Желенков Борис Владимирович, доцент, заведующий кафедры «Вычислительные системы, сети и информационная безопасность», г. Москва
Замуруев Александр Владимирович, аспирант, Российский Университет Транспорта (МИИТ), г. Москва
Мымрин Павел Михайлович, аспирант, Российский Университет Транспорта (МИИТ), г. Москва

ИССЛЕДОВАНИЕ ВЛИЯНИЯ МЕТОДОВ ПРЕДОБРАБОТКИ ДАННЫХ НА ЭФФЕКТИВНОСТЬ НЕЙРОСЕТЕВОГО АНАЛИЗА ТОНАЛЬНОСТИ ИНТЕРНЕТ-ОТЗЫВОВ

Аннотация

Проведено экспериментальное тестирования различных комбинаций алгоритмов предварительной обработки текста и нейросетевых архитектур на массивах данных разного объема. Проведены испытания с целью эмпирической оценки влияния простой очистки, стемминга и лемматизации на результирующее качество классификации и временные ресурсы, необходимые для обучения моделей. Определено, что сверточные нейросети в базовом виде в сочетании со стеммингом обеспечивают наилучший баланс точности и скорости обучения, а основным источником ошибок и нестабильности на этапе классификации являются рекуррентные сети, склонные к переобучению на несбалансированных текстовых выборках. В заключении можно сказать, что превосходство сверточных архитектур обусловлено спецификой интернет-отзывов, представляющих собой короткие зашумленные тексты со слабой синтаксической связанностью, вследствие чего необходимость использования локальных n-грамм для извлечения эмоционально окрашенной лексики была подтверждена.

Ключевые слова: нейросетевые архитектуры, методы предобработки, анализ тональности, интернет-отзывы.

ВВЕДЕНИЕ

Стремительное развитие веб-технологий и платформ электронной коммерции привело к резкому увеличению объемов генерируемого пользователями контента (User Generated Content, UGC). Отзывы клиентов о товарах и услугах стали ключевым источником данных для бизнес-аналитики, управления репутацией и мониторинга качества. В связи с этим задача автоматического анализа тональности (Sentiment Analysis) коротких текстов приобрела высокую практическую значимость.

Основная проблема при обработке русскоязычных интернет-отзывов заключается в их высокой «зашумленности». Авторы часто используют сленг, эмодзи, ненормативную лексику, допускают орфографические ошибки и нарушают правила синтаксиса. Кроме того, стандартные алгоритмы предобработки (например, автоматическое удаление стоп-слов из библиотек типа NLTK) нередко искажают контекст: удаление отрицательной частицы «не» или союзов может полностью инвертировать тональность высказывания с позитивной на негативную и наоборот.

В то же время современные глубокие нейросети (Deep Neural Networks) демонстрируют высокую точность в задачах NLP, однако их использование сопряжено со значительными вычислительными затратами. В инженерной практике при проектировании прикладных систем разработчики сталкиваются с дилеммой «точность против производительности».

Цель данного исследования – оценка влияния различных стратегий текстовой предобработки и типов нейросетевых архитектур (MLP, CNN, LSTM) как на результирующую метрику качества классификации (Accuracy, F1-score), так и на время, необходимые для обучения моделей.

1. ИСХОДНЫЕ ДАННЫЕ

Для достижения поставленной цели и обеспечения воспроизводимости результатов был разработан комплексный план экспериментального исследования. Методология базируется на сквозном тестировании различных комбинаций алгоритмов предварительной обработки и нейросетевых архитектур на массивах данных разного объема. Все вычисления производились в идентичной программно-аппаратной среде, что позволило объективно сопоставить временные затраты на обучение моделей. Ниже детально описаны использованные датасеты, примененные стратегии нормализации текста и математические конфигурации исследуемых нейросетей.

