DOI: 10.55648/1998-6920-2025-19-3-122-138 УДК 004.8

Решение задачи классификации слухов в интернет-новостях

А. Д. Худобин, И. Е. Воронина

Воронежский государственный университет (ВГУ)

Аннотация: Рассматривается подход к решению задачи классификации слухов в новостях на основе продукционных правил. Непроверенная информация, появляющаяся на новостных сайтах, имеет характер информационного мусора и способна в отдельных случаях нанести существенный вред потребителям. Решаемая задача носит нетривиальный характер, актуальна и не имеет стандартного решения.

Ключевые слова: слухи, классификация, LLM, машинное обучение, нейронные сети, TF-IDF, Bag-of-words, Word2Vec, N-граммы, продукционные правила, DeepSeek-R1, e5-large.

Для цитирования: Худобин А. Д., Воронина И. Е. Решение задачи классификации слухов в интернет-новостях // Вестник СибГУТИ. 2023. Т. 19, № 3. С. 122–138. https://doi.org/10.55648/1998-6920-2025-19-3-122-138.

Статья поступила в редакцию 13.04.2025;
переработанный вариант – 20.05.2025;
принята к публикации 30.05.2025.

1. Введение

Новостные сайты в интернет-пространстве содержат достаточное количество непроверенной информации. Это становится серьезной проблемой.

Во многих современных СМИ «желтой» направленности есть рубрики с названиями: «Слухи», «Сплетни», «Версия». Формально подобные рубрики анонсируют, что в них содержится непроверенная информация, а фактически они часто используются для очернения действующих лиц [1]. Сейчас новости, которые можно отметить, как слухи, встречаются в СМИ не только в отдельных, предназначенных для них разделах, но и там, где должны быть факты [2].

Следует заметить, что за основу были взяты работы специалистов в области лингвистической экспертизы «Воронежская ассоциация экспертов-лингвистов имени профессора И. А. Стернина» [1].

Слух – высказывание, имеющее особую композицию и стилистические особенности. Например, слухи часто содержат такие маркеры, как «СМИ узнали», «источник сообщил», «говорят», «по слухам». Слух – это массовое распространение информации по актуальным темам, достоверность которой не установлена. Текст, который отмечен как слух, может восприниматься читателями как содержащий недостоверные сведения – после прочтения такого текста в сознании читателя может оставаться вероятностный семантический компонент «недостоверность». Однако, чем длиннее текст, тем больше вероятность того, что читатель поверит в утверждения, написанные в новости [3–9].

Факт – событие, которое уже произошло, его достоверность установлена.

Объемы информации таковы, что актуальна задача автоматического разделения слухов и фактов. Это важно, потому что слухов в новостях становится все больше, а реализаций,

способных их удалить, нет. Для решения данной задачи необходимо разработать подход, позволяющий выполнить классификацию слухов и фактов.

Публикаций, посвящённых выделению слухов, найдено не было, но есть статьи, в которых решаются близкие по тематике задачи. Например, задача классификации речевой агрессии [10] близка к настоящей статье, потому что подразумевает многоклассовую классификацию текста, задача классификации спама [11] также довольно близка, но подразумевает бинарную классификацию текста.

Преимущество настоящей статьи заключается в том, что задача решается впервые, статей на эту тему нет.

Статья организована следующим образом. Во втором разделе описан датасет и методология решения. В третьем разделе приведены результаты решения задачи различными методами, в четвертом разделе обсуждаются ограничения предложенных подходов.

2. Описание датасета и методология решения

Обсудим сложности решения и представим предлагаемый подход. Для решения задачи отделения слухов от фактов предлагается использовать методы машинного обучения и нейронные сети. Важной частью работы является создание исходного набора данных, которого нет. Для этого необходимо выгрузить новости с различных сайтов и провести разметку данных вручную, несмотря на очевидную трудоемкость процесса. В ходе работы возникла проблема: новости, относящиеся к классу «Политика», не получалось разделить на слухи и факты [2]. Поэтому понадобилось разработать набор правил, который позволил бы выполнить эти действия. Правила должны показать, существует ли возможность корректного разбиения на классы.

Предлагается следующий подход к решению:

1. Выгрузить новости с различных сайтов, провести разметку данных вручную.
2. Сформулировать продукционные правила для каждого класса.
3. Провести предварительную обработку данных.
4. Провести векторизацию различными способами (TF-IDF, Bag-of-words, Word2Vec, предобученный Word2Vec) для выбора наиболее подходящего.
5. Решить задачу классификации, используя нейронные сети, методы машинного обучения и языковые модели.
6. Провести сравнение методов и сделать выводы об их эффективности.

Рассмотрим набор данных. В [2] приводятся результаты создания набора данных из 6 классов: «Политика», «Политика слухи», «Спорт», «Спортивные слухи», «Наука», «Научные слухи». Новости были выгружены с сайтов `sports.ru`, `argumenti.ru`. Датасет содержал 12178 новостей. С сайта `sports.ru` были выгружены новости с июля 2023 года по январь 2024 года. С сайта `argumenti.ru` были выгружены новости с августа 2016 года по октябрь 2023 года.

В табл. 1 показано, сколько данных содержал каждый класс и сколько новостей есть в тестовом и тренировочном наборах данных из [2].

Таблица 1. Данные в старом датасете

Наука	Наука слухи	Спорт	Спорт слухи	Политика	Политика слухи
1639	2534	1999	1980	3745	281
Тренировочная часть			Тестовая часть		
8524			3654		

В датасет добавлены классы «Шоу-бизнес» и «Шоу-бизнес слухи». Новости для них были выгружены с сайтов `7days.ru` и `argumenti.ru`. С сайта `7days.ru` были выгружены новости с января 2013 года по август 2024 года.

Набор данных содержит 6570 новостей.

В табл. 2 показано, сколько данных содержит каждый класс и сколько новостей есть в тестовом и тренировочном наборах данных.

