УДК 004.85:519.2

Новиков Иван Владимирович, *магистрант,* *Белгородский государственный национальный исследовательский университет, Россия, г. Белгород*

ВЫЯВЛЕНИЕ НЕДОСТОВЕРНЫХ НОВОСТЕЙ: КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, TF-IDF И ВИЗУАЛИЗАЦИЯ ПРИЗНАКОВ

АННОТАЦИЯ

В статье рассматривается проблема выявления недостоверных новостей в условиях стремительного роста цифрового контента и высокой скорости распространения информации. Целью работы является анализ текстовых и стилистических признаков, позволяющих дифференцировать реальные и фейковые новости. Исследование проведено на датасете GonzaloA/Fake_News, содержащем 40587 новостных записей с бинарными метками достоверности. В ходе корреляционного анализа установлено, что наиболее сильную отрицательную связь с классом фейковых новостей демонстрируют такие признаки, как количество восклицательных знаков в заголовке, количество восклицательных знаков в тексте, количество предложений в заголовке, а также количество вопросительных знаков в заголовке. Применение метода главных компонент для визуализации многомерного TF-IDF пространства показало, что классы реальных и недостоверных новостей образуют сложную нелинейную структуру с зонами значительного перекрытия, что указывает на наличие пограничных случаев, трудноразличимых на основе только стилистических характеристик. Полученные результаты подтверждают необходимость использования более сложных контекстно-зависимых моделей для эффективного решения задачи классификации недостоверного контента.

ABSTRACT

This paper addresses the problem of fake news detection in the context of rapid digital content growth and high-speed information dissemination. The study aims to analyze textual and stylistic features that help distinguish real news from fake news. The

research is conducted on the GonzaloA/Fake_News dataset, which includes 40,587 news records with binary veracity labels. Correlation analysis reveals that the strongest negative association with the fake news class is demonstrated by the number of exclamation marks, question marks, and sentences in the title. Principal Component Analysis (PCA) visualization of the multidimensional TF-IDF space shows that real and fake news classes form a complex structure with significant overlap zones, indicating the presence of borderline cases that are difficult to distinguish based solely on stylistic features. The obtained findings confirm the need for more sophisticated context-dependent models for effective fake news classification.

Ключевые слова: недостоверная информация, корреляционный анализ, TF-IDF, метод главных компонент, визуализация данных, текстовые признаки, классификация текстов, обработка естественного языка

Keywords: misinformation, correlation analysis, TF-IDF, principal component analysis (PCA), data visualization, textual features, text classification, natural language processing

Современное информационное пространство характеризуется стремительным ростом объемов цифрового контента и высокой скоростью распространения новостей через интернет СМИ и социальные платформы. Это создает благоприятные условия не только для оперативного информирования общества, но и для распространения недостоверной, искаженной или преднамеренно ложной информации. Проблема фейковых новостей становится особенно актуальной в условиях политической, экономической и социальной нестабильности, поскольку подобный контент способен оказывать существенное влияние на общественное мнение, формировать ложные представления о событиях и даже провоцировать социальную напряженность.

Современные методы выявления недостоверной информации включают анализ текстового содержания, проверку источников, исследование распространения новостей в социальных сетях, а также использование методов машинного обучения и обработки естественного языка. Однако существующие

подходы имеют ряд ограничений: они могут требовать больших размеченных датасетов, не всегда учитывают контекст и динамику распространения информации, а также демонстрируют недостаточную точность при работе с новыми или ранее неизвестными типами фейков.

Датасет GonzaloA/Fake_News представляет собой коллекцию новостных статей на английском языке, собранную и размещённую на платформе Hugging Face в рамках исследовательского проекта по идентификации фейковых новостей с использованием трансформерных моделей. Общий объём данных составляет 40587 уникальных новостных записей, каждая из которых содержит три основных поля: заголовок новости (title), полный текст статьи (text) и бинарную метку достоверности (label), где значение 1 соответствует реальной новости, а значение 0 – фейковой. Датасет структурирован таким образом, что все записи не имеют пропущенных значений, что подтверждается результатами первичного анализа, показавшего отсутствие null-значений по всем трём колонкам. Это делает датасет полностью готовым к использованию без необходимости выполнения дополнительных операций по очистке или вставки пропусков, что существенно упрощает процесс предобработки и позволяет сосредоточиться непосредственно на решении задачи классификации.