1.1. Наборы данных

Для проведения экспериментов использовались три размеченных датасета русскоязычных отзывов различных объемов. Данные для обучения представлены в виде .CSV таблиц, в которых данные находятся в виде строк (первая строка заголовок отсутствует) «ОТЗЫВ, ОЦЕНКА».

- IMDB.csv – выборка малого объема (6 000 записей), состоящая из отзывов на кинематограф;
- clothes.csv – выборка среднего объема (12 000 записей), состоящая из отзывов на одежду с онлайн-магазинов;
- phone.csv – выборка большого объема (21 000 записей), состоящая из отзывов на различные телефоны.

Данный выбор был сделан по двум причинам:

1. Различная величина данных в датасетах – позволит оценить реакцию моделей на различный объем данных для обучения, чтобы выявить оптимальный

диапазон данных, при котором не будет возникать переобучений или заучиваний.

2. Различная эмоциональная окраска – датасеты взяты из различных областей, что позволит проверить лексическую универсальность моделей.

1.2. Стратегии предобработки текста

В исследовании сравнивались три подхода к нормализации исходного текста:

- Simple (Простая очистка) – базовая токенизация, удаление пунктуации и приведение к нижнему регистру без глубокого морфологического анализа;
- Stemmed (Стемминг) – отсечение окончаний и суффиксов слов для приведения их к псевдооснове;
- Lemmatized (Лемматизация) – полноценное приведение слов к начальной словарной форме с учетом контекста и части речи.

Данные методы были выбраны не случайно. Сравнение базовой очистки, стемминга и лемматизации покажет, оправдывает ли прирост точности классификации усложнение алгоритмов предварительной обработки, или же современные нейросетевые архитектуры способны эффективно работать с менее обработанными данными.

1.3. Исследуемые архитектуры

Для классификации были спроектированы и обучены следующие модели искусственных нейронных сетей:

- MLP (Многослойный перцептрон) – базовая (MLP_B) и модифицированная (MLP_D) версии полносвязных сетей;

- CNN (Сверточная нейросеть) – базовая (CNN_B) и расширенная (CNN_E) конфигурации с применением 1D-сверток для извлечения локальных признаков;
- LSTM (Долгая краткосрочная память) – базовая (LSTM_B) и оптимизированная (LSTM_E) рекуррентные сети, предназначенные для учета долгосрочных зависимостей в последовательностях текста.

Для комплексной оценки моделей применяется следующий набор метрик:

- Качество классификации – Accuracy (точность);
- Качество определенного класса – recall (точность для каждого из трех классов – Positive, Neutral, Negative).

1.4. Методика исследования

В основу данной работы положен сравнительный эксперимент, направленный на выявление оптимального сочетания архитектуры нейронной сети, способа предобработки текста и объема входных данных. Исследование реализуется в четыре этапа:

1. Подготовка и стратификация данных:

На первом этапе формируется экспериментальная база из трех наборов данных различного объема и тематики.

2. Многоуровневая предобработка текста:

Для каждого датасета применяются три независимых сценария обработки текстовой информации.

3. Построение и обучение моделей:

В рамках исследования сравниваются три архитектурных подхода к анализу текста.

4. Оценка и сравнительный анализ:

Заключительный этап включает в себя кросс-валидацию и замер метрик эффективности (Accuracy, recall) для каждой комбинации «Датасет + Метод

обработки + Модель». Это позволит определить, как глубина морфологического анализа влияет на точность классификации в зависимости от размера выборки.

1.5. Набор инструментов для решения задач

Эксперименты проводились с использованием современных технологий обработки естественного языка и машинного обучения. Основой системы выбран язык программирования Python версии 3.10.11. Для работы с нейросетевыми архитектурами использовался фреймворк TensorFlow в версии 2.19.0 в сочетании с высокоуровневым API Keras (версия 3.9.2).

Обработка текстовых данных реализована с помощью комбинации нескольких специализированных библиотек. Для токенизации и базовой обработки текста применялась библиотека NLTK (версия 3.9.1), предоставляющая надежные алгоритмы работы с естественным языком. Лемматизация для русского языка выполнена с использованием связки pymorphy2 (версия 0.9.1) и pymorphy2–dicts–ru (версия 2.4.417127.4579844).