Таблица 2. Данные в датасете

Наука	Наука слухи	Спорт	Спорт слухи	Шоу-бизнес	Шоу-бизнес слухи
1095	1095	1095	1095	1095	1095
Тренировочная часть			Тестовая часть		
5256			1314		

В [2] данные были несбалансированными. Это приводит к тому, что модель в качестве ответа выдает тот класс, в котором больше элементов [12]. В новом наборе данных количество новостей в каждом классе одинаково.

Основное правило для проведения разметки: если событие уже произошло – это факт, если кто-то говорит, что событие должно произойти когда-то в будущем (завтра, через месяц, через год) – это слух. Для понимания, что событие действительно произошло, выбираются источники, которым мы можем доверять, например, в новостях науки – это НАСА и Роскосмос, а в новостях шоу-бизнеса – сам человек, его представитель, его близкий родственник. Жена, муж, дети, родители (близкие родственники могут быть в разных отношениях, поэтому каждый источник приходится рассматривать отдельно и выбирать, можно ли ему доверять). Если в новости говорят, что событие произошло, но есть пометка, что это инсайдерская информация – это слух. На этой основе были составлены продукционные правила, которые дополнялись во время разметки. Дополнения нужны, чтобы показать, какому классу необходимо присвоить некоторые похожие друг на друга новости, которые часто встречаются. Например, было прописано правило в новостях науки, что если новость об уфологах, то это всегда слух.

Представим набор продукционных правил, который был разработан для разбиения новостей на классы. [13]

Правило 1 для каждого абзаца:

Если абзац не содержит «недавно/давно появилась новость» и абзац не содержит «читайте далее новость», то абзац_смысл=да.

(Правило необходимо для того, чтобы отбросить текст, в котором идет описание предыдущих новостей, так как необходимо сделать вывод по конкретной новости, а не по связке).

Правило 2:

Если абзац_смысл==да и в новости человек говорит о себе, и событие, о котором говорят в новости, уже произошло, то шоу-бизнес факт.

Правило 3:

Если абзац_смысл==да и в новости говорит официальный представитель, и событие, о котором говорят в новости, уже произошло, то шоу-бизнес факт.

Правило 4:

Если абзац_смысл==да и новость содержит опровержение, то шоу-бизнес факт.

Правило5:

Если абзац_смысл==да и в новости говорит не сам человек, и в новости говорит не официальный представитель, то шоу-бизнес слух.

Правило 6:

Если абзац_смысл==да и (новость содержит «по слухам» или новость содержит «СМИ узнали» или новость содержит «по слухам»), то шоу-бизнес слух.

Правило 7:

Если абзац_смысл==да и человек говорит о событии, которое еще не произошло, то шоу-бизнес слух.

Правило 8:

Если абзац_смысл==да и (новость содержит «широкая одежда» или новость содержит «тайная свадьба»), то шоу-бизнес слух.

Правило 9:

Если абзац_смысл==да и в новости пишут об уфологах, то наука слух.

Правило 10:

Если абзац_смысл==да и (в новости говорит представитель НАСА или в новости говорит представитель Роскосмоса) и говорят о планах на что-то, то наука слух.

Правило 11:

Если абзац_смысл==да и (в новости говорит представитель НАСА или в новости говорит представитель Роскосмоса) и говорят о результатах какого-либо процесса, то наука факт.

Правило 12:

Если абзац_смысл==да и (в новости говорит представитель НАСА или в новости говорит представитель Роскосмоса) и называют точную дату начала какого-либо процесса, то наука факт.

Правило 13:

Если абзац_смысл==да и в новости пишут о раскопках древних сооружений, то наука факт.

Правило 14:

Если абзац_смысл==да и в новости пишут об уже произошедшем запуске ракеты (спутника), то наука факт.

Рассмотрим этап предварительной обработки данных. Для решения задачи классификации необходимо:

1. Очистить новости, удалить лишние символы.
2. Обрезать каждую новость, оставить только первые 1350 символов.
3. Разбить новости на токены.
4. Провести лемматизацию всех токенов.

Стоп-слова не удаляются из новостей, так как они могут показывать, что новость является слухом [14].

Для лемматизации использовался морфологический анализатор pymorphy, так как в [15] он показал лучшие результаты в количестве верно полученных начальных форм.

Остаются только первые 1350 символов, потому что с таким количеством были получены наилучшие результаты при решениях с Word2Vec, TF-IDF и Bag-of-words. Кроме этого, языковые модели имеют ограничение на максимальное количество токенов на вход. Языковые модели, которые используются в работе, могут принять примерно 2500 символов, поэтому для решений с ними в любом случае необходимо обрезать новости. Так как лучшие результаты классификации были получены с новостями, в которых есть только первые 1350 символов, можем сделать вывод, что для классификации новости достаточно ее части, но эта часть должна быть не слишком мала, чтобы остался смысл.

3. Методы решения задачи

Результаты оцениваются с помощью метрик: accuracy, precision, recall, f1-score. Accuracy (точность) – доля правильно предсказанных объектов от общего числа объектов. Precision (точность для класса) – доля правильно предсказанных объектов класса из всех

объектов, которые модель отнесла к классу. Recall (полнота) – доля правильно предсказанных объектов класса от общего количества объектов, которые действительно принадлежат классу. F1-score – гармоническое среднее между точностью и полнотой. Точность – основная метрика для оценки, потому что количество новостей в каждом классе одинаково, датасет сбалансирован. Полнота и f1-score – вспомогательные метрики для оценки. Если полнота и точность для класса примерно равны при использовании сбалансированного датасета, то модель примерно одинаково эффективно обнаруживает нужные объекты и избегает ошибочного включения объектов, относящихся к другим классам. Будем использовать следующие методы машинного обучения: градиентный бустинг, гребневая регрессия, случайный лес, метод опорных векторов, стохастический градиентный спуск, логистическая регрессия, метод k ближайших соседей, дерево решений.

TF-IDF – частотность термов по отношению к обратной частотности документа. TF-IDF используется, чтобы понять, насколько важно каждое из слов для общего смысла фрагмента текста в наборе документов [16].

Bag-of-words – текст представляется в виде мешка его слов. Порядок слов не учитывается, но хранится их количество. Количество вхождений слова называется частотностью терма [17].