Корреляционный анализ количественных признаков с целевой переменной позволяет количественно оценить вклад каждой метрики в различение классов. Положительная корреляция с меткой реальных новостей наблюдается для средней длины слова в заголовке и тексте, хотя эти коэффициенты невелики: 0.043 и 0.017 соответственно. Отрицательную корреляцию, то есть связь с фейковыми новостями, демонстрирует большинство остальных признаков. Наиболее сильная отрицательная корреляция зафиксирована для количества восклицательных знаков в заголовке (коэффициент -0.228), количества восклицательных знаков в тексте (-0.227), количества предложений в заголовке (-0.225), количества вопросительных знаков в заголовке (-0.119), а также для общего количества слов в тексте (-0.057) и доли заглавных букв в тексте (-0.026). Интересно, что количество предложений в тексте и доля заглавных букв в тексте

также коррелируют с фейковыми новостями, хотя и с меньшей силой. Корреляционный анализ признаков недостоверных новостей представлен на рисунке 1.

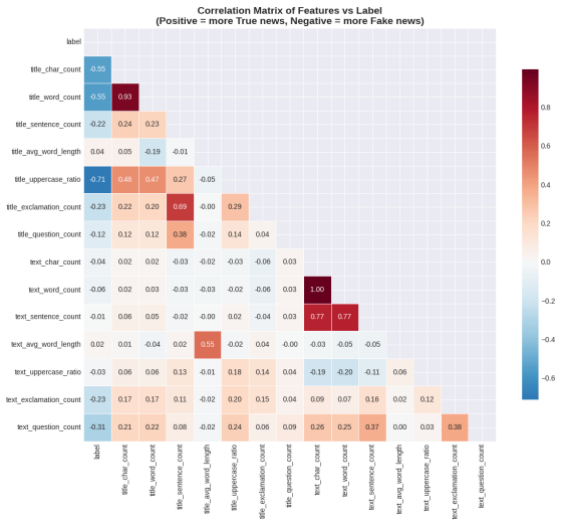

Рисунок 1. Корреляционный анализ признаков недостоверных новостей

После выполнения TF-IDF векторизации с максимальным количеством признаков, ограниченным пятью тысячами наиболее информативных униграмм и биграмм, была получена матрица размерностью 40587 на 5000, что представляет собой достаточно разреженное, но информативное представление текстовых данных. Анализ средних TF-IDF значений слов показал, что наиболее весомым термином во всём корпусе является "trump" со значением 0.0372, что неудивительно, учитывая политическую направленность большинства новостей в датасете. За ним следуют такие общеупотребительные слова как "said", "president", "people", а также "reuters", что свидетельствует о присутствии в выборке значительного количества материалов от авторитетного новостного агентства. Особого внимания заслуживает наличие биграммы "donald trump" в списке топ-20, что подтверждает тесную связь политической тематики с обсуждаемыми новостями. Результаты векторизации TF-IDF представлены на рисунке 2.

```
Fitting TF-IDF vectorizer...
TF-IDF matrix shape: (40587, 5000)
Number of features (n-grams): 5000
Vocabulary size: 5000

=== Top 20 words by average TF-IDF score ===
trump               : 0.0372
said                : 0.0341
president           : 0.0237
people              : 0.0197
reuters             : 0.0192
donald              : 0.0176
state               : 0.0174
donald trump        : 0.0171
house               : 0.0170
new                 : 0.0169
just                : 0.0157
republican          : 0.0157
government          : 0.0155
obama               : 0.0154
states              : 0.0151
clinton             : 0.0149
told                : 0.0148
white               : 0.0148
video               : 0.0148
year                : 0.0145
```

Рисунок 2. Результаты векторизации TF-IDF

Применение метода главных компонент для визуализации многомерного TF-IDF пространства показало, что первые 50 компонент объясняют лишь 13.9 процента общей дисперсии данных, что характерно для разреженных текстовых признаковых пространств с высокой размерностью. Двумерная проекция, объясняющая всего 1.49 процента дисперсии, визуально демонстрирует формирование двух соприкасающихся облаков точек, соответствующих фейковым и реальным новостям. Важно отметить, что хотя существуют области, где преобладают исключительно точки одного класса, значительная часть облаков перекрывается, что указывает на наличие пограничных случаев, где стилистические характеристики фейковых и реальных новостей становятся трудноразличимыми. Трёхмерная проекция подтверждает эту картину, показывая, что классы образуют сложную нелинейную структуру с зонами смешения и относительно чистыми кластерными областями. Визуализация PCA представлена на рисунке 3.

