УДК 004.89

АВТОМАТИЗИРОВАННАЯ ПРОВЕРКА ДОСТОВЕРНОСТИ ТЕКСТОВОЙ ИНФОРМАЦИИ В МОБИЛЬНОМ ПРИЛОЖЕНИИ НА ОСНОВЕ ИЗВЛЕЧЕНИЯ УТВЕРЖДЕНИЙ И ДОКАЗАТЕЛЬНОЙ ВЕРИФИКАЦИИ

Кошкарев К. П.[1] - студент, **Майстренко А. Н.** [1] – студент
Научный руководитель – доцент, к.т.н., Федоров Д.А. [1]
[1]Университет ИТМО
koshkarit@gmail.com

Введение

В цифровой среде люди ежедневно сталкиваются с колоссальными объемами информации, которая содержит факты, субъективные суждения и заведомо ложные материалы, которые распространяют как обычные люди, так и авторитетные источники. При этом у обычного пользователя нет времени и навыков для самостоятельной проверки фактов, а существующие ручные решения факт-чекинга нельзя использовать повседневно.

На настоящий момент подходы к автоматической проверке достоверности информации можно разделить на два вида:

1) Поиск уже опубликованных фактчеков, где утверждение сравнивается с базой уже проверенных материалов.
2) Автоматизированная проверка утверждений [1].

Большинство существующих решений ориентированы на англоязычные данные, а также выдают пользователю процент достоверности, не давая явную доказательную базу, что снижает доверие к таким сервисам и увеличивает риск ошибок.

В зарубежной практике за основу берутся постановки, называемые конвейерными: "claim -> retrieval -> evidence -> verification". Также используются бенчмарки по проверке утверждений, методы NLI/entailment для выводов по доказательствам, развиваются подходы Retrieval-Augmented [2] и методы повышения устойчивости к парафразам и тонким изменениям фактов [3].

В отечественной практике часто встречаются решения, которые ограничены классификацией "правда/ложь" без полноценных доказательств.

Основная часть

Предлагается создать многоэтапный конвейер, который будет проверять информацию в формате мобильного приложения, будучи ориентированным на объяснимость и практическую применимость. Поэтапно алгоритм выглядит следующим образом:

1) **Нормализация ввода и выделение проверяемых утверждений.** Пользователь может вставить короткий текст, ссылку или фрагмент статьи. Система выполнит очистку текста и при помощи LLM выделит список кратких утверждений, которые содержат проверяемый факт, отделены от субъективных оценочных суждений и при необходимости нормализованы.
2) **Первый слой - поиск ранее проверенных утверждений.** Для каждого утверждения система выполнит поиск по достоверным агрегаторам/реестрам и внутреннему кэшу уже обработанных запросов. Если факт-чек релевантен, пользователь сразу получает ответ - вердикт, ссылку на материал и другую необходимую информацию.
3) **Основной слой - проверка по доказательствам**. Если готовых фактчеков нет, выполняем поиск кандидатов на источник, отбираем доказательства, верифицируем строго опираясь на выбранные доказательства, формируем вывод [4].

4) **Оценка уверенности и качество источников.** Для избежания иллюзии абсолютной истины, к результату прикрепляется уровень уверенности, который определяется такими параметрами, как:

- Прямота и полнота доказательств
- Количество независимых источников
- Авторитетность источника (официальные/научные/новостные)

5) **Объяснимость результата.** Помимо простого вердикта пользователь увидит:

- Цитаты из источников
- Ссылки на первоисточники
- Комментарии с пояснениями

Оптимальность достигается путем сочетания двух слоев, описанных выше. Это снижает как стоимость проверки, так и время выполнения запроса. Помимо этого, улучшают работу следующие методы:

- Использование LLM в качестве инструмента структурирования (не строит свой вердикт, то есть не является определителем истины)
- Кэширование частых утверждений и источников
- Доменная настройка источников, к примеру, по запросам медицинских фактов приоритет отдается научным и официальным ресурсам, что снижает шум на входе.

Выводы

Автоматический факт-чекинг в прикладном мобильном формате должен быть объяснимым. Наиболее практичным является двухслойный подход - поиск существующих фактчеков + проверка по источникам. Для пользовательского доверия концентрируем внимание на прозрачности - отображаем ссылки/цитаты, боремся с неопределенностью путем уточнений. Результаты могут применяться:

- Пользователями для быстрой проверки спорных утверждений из социальных сетей, новостей;
- Компаниями как инструмент первичной верификации контента;
- Редакциями/инфлюэнсерами/контент-мейкерами как инструмент первичного отбора утверждений, требующих ручной проверки.

Предложения по внедрению и испытаниям:
MVP испытание на ограниченном наборе сценариев, а далее пилот на домене. После проводится сбор обратной связи от пользователей. И заключительный этап - масштабирование - добавление новых доменов и источников, расширение языковой поддержки, внедрение улучшенных алгоритмов.

Список использованных источников:

1. James Thorne, Andreas Vlachos, Automated Fact Checking: Task Formulations, Methods and Future Directions // ACL Anthology. – 2018. – С. 3346–3350.
2. Lewis Patrick, Perez Ethan, Piktus Aleksandra, Petroni Fabio, Karpukhin Vladimir, Goyal Naman, Küttler Heinrich, Lewis Mike, Yih Wen-tau, Rocktäschel Tim, Riedel Sebastian, Kiela Douwe., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Cornell University – 2020. – С. 2–10.
3. Schuster Tal, Fisch Adam, Barzilay Regina., Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence // ACL Anthology – 2021. – С. 624–632.
4. Thorne James, Vlachos Andreas, Christodoulopoulos Christos, Mittal Arpit., FEVER: a Large-scale Dataset for Fact Extraction and VERification // ACL Anthology. – 2018. – С. 809– 817.

Кошкарев К.П. (автор) Подпись

Майстренко А.Н. (автор) Подпись

Федоров Д.А. (научный руководитель) Подпись