N-грамма – последовательность токенов из N элементов [18].

В табл. 3 представлена точность классификации методами машинного обучения для способов векторизации TF-IDF и Bag-of-words. Указан только лучший результат для каждого способа векторизации. (N-граммы) в первом столбце показывают, какие n-граммы использовались для метода векторизации. Например, (1, 3) – униграммы, биграммы, триграммы, (2, 2) – только биграммы.

Таблица 3. Точность классификации для TF-IDF и Bag-of-words

Способ векторизации, (N-граммы)	Общая точность	Наука		Шоу-бизнес		Метод машинного обучения
		Факт	Слух	Факт	Слух	
TF-IDF (1, 1)	0,820	0,693	0,807	0,741	0,716	Метод опорных векторов
TF-IDF (1, 2)	0,829	0,730	0,822	0,775	0,722	Гребневая регрессия
TF-IDF (1, 3)	0,820	0,728	0,818	0,744	0,695	Градиентный спуск
TF-IDF (2, 2)	0,792	0,693	0,810	0,701	0,658	Градиентный спуск
TF-IDF (1, 6)	0,807	0,712	0,811	0,763	0,665	Градиентный спуск
BOW (1, 1)	0,787	0,663	0,749	0,691	0,687	Логистическая регрессия
BOW (1, 2)	0,802	0,695	0,775	0,724	0,695	Гребневая регрессия
BOW (1, 3)	0,804	0,694	0,797	0,690	0,690	Логистическая регрессия
BOW (2, 2)	0,780	0,671	0,803	0,690	0,688	Градиентный спуск

Видим, что точность для решений с использованием TF-IDF выше, чем точность для решений с Bag-of-words. Так происходит, потому что Bag-of-words не смотрит на значимость слова в корпусе.

TF-IDF при использовании униграмм и биграмм даёт лучшие результаты, при добавлении триграмм размерность возрастает и это приводит к переобучению.

В табл. 4 представлены полнота и f1-score для способов векторизации TF-IDF и Bag-of-words. (N-граммы) в первом столбце показывают, какие n-граммы использовались для метода векторизации. Например, (1, 3) – униграммы, биграммы, триграммы, (2, 2) – только биграммы. Показаны результаты нескольких решений с использованием TF-IDF и одного с использованием BOW.

Таблица 4. Полнота и f1-score для TF-IDF и Bag-of-words

Способ векторизации, (N-граммы)	Метрики	Наука		Шоу-бизнес	
		Факт	Слух	Факт	Слух
TF-IDF (1, 2)	Recall	0,79	0,76	0,69	0,81
	F1-score	0,76	0,79	0,73	0,76
TF-IDF (1, 6)	Recall	0,77	0,75	0,59	0,83
	F1-score	0,74	0,78	0,67	0,74
BOW (1, 3)	Recall	0,77	0,73	0,68	0,69
	F1-score	0,73	0,76	0,68	0,69

Можем заметить, что полнота и точность для каждого класса шоу-бизнеса для способа векторизации BOW(1, 3) примерно равны, но f1-score низкий. У остальных способов векторизации полнота и точность отличаются значительно.

Word2Vec используется для получения векторных представлений слов с невысокой фиксированной размерностью. Векторное представление основано на контекстной близости: слова, встречающиеся в тексте рядом, имеют высокое косинусное сходство. В Word2Vec реализованы два основных алгоритма обучения:

1. Skip-gram (SG) – выходные слова предсказываются на основе входного слова.
2. CBOW – выходное слово предсказывается по входным словам, то есть предугадывается текущее слово, основываясь на окружающем контексте [19–20].

С использованием Word2Vec было проведено сравнение предобученных моделей (RUWIKI, WIKI-LENTA, CC.RU) и собственных, созданных на основе обучающего набора данных. В собственных моделях проводилось сравнение нескольких параметров модели: количество эпох, размер окна, алгоритм обучения. Эпоха – итерация прохода текста. Размер окна – количество слов в контексте текущего слова.

В табл. 5 показана точность классификации методами машинного обучения для способа векторизации Word2Vec. Указан только лучший результат для каждого способа векторизации.

Таблица 5. Точность классификации для Word2Vec

Способ векторизации, количество эпох, окно	Общая точность	Наука		Шоу-бизнес		Метод машинного обучения
		Факт	Слух	Факт	Слух	
CBOW, 100, 2	0,800	0,678	0,768	0,724	0,720	Логистическая регрессия
CBOW, 100, 5	0,796	0,667	0,761	0,741	0,733	Логистическая регрессия
CBOW, 300, 2	0,800	0,672	0,760	0,727	0,741	Логистическая регрессия
CBOW, 300, 5	0,796	0,667	0,804	0,741	0,708	Гребневая регрессия
CBOW, 10, 2	0,778	0,684	0,804	0,709	0,709	Градиентный

						бустинг
CBOW, 10, 5	0,769	0,637	0,773	0,718	0,673	Метод опорных векторов
CC.RU, 10, 5	0,737	0,664	0,801	0,634	0,617	Градиентный бустинг
RUWIKI, 10, 5	0,791	0,680	0,806	0,708	0,727	Логистическая регрессия
WIKI-LENTA, 10, 5	0,790	0,677	0,802	0,730	0,700	Градиентный спуск
SG, 10, 2	0,772	0,637	0,793	0,733	0,677	Метод опорных векторов
SG, 10, 5	0,772	0,655	0,795	0,820	0,606	Градиентный спуск
SG, 100, 2	0,796	0,674	0,790	0,744	0,694	Метод опорных векторов
SG, 100, 5	0,802	0,674	0,804	0,762	0,724	Логистическая регрессия
SG, 300, 2	0,788	0,652	0,779	0,765	0,684	Градиентный спуск
SG, 300, 5	0,802	0,671	0,811	0,759	0,722	Логистическая регрессия

При увеличении количества итераций в собственном Word2Vec, увеличивается точность. Изменение размеров окна не оказывает большого влияния на результат.