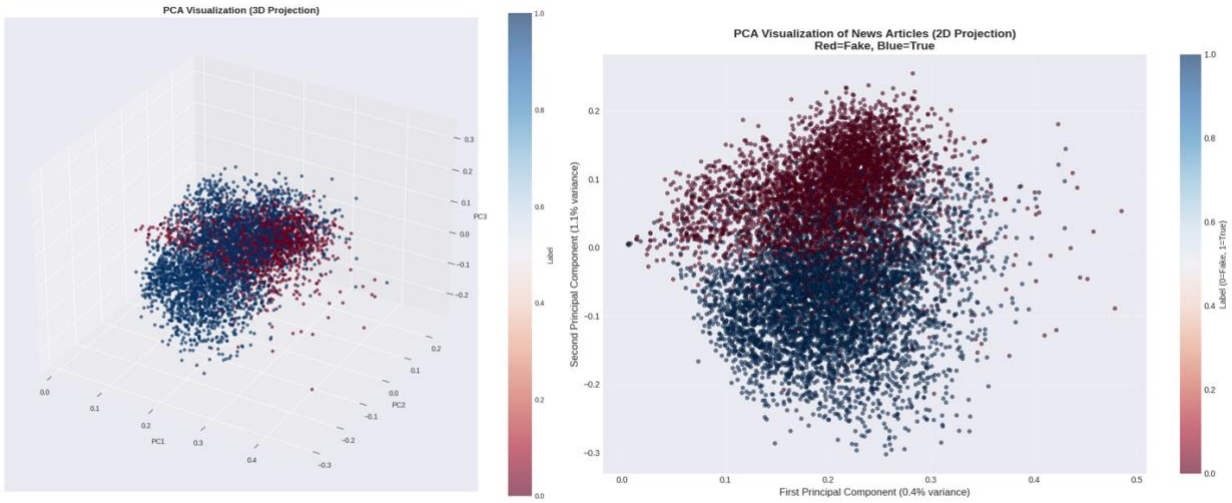

Рисунок 3. Визуализация PCA

Проведённое исследование подтверждает, что задача выявления недостоверных новостей в современном информационном пространстве остаётся сложной и многогранной, несмотря на использование продвинутых методов анализа текста и машинного обучения. Анализ датасета GonzaloA/Fake_News на объёме 40587 записей показал, что простые количественные признаки, такие как количество восклицательных или вопросительных знаков, длина предложений и доля заглавных букв, демонстрируют хотя и слабую, но статистически значимую корреляцию с классом фейковых новостей. Наиболее сильная отрицательная связь с достоверностью выявлена для пунктуационных маркеров эмоциональности (коэффициент до -0.228), что указывает на характерную стилистическую агрессивность недостоверного контента. Визуализация многомерного TF-IDF пространства с помощью метода главных компонент, объясняющая лишь около 1.5% дисперсии в двумерной проекции, наглядно продемонстрировала, что классы реальных и фейковых новостей не образуют полностью разделимых кластеров: их облака точек значительно пересекаются, формируя сложную нелинейную структуру с пограничными зонами, где стилистические различия становятся минимальными. Полученные результаты подчёркивают необходимость перехода от изолированного анализа статических текстовых

признаков к более сложным контекстно-зависимым моделям, способным улавливать семантические и дискурсивные особенности, а также учитывать динамику распространения информации, что особенно актуально для работы с новыми или ранее неизвестными типами недостоверных новостей.

Литература

1. Зембатова М. А. Современный взгляд на искусственный интеллект, машинное обучение и глубокое обучение // Научный потенциал. – 2025. – С. 10-12.

2. Колосов М. И. Обзор методов и технологий обнаружения и борьбы с недостоверными новостями и искаженными данными в социальных сетях // Научно-техническая информация. Серия 1: Организация и методика информационной работы. – 2023. – № 8. – С. 14-21.

3. Салимгареева Д. А. Метод ГЛАВНЫХ КОМПОНЕНТ // NovaInfo.Ru. – 2022. – Т. 3, № 58. – С. 5-9.

4. Слепенко А. И. Мера векторов TF-IDF в анализе сходства текстов // Молодые ученые в решении актуальных проблем науки : сборник материалов Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, Красноярск, 17–18 апреля 2025 года. – Красноярск: Сибирский государственный университет науки и технологий им. акад. М.Ф. Решетнева, 2025. – С. 805-807.