соседние точки)
2. algorithm: auto (тип алгоритма, используемый для построения модели ближайших соседей)
3. metric: cosine (указывает, как модель измеряет расстояние между точками, в данном случае — косинусное расстояние)

На рисунке 3 представлены результаты метрик accuracy и среднего времени выполнения запроса для доработанных версий алгоритма TF-IDF в сравнении с базовой версией алгоритма. Как видно из графиков, лучшей метрики accuracy удалось достичь при комбинированном подходе — предварительной предобработки текста, а также использовании алгоритма TF-IDF в паре с алгоритмом KNN.

Выводы

В ходе проведения исследования были успешно выполнены следующие задачи:

1. Определены бизнес-ограничения.
2. Сделан обзор литературы.
3. Проверена работа изученных методов на выбранном датасете.
4. Определен и доработан наиболее эффективный алгоритм поиска на выбранном датасете.
5. Сделана валидация полученных результатов.

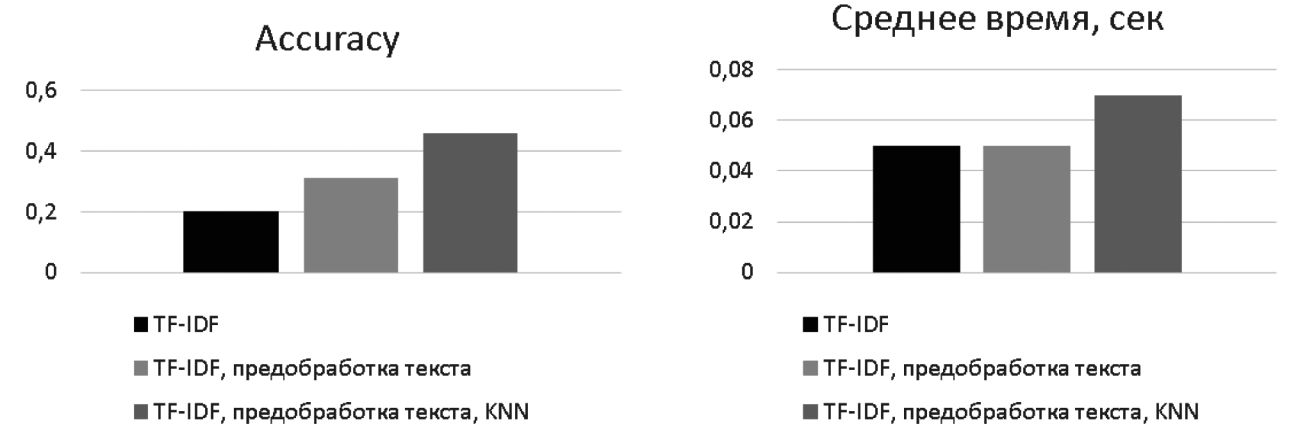

Рис. 3. Результаты метрик Accuracy, Среднее время

Цель исследования выполнена успешно, определен оптимальный алгоритм поиска, анализа и фильтрации информации для малых датасетов с учетом бизнес-ограничений, а также ограничений ИТ-ресурсов. Данным алгоритмом стал доработанный алгоритм TF-IDF.

ЛИТЕРАТУРА

1. Методы конкуренции в бизнесе. — Текст: электронный // Официальный сайт издателя: https://planfact.io/ URL: https://planfact.io/blog/posts/metody-konkurencii-v-biznese (дата обращения: 27.03.2021).
2. Обзор методов автоматической обработки текстов на естественном языке / С.Д. Белов, Д.П. Зрелова, П.В. Зрелов, В.В. Кореньков // Системный анализ в науке и образовании: сетевое научное издание. — 2020. — № 3. — С. 8–22. — URL: http://sanse.ru/download/401.
3. Извлечение признаков из текстовых данных с использованием TF-IDF. — Текст: электронный // Официальный сайт издателя: https://habr.com/ URL: https://habr.com/en/companies/otus/articles/755772/ (дата обращения: 10.12.2024).
4. Кревский М.И., Бождай А.С. Сложные векторные модели бизнес-процессов в задаче классификации // Модели, системы, сети в экономике, технике, природе и обществе. 2023. № 3. С. 142–154. doi: 10.21685/2227–8486-2023-3-10
5. Жаксыбаев Д.О., Мизамова Г.Н. Алгоритмы обработки естественного языка для понимания семантики текста. Труды ИСП РАН, том 34, вып. 1, 2022 г., стр. 141–150. DOI: 10.15514/ISPRAS–2022–34(1)–10
6. Гукасян Ц.Г. Векторные модели на основе символьных н-грамм для морфологического анализа текстов. Труды ИСП РАН, том 32, вып. 2, 2020 г., стр. 7–14. DOI: 10.15514/ISPRAS–2020–32(2)–1
7. Лыченко Н.М., Сороковая А.В. Сравнение эффективности методов векторного представления слов для определения тональности текстов. Математические структуры и моделирование. 2019. № 4(52). С. 97–110. УДК 004.89. DOI 10.24147/2222–8772.2019.4.97–110.
8. Осипова Ю.А., Лавров Д.Н. Применение кластерного анализа методом К-средних для классификации текстов научной направленности. // Математические структуры и моделирование. // 2017. № 3(43). С. 108–121. УДК 004.93. DOI: 10.25513/2222–8772.2017.3.108–121.
9. Akiba T. [et al.]. Optuna: A next generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019. P. 2623–2631.

Журнал «Современная наука: актуальные проблемы теории и практики»