Можем заметить, что предобученные Word2Vec значительно проигрывают в точности классификации шоу-бизнеса собственным моделям. На новостях науки большой разницы в результатах нет. Скорее всего, эти модели просто не обучались на новостях шоу-бизнеса.

Фиксированная размерность является преимуществом Word2Vec в сравнении с TF-IDF и Bag-of-words, так как в результате работы TF-IDF и Bag-of-words на выходе могут получиться векторы размерностью в десятки тысяч измерений, что усложнит работу [21]. В табл. 3 TF-IDF (1, 6) имеет точность ниже, чем TF-IDF (1, 2). Это связано с тем, что в TF-IDF (1, 6), размерность становится очень большой.

В табл. 6 показаны полнота и f1-score для способа векторизации Word2Vec. Приведены результаты нескольких решений с использованием предобученных Word2Vec и одного с использованием собственного Word2Vec.

Таблица 6. Полнота и f1-score для Word2Vec

Способ векторизации, количество эпох, окно	Метрики	Наука		Шоу-бизнес	
		Факт	Слух	Факт	Слух
RUWIKI, 10, 5	Recall	0,79	0,70	0,73	0,70
	F1-score	0,73	0,75	0,72	0,72
SG, 100, 5	Recall	0,66	0,84	0,87	0,56
	F1-score	0,71	0,79	0,76	0,67
WIKI-LENTA, 10, 5	Recall	0,80	0,73	0,71	0,73
	F1-score	0,75	0,78	0,72	0,72

Видим, что у моделей, которые используют предобученный Word2Vec, точность для каждого класса и полнота практически равны. F1-score для шоу-бизнеса выше, чем у BOW (1, 3). В моделях, которые используют собственный Word2Vec, точность и полнота для каждого класса отличаются значительно.

Рассмотрим решения с использованием нейронных сетей. RNN – рекуррентная нейронная сеть. LSTM – рекуррентная нейронная сеть, которая способна запоминать значения как на короткие, так и на длинные промежутки времени [22].

В табл. 7 показана точность классификации нейронными сетями для способов векторизации SG, 100, 5, RUWIKI, 10, 5.

Таблица 7. Точность классификации нейронными сетями LSTM и RNN

Способ векторизации, нейронная сеть	Общая точность	Наука		Шоу-бизнес	
		Факт	Слух	Факт	Слух
SG, 100, 5, LSTM	0,757	0,653	0,764	0,570	0,679
SG, 100, 5, RNN	0,588	0,516	0,694	0,587	0,582
RUWIKI, 10, 5, LSTM	0,769	0,779	0,703	0,628	0,618
RUWIKI, 10, 5, RNN	0,544	0,464	0,667	0,528	0,526

Видим, что рекуррентная нейронная сеть не справилась с задачей, а сеть LSTM показала приемлемый результат. Можем сделать вывод, что RNN совсем не подходит для задачи классификации текстов, так как не может запоминать значения на длинные промежутки времени.

Рассмотрим решения с использованием больших языковых моделей. Для больших языковых моделей не выполняется лемматизация, так как они разбивают слова, которые не содержатся у них в словаре, на набор токенов из меньшего количества букв. Кроме этого, они имеют механизм внимания, который дает им лучше понимать контекст, а при использовании лемматизации контекст может теряться.

Задача классификации была решена с использованием LLM моделей: distilbert-base-multilingual-cased, `rubert-base-cased` [23], intfloat/multilingual-e5-large [24], deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B [25]. На выходе используется логистическая регрессия с различными гиперпараметрами.

В табл. 8 показана точность классификации для предобученных языковых моделей с логистической регрессией (для intfloat/multilingual-e5-large дополнительно использовались метод опорных векторов и гребневая регрессия) на выходе и различными значениями коэффициента C, который отвечает за регуляризацию.

Таблица 8. Точность классификации для языковых моделей

LLM модель, метод машинного обучения, параметр	Общая точность	Наука		Шоу-бизнес	
		Факт	Слух	Факт	Слух
distilbert-base-multilingual-cased, логистическая регрессия, C=1	0,733	0,645	0,726	0,668	0,665
distilbert-base-multilingual-cased, логистическая регрессия, C=0,1	0,737	0,635	0,750	0,707	0,670
distilbert-base-multilingual-cased, логистическая регрес-	0,740	0,628	0,750	0,662	0,668

сия, C=10					
rubert-base-cased, логистическая регрессия, C=1	0,780	0,652	0,803	0,717	0,682
rubert-base-cased, логистическая регрессия, C=0,1	0,772	0,633	0,797	0,724	0,688
rubert-base-cased, логистическая регрессия, C=10	0,762	0,630	0,777	0,690	0,653
intfloat/multilingual-e5-large, логистическая регрессия, C=1	0,804	0,673	0,802	0,731	0,711
intfloat/multilingual-e5-large, логистическая регрессия, C=10	0,798	0,671	0,794	0,730	0,700
intfloat/multilingual-e5-large, логистическая регрессия, C=0,1	0,817	0,697	0,812	0,751	0,722
intfloat/multilingual-e5-large, логистическая регрессия, C=0,01	0,830	0,712	0,831	0,776	0,741
intfloat/multilingual-e5-large, метод опорных векторов, C=1	0,836	0,714	0,862	0,797	0,731
intfloat/multilingual-e5-large, гребневая регрессия, alpha=1	0,836	0,722	0,833	0,793	0,754
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B, логистическая регрессия, C=1	0,745	0,620	0,773	0,706	0,682
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B, логистическая регрессия, C=10	0,739	0,598	0,732	0,711	0,692
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B, логистическая регрессия,	0,750	0,631	0,728	0,752	0,692

C=0,1					
deepseek-ai/ DeepSeek-R1-Distill-Qwen-1.5B, логистическая регрессия, C=0,01	0,763	0,652	0,782	0,749	0,723

Модель rubert-base-cased, обученная на части русскоязычной википедии и новостных сайтах, показывает результаты, очень похожие на предобученные модели Word2Vec.

Модель rubert-base-cased показывает результат лучше, чем модель distilbert-base-multilingual-cased, которая обучена на данных википедии на различных языках.

Для distilbert-base-multilingual-cased лучшая точность получена при слабой регуляризации, для intfloat/multilingual-e5-large – при сильной регуляризации, deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B – при сильной регуляризации.

Модель intfloat/multilingual-e5-large показывает результаты лучше, чем модель deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B, хотя имеет в 3 раза меньше параметров (intfloat/multilingual-e5-large – 560 миллионов, deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B – 1,78 миллиарда).

В табл. 9 показаны полнота и f1-score для языковых моделей. Приведены результаты решений с использованием intfloat/multilingual-e5-large, deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B и rubert-base-cased.

Таблица 9. Полнота и f1-score для языковых моделей

LLM модель, метод машинного обучения, параметр	Метрики	Наука		Шоу-бизнес	
		Факт	Слух	Факт	Слух
intfloat/ multilingual-e5-large, логистическая регрессия, C=0,01	Recall	0,82	0,73	0,74	0,79
	F1-score	0,76	0,78	0,76	0,76
deepseek-ai/ DeepSeek-R1-Distill-Qwen-1.5B, логистическая регрессия, C=0,01	Recall	0,77	0,67	0,69	0,74
	F1-score	0,70	0,72	0,72	0,73
rubert-base-cased, логистическая регрессия, C=1	Recall	0,80	0,66	0,66	0,74
	F1-score	0,72	0,72	0,69	0,71

Заметим, что для всех моделей разница между полнотой и точностью для каждого класса больше для классов науки.

Проведем Fine-tuning. Fine-tuning – это тонкая настройка языковой модели. Позволяет дообучить модель на наборе данных и изменить некоторые веса в модели.

Проведем тонкую настройку для языковой модели intfloat/multilingual-e5-large.

В табл. 10 показана точность классификации для тонкой настройки языковой модели intfloat/multilingual-e5-large.

Таблица 10. Точность классификации для тонкой настройки intfloat/multilingual-e5-large

Общая точность	Наука		Шоу-бизнес	
	Факт	Слух	Факт	Слух
0,857	0,745	0,836	0,790	0,820

Видим, что точность после тонкой настройки сильно улучшилась.

В табл. 11 показаны полнота и f1-score для тонкой настройки языковой модели intfloat/multilingual-e5-large.

Таблица 11. Полнота и f1-score для тонкой настройки `intfloat/multilingual-e5-large`

Метрики	Наука		Шоу-бизнес	
	Факт	Слух	Факт	Слух
Recall	0,81	0,78	0,83	0,78
F1-score	0,78	0,80	0,81	0,80

Видим, что после тонкой настройки получен самый высокий средний f1-score для каждого класса.

Можем заметить, что на всех моделях в работе точность для класса «Наука факт» низкая, а разница между полнотой и точностью для каждого класса науки больше, чем разница между полнотой и точностью для каждого класса шоу-бизнеса. Скорее всего, это объясняется тем, что в новостях, связанных с наукой, можно провести более корректное разбиение на классы. Для классов, связанных с шоу-бизнесом, точность и полнота для каждого класса примерно равны во многих моделях (intfloat/multilingual-e5-large с тонкой настройкой, WIKI-LENTA, 10, 5, RUWIKI, 10, 5), поэтому в новостях шоу-бизнеса разбиение на классы проведено удачно. Кроме этого, новости шоу-бизнеса сложнее анализировать и проводить их разбиение, но точность у классов шоу-бизнеса и науки примерно одинакова, это тоже говорит о том, что разбиение новостей науки можно провести лучше.

Для языковых моделей было проведено сравнение различных гиперпараметров, теперь проведем сравнение для моделей, которые использовали методы векторизации TF-IDF, Bag-of-words, Word2Vec.

Будем проводить поиск с использованием кросс-валидации. Разобьем весь датасет на 10 частей и составим 10 выборок, в каждой будет 9 тренировочных и 1 тестовая часть и все тестовые часть будут различными. Найдем точность на каждой, затем найдем среднее, выберем лучшую из моделей с различными гиперпараметрами. Создадим эту модель для нашей выборки с тренировочной и тестовой частями и сравним результаты с уже имеющимися выше.

При решении без подбора гиперпараметров в логистической регрессии отличалось от стандартного значения max_iter=1000, в методе опорных векторов – kernel=linear.

Для логистической регрессии использовались C= [0.1, 1, 10, 100], max_iter=[1000, 10000]. Для гребневой регрессии: alpha=[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0].

Будем подбирать гиперпараметры для методов машинного обучения, которые показали лучшую точность для метода векторизации.

В табл. 12 показана точность классификации для моделей и указаны гиперпараметры, подобранные для них.

Таблица 12 Точность классификации с гиперпараметрами

Метод векторизации, метод машинного обучения, (гиперпараметры)	Общая точность	Общая точность для модели без подбора гиперпаметров
SG, 100, 5, Логистическая регрессия (max_iter=1000), модели совпали	0,802	0,802
TF-IDF (1, 1), метод опорных векторов, (kernel=linear), модели совпали	0,820	0,820
TF-IDF (1, 2), гребневая регрессия, (alpha=0,6)	0,827	0,829
BOW (1, 3), логистическая регрессия, (стандартные, max_iter=100)	0,808	0,804

Видим, что подбор гиперпараметров не всегда дает прирост точности, многие модели уже имели оптимальные параметры, несмотря на то, что подбор до этого не производился.

Для оценки влияния гиперпараметров на результаты проведём ещё один эксперимент. Для большой языковой модели intfloat/multilingual-e5-large, TF-IDF (1, 3), SG, 100, 5 проведём сравнение результатов с логистической регрессией на выходе и гиперпараметром C, который отвечает за регуляризацию. Будем использовать гиперпараметры 0.01, 0.1, 1, 10.

В табл. 13 приведены результаты сравнения e5-large, TF-IDF и Word2Vec с гиперпараметрами 0.01, 0.1, 1, 10 и разностью между максимальной точностью и минимальной для каждой модели.

Таблица 13. Сравнение влияния гиперпараметров на точность модели

Модель	C=0,01	C=0,1	C=1	C=10	Разность
intfloat/multilingual-e5-large, логистическая регрессия	0,830	0,817	0,804	0,798	0,032
SG, 100, 5, логистическая регрессия	0,737	0,779	0,803	0,791	0,066
TF-IDF (1, 3), логистическая регрессия	0,645	0,762	0,804	0,817	0,172

Видим, что самая большая разность между максимальной точностью и минимальной у модели с векторизацией TF-IDF, самая маленькая – у intfloat/multilingual-e5-large. Рассмотрим TF-IDF подробнее. Она при униграммах, биграммах и триграммах имеет большую размерность. Размерность для TF-IDF – количество уникальных n-грамм. При C=0.01 (сильная регуляризация) получена наименьшая точность, так как n-граммы, которые использовались редко, становятся значениями, близкими к 0, и никак не влияют на точность. Оказывают влияние только n-граммы, которые встречались чаще всего. Из-за этого тексты становятся слишком похожи друг на друга. При увеличении C точность монотонно возрастает. Заметим, что при использовании TF-IDF нет переобучения даже при C=10, и редкие n-граммы оказывают большое влияние на точность.

С Word2Vec при C=10 заметно переобучение, точность при C=1 больше. Точность монотонно возрастает при увеличении C от 0.01 до 1.

Intfloat/multilingual-e5-large имеет размерность 1024. Точность монотонно убывает при увеличении C, но разброс результатов гораздо меньше, чем у TF-IDF. Модель уже предобучена на большом количестве данных и понимает контексты. При сильной регуляризации она не теряет редкие значения полностью, так как имеет для каждого значения вектор размерности 1024, а TF-IDF для каждой n-граммы имеет 1 значение, которое не оказывает влияние на результат при стремлении к 0. Так как для intfloat/multilingual-e5-large все значения оказывают влияние на результат, разница в точности при различных гиперпараметрах меньше, чем у TF-IDF.

4. Ограничения предложенных подходов

В работе создан набор данных, его можно расширять и добавлять новые классы, но классы должны быть такими, чтобы была возможность составить правила для разбиения на слухи и факты. В [2] была произведена попытка разделения новостей политики, но было слишком сложно составить какие-либо правила, которые разделяли новости, поэтому этот класс пришлось удалить.

Для нормальных результатов с TF-IDF и Bag-of-words лучше работать с небольшим датасетом, особенно при использовании n-грамм. Чем больше датасет, тем больше размерность TF-IDF и Bag-of-words, тем хуже могут стать результаты.

Предобученные модели Word2Vec, использованные в работе, не были обучены для работы с новостями шоу-бизнеса, поэтому они выдают низкую точность на таких новостях. Изменить это может fine-tuning.

5. Заключение

Для решения задачи классификации создан набор данных, состоящий из шести различных классов. Датасет содержит новости науки, политики и спорта. Все новости разбиты на классы самостоятельно с помощью продукционных правил. Представлены сравнительные результаты применения различных способов векторизации: TF-IDF, Bag-of-words (с использованием различных n-грамм), Word2Vec. Для Word2Vec было проведено сравнение предобученных на больших корпусах данных моделей и собственных. Собственные модели использовали два алгоритма обучения: skip-gram и CBOW, кроме этого, сравнивались результаты с различным количеством эпох обучения и размером окна. Задача решалась различными методами машинного обучения, нейронными сетями LSTM и RNN, большими языковыми моделями (multilingual-e5-large, DeepSeek-R1-Distill-Qwen-1.5B, distilbert-base-multilingual-cased, rubert-base-cased). Решения с использованием способа векторизации TF-IDF оказались лучше решений с использованием Word2Vec. Решение с использованием собственных Word2Vec лучше решений с использованием предобученных Word2Vec. При увеличении количества итераций в собственном Word2Vec увеличивается точность. Решения с использованием методов машинного обучения показали себя лучше решений с использованием нейронных сетей. Решения с использованием RNN не справились с задачей и показали плохие результаты. Решения с использованием rubert-base-cased показали схожие результаты с предобученными моделями Word2Vec. Решения с использованием multilingual-e5-large показали лучшие результаты. Решения с использованием DeepSeek-R1-Distill-Qwen-1.5B сильно уступили multilingual-e5-large. Тонкая настройка большой языковой модели даёт большой прирост точности. Большие языковые модели показывают лучшие результаты на небольшом датасете при сильной регуляризации, а методы векторизации (TF-IDF, Word2Vec) – при слабой.

В дальнейших исследованиях будет расширено количество классов в датасете, увеличено количество новостей в каждом классе, проведено сравнение других языковых моделей, например, LaBSE.

Литература

1. Лингвистическая экспертиза в трудах Воронежской ассоциации экспертов-лингвистов : монография / Ж. В. Грачева [и др.] ; под ред. М. Е. Новичихиной, А. В. Рудаковой. – Воронеж : Издательский дом ВГУ, 2023. – 357 с.
2. *Худобин А. Д.* TF-IDF, Bag-of-words, Word2Vec и N-граммы для решения задачи классификации слухов в новостях / А. Д. Худобин, И. Е. Воронина // Информатика: проблемы, методы, технологии. – 2024. – С. 1285–1294.
3. *Осетрова Е. В.* Слухи в парадигме лингвистической генристики / Е. В. Осетрова // Жанры речи. – 2015. – Т. 12, № 2. – С. 80–89.
4. *Osetrova E. V.* Rumours as a Subject of Scientific Analysis: Social Psychology, History, Philology / E. V. Osetrova // Journal of Siberian Federal University. Humanities & Social Sciences. – 2013. – № 9. – С. 1265–1280.
5. *Русакова И. Б.* Концепт слух и его репрезентация в русских и английских пословицах / И. Б. Русакова // Международный аспирантский вестник. Русский язык за рубежом. – 2024. – № 1. – С. 60–67.
6. *Иванова С. В.* Жанр светских слухов в дискурсе англоязычных массмедиа / С. В. Иванова, Г. Ш Хакимова // Russian Journal of Linguistics. – 2020. – Т. 24, № 2. – С. 386–418.
7. *Khakimova G.* Rumour Text Constructing Techniques In Media Discourse: Case Study Of Gossip Columns / The European Proceedings of Social and Behavioural Sciences. – 2019. – С. 346–357.
8. *Хакимова Г. Ш.* Концепт «слух» в фокусе лексикографического анализа. Часть 1 / Г. Ш. Хакимова // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – № 2. – С. 29–36.
9. *Хакимова Г. Ш.* Концепт «слух» в фокусе лексикографического анализа. Часть 2 / Г. Ш. Хакимова // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – № 3. – С. 38–46.
10. *Пастревич М. К.* Автоматическая классификация речевой агрессии / М. К. Пастревич, И. Е. Воронина // Актуальные проблемы прикладной математики, информатики и механики. – 2023. – С. 1359–1365.
11. *Roumeliotis K. I.* Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification / K. I. Roumeliotis, N. D. Tselika, D. K. Nasiopoulos // Electronics. – 2024. – vol. 13, no. 11. – С. 2034–2058.
12. *Li X.* Dice Loss for Data-imbalanced NLP Tasks / X. Li [и др.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020. – С. 465–476.
13. *Худобин А. Д.* Подход к решению задачи классификации слухов в интернет-новостях / А. Д. Худобин, И. Е Воронина // Актуальные проблемы прикладной информатики, математики и механики. – 2024. – С. 334–340.
14. *Duncan J.* The Sensitivity of Word Embeddings-based Author Detection Models to Semantic-preserving Adversarial Perturbations / J. Duncan [и др.] // arXiv preprint arXiv:2102.11917. – 2021. – 22 с.
15. *Полицына, Е. В.* Анализ качества работы и расширение возможностей инструментов морфологического анализа текстов на русском языке / Е. В. Полицына, С. А. Полицын, А. С. Поречный, А. Н. Рыкунов // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2023. – № 2. – С.171-180.
16. *Хобсон Л.* Обработка естественного языка в действии / Л. Хобсон, Х. Хапке, К. Ховард. – СПб.: Питер, 2020. – 576 с.
17. *Mohare R. V.* ‘Bag of Words’ to ‘Bag of Concepts’: Improving Text Categorization using SVM / R. V. Mohare [и др.] // Nanotechnology Perceptions. – 2024. – № S6. – С. 419–426.

18. *Suhasini V.* A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data / V. Suhasini, Dr. N. Vimala // Turkish Journal of Computer and Mathematics Education. – 2021. – vol. 12, no. 06. – С. 5710–5723.
19. *Cahyani D. E.* Performance comparison of TF-IDF and Word2Vec models for emotion text classification / D. E. Cahyani, I. Patasik // Bulletin of Electrical Engineering and Informatics. – 2021. – № 5 – С. 2780–2788.
20. *Abubakar H. D.* Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec / H. D. Abubakar, M. Umar, M. A. Bakale // Sule Lamido University Journal of Science & Technology. – 2022. – № 1 & 2. – С. 27–33.
21. *Линькова Г. В.* Сравнение методов векторных представлений текстов в задачах классификации вакансий / Г. В. Линькова, Д. С. Ботов // Научное сообщество студентов XXI столетия. Технические науки. – 2018. – № 6 – С. 96–113.
22. *Santos R. F.* Long Term-short Memory Neural Networks and Word2vec for Self-admitted Technical Debt Detection / R. F. Santos [и др.] // Proceedings of the 22nd International Conference on Enterprise Information Systems. – 2020. – С. 157–165.
23. *Devlin J.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [и др.] // Proceedings of NAACL-HLT. – 2019. С. 4171–4186.
24. *Wang L.* Multilingual E5 Text Embeddings: A Technical Report / L. Wang [и др.] // arXiv preprint arXiv:2402.05672. – 2024. – 6 с.
25. Bi X. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism / X. Bi [и др.] // arXiv preprint arXiv:2401.02954. – 2024. – 48 с.

Худобин Александр Дмитриевич

магистрант 2-го года обучения кафедры программного обеспечения и администрирования информационных систем Воронежского государственного университета. (Воронежский государственный университет, Университетская пл., 1, 394018, Воронеж, Российская Федерация), тел. +7(960)1013772, e-mail: 2001kad@mail.ru, ORCID ID: 0009-0004-5786-8107.

Воронина Ирина Евгеньевна

д-р. техн. наук, доц., профессор кафедры программного обеспечения и администрирования информационных систем Воронежского государственного университета. (Воронежский государственный университет, Университетская пл., 1, 394018, Воронеж, Российская Федерация), тел. +7(903)6504410, e-mail: irina.voronina@gmail.com, ORCID ID: 0000-0003-0040-5764.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: *Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.*

Solving the problem of rumors classification in online news

Alexandr D. Khudobin, Irina E. Voronina

Voronezh state university (VSU)

Abstract: The article considers an approach to solving the problem of classifying rumors in the news based on production rules. Unverified information appearing on news sites has the nature of information garbage and is capable of causing significant harm to consumers in some cases. The problem being solved is non-trivial, relevant and has no standard solution.

Keywords: rumors, classification, LLM, machine learning, neural networks, TF-IDF, Bag-of-words, Word2Vec, N-grams, production rules, DeepSeek-R1, e5-large.

For citation: Khudobin A. D., Voronina I. E. Reshenie zadachi klassifikacii sluhov v internet-novostyah [Solving the problem of rumours classification in online news]. *Vestnik SibGUTI*, 2025, vol. 19, no. 3, pp. 122-138. https://doi.org/10.55648/1998-6920-2025-19-3-122-138.

The article was submitted: 13.04.2025;
revised version: 20.05.2025;
accepted for publication 30.05.2025.

References

16. Gracheva Zh. V. Lingvisticheskaja jekspertiza v trudah Voronezhskoj associacii jekspertov-lingvistov : monografija [Linguistic expertise in the works of the Voronezh Association of Linguistic Experts: monograph]. Voronezh, Izdatel'skij dom Voronezhskogo gosudarstvennogo universiteta, 2023. 357 p.
17. Khudobin A. D., Voronina I. E. TF-IDF, Bag-of-words, Word2Vec i N-grammy dlja reshenija zadachi klassifikacii sluhov v novostjah [TF-IDF, Bag-of-words, Word2Vec and N-grams for News Rumors Classification Problem]. *Informatika: problemy, metody, tehnologii*, 2024, pp. 1285–1294.
18. Osetrova E. V. Rumors in the paradigm of linguistic genristics. *International Journal "Speech Genres"*, 2015, vol. 12, no. 2, pp. 80–89.
19. Osetrova E. V. Rumours as a Subject of Scientific Analysis: Social Psychology, History, Philology. *Journal of Siberian Federal University. Humanities & Social Sciences*, 2013, no. 9, pp. 1265–1280.
20. Rusakova I. B. Koncept sluh i ego reprezentacija v russkih i anglijskih poslovicah [The concept of hearing and its representation in Russian and English proverbs]. *Mezhdunarodnyj aspirantskij vestnik. Russkij jazyk za rubezhom*, 2024, no. 1, pp. 60–67.
21. Ivanova S. V, Khakimova G. Celebrity gossip as a genre in English-language mass media discourse. *Russian Journal of Linguistics*, 2020, vol. 24, no. 2, pp. 386–418.
22. Khakimova G. Rumour Text Constructing Techniques In Media Discourse: Case Study Of Gossip Columns. *The European Proceedings of Social and Behavioural Sciences*, 2019, pp. 346–357.
23. Khakimova G. CONCEPT "RUMOUR" IN TERMS OF LEXICOGRAPHIC ANALYSIS. PART 1. *Bulletin of the South Ural State University series Linguistics*, 2016, vol. 13, no. 2, pp, 29–36
24. Khakimova G. CONCEPT "RUMOUR" IN TERMS OF LEXICOGRAPHIC ANALYSIS. PART 2. *Bulletin of the South Ural State University series Linguistics*, 2016, vol. 13, no. 3, pp, 38–46
25. Pastrevich M. K., Voronina I. E. Avtomaticheskaja klassifikacija rechevoj agressii [Automatic classification of speech aggression]. Aktual'nye problemy prikladnoj matematiki, informatiki i mehaniki, 2023, pp. 1359–1365.
26. Roumeliotis K. I., Tselika N. D., Nasiopoulos D. K. Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification. Electronics, 2024. vol. 13, no. 11, pp. 2034–2058.
27. Li X. et. al. Dice Loss for Data-imbalanced NLP Tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 465–476.
28. Khudobin A. D., Voronina I. E. Podhod k resheniju zadachi klassifikacii sluhov v internet-novostjah [An approach to solving the problem of rumors classification in online news]. Aktual'nye problemy prikladnoj matematiki, informatiki i mehaniki, 2024, pp. 334–340.
29. Duncan J. et. al. The Sensitivity of Word Embeddings-based Author Detection Models to Semantic-preserving Adversarial Perturbations. 2021, 22 p., available at: https://arxiv.org/pdf/2102.11917 (accessed 13.04.2025).
30. Policyna, E. V., Policyn S. A., Porechnyj A. S., Rykunov A. N. Analiz kachestva raboty i rasshirenie vozmozhnostej instrumentov morfologicheskogo analiza tekstov na russkom jazyke [Analysis of the quality of work and expansion of the capabilities of tools for mor-

phological analysis of texts in Russian]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Sistemnyj analiz i informacionnye tehnologii, 2023, no. 2, pp. 171–180.
31. Hobson L., Hapke H., Hovard K. Obrabotka estestvennogo jazyka v dejstvii [Natural Language Processing in Action]. Saint Petersburg, Piter, 2020, 576 p.
32. Mohare R. V. et. al. 'Bag of Words' to 'Bag of Concepts': Improving Text Categorization using SVM. *Nanotechnology Perceptions*, no. S6, pp.419–426.
33. Suhasini V., Vimala Dr. N. A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data. *Turkish Journal of Computer and Mathematics Education*, 2021, vol. 12, no. 06, pp. 5710–5723.
34. Cahyani D. E., Patasik I. Performance comparison of TF-IDF and Word2Vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, 2021, no. 5, pp. 2780–2788.
35. Abubakar H. D., Umar M., Bakale M. A. Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *Sule Lamido University Journal of Science & Technology*, 2022, no. 1 & 2, pp. 27–33.
36. Lin'kova G. V., Botov D. S. Sravnenie metodov vektornyh predstavlenij tekstov v zadachah klassifikacii vakansij [Comparison of Text Vector Representation Methods in Job Classification Problems]. Nauchnoe soobshhestvo studentov XXI stoletija. Tehnicheskie nauki, 2018, no. 6, pp. 96–113.
37. Santos R. F. et. al. Long Term-short Memory Neural Networks and Word2vec for Self-admitted Technical Debt Detection. *Proceedings of the 22nd International Conference on Enterprise Information Systems*, 2020, pp. 157–165.
38. Devlin J. et. al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
39. Wang L. et. al. Multilingual E5 Text Embeddings: A Technical Report. 2024, 6 p., available at: `https://arxiv.org/pdf/2402.05672` (accessed 13.04.2025).
40. Bi X. et. al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. 2024, 48 p., `https://arxiv.org/pdf/2401.02954` (accessed 13.04.2025).

Alexandr D. Khudobin

2nd year Master's student of Software Development and Information Systems Administration Department of Applied Mathematics, Informatics and Mechanics Faculty Voronezh State University. (Voronezh State University, Universitetskaya sq., 1, 394018, Voronezh, Russian Federation), phone: +7(960)1013772, e-mail: `2001kad@mail.ru`, ORCID ID: 0009-0004-5786-8107.

Irina E. Voronina

DSc in Technical Sciences, Professor of Software Development and Information Systems Administration Department of Applied Mathematics, Informatics and Mechanics Faculty. (Voronezh State University, Universitetskaya sq., 1, 394018, Voronezh, Russian Federation), phone: +7(903)6504410, E-mail: `irina.voronina@gmail.com` ORCID ID: 0000-0003-0040-5764.