2. РЕЗУЛЬТАТЫ

Для анализа эффективности исследуемых моделей было проведено сквозное тестирование 6 архитектур в комбинации с тремя методами предобработки на трех наборах данных разного объема. Далее в подразделах рассмотрим обучение на каждом датасете. Для информации будут приведены название модели, метод предобработки, точность модели, точность определения положительных и негативных отзывов в обучающей выборке.

2.1. Результаты на малом наборе данных (IMDB)

На небольшом объеме данных все модели показали относительно низкую точность. Данные представлены ниже в таблице 1, где отображены наилучшие 5 результатов обучений:

Таблица 1. Сравнительные результаты на датасете IMDB (6 000 записей).

Модель	**Предобработка**	**Accuracy**	**Pos Recall**	**Neg Recall**	**Neutral Recall**
CNN_B	stemmed	0,6536	0,7042	0,7483	0,5083
CNN_E	stemmed	0,6464	0,6575	0,7375	0,5442
CNN_B	lemmatized	0,6444	0,6342	0,7725	0,5267
CNN_B	simple	0,6405	0,7233	0,7317	0,4667
LSTM_E	lemmatized	0,6369	0,6900	0,7150	0,5058

Полученные данные связаны со сложностью тональной разметки и недостаточным объемом обучающей выборки для глубоких сетей. Сверточные

сети CNN_B и CNN_B на стемминге показали лучшие результаты, показав хорошее определение позитивных и негативных отзывов, однако нейтральные отзывы поддавались определению одинаково хуже для всех моделей и предобработок.

2.2. Результаты на среднем наборе данных (Clothes)

С увеличением объема данных качество классификации существенно возросло для всех архитектур (кроме LSTM). Данные приведены в таблице 2:

Таблица 2. Сравнительные результаты на датасете clothes (12 000 записей).

Модель	Предобработка	Accuracy	Pos Recall	Neg Recall	Neutral Recall
CNN_B	stemmed	0,8725	0,8721	0,9804	0,7650
CNN_B	stemmed	0,8645	0,8012	0,9888	0,8033
CNN_B	simple	0,8613	0,7571	0,9783	0,8483
MLP_B	stemmed	0,8568	0,7650	0,9612	0,8441
CNN_E	simple	0,8525	0,8171	0,9667	0,7738
MLP_B	simple	0,8486	0,6617	0,9671	0,9171
LSTM_B	stemmed	0,3334	1,0000	0,0000	0,0000
LSTM_B	simple	0,3334	1,0000	0,0000	0,0000
LSTM_B	lemmatized	0,3334	1,0000	0,0000	0,0000

Максимальная точность зафиксирована у CNN_B со стеммингом. При этом модели LSTM на всех видах предобработки показала некорректное

поведение (Accuracy ~ 0.33, Recall для позитивного класса = 1.0, для остальных = 0.0), что свидетельствует о переобучении и сваливании модели в мажоритарный класс. Так же выделяется тенденция четкого определения негативных отзывов. Причина такого поведения потенциально в неоднородной структуре отзывов на одежду, где четко выражен негатив на брак товара, в то время как позитивные и нейтральные отзывы имеют различный эмоциональный окрас.

2.3. Результаты на большом наборе данных (Phone)

Наибольший интерес для высоконагруженных IT-систем представляют результаты на датасете максимального объема (Таблица 3).

Таблица 3. Сравнительные результаты на датасете Phone (21 000 записей)

Модель	**Предобработка**	**Accuracy**	**Pos Recall**	**Neg Recall**	**Neutral Recall**
CNN_B	simple	0,8112	0,6393	0,8170	0,9772
CNN_B	stemmed	0,8001	0,6723	0,8259	0,9021
CNN_E	simple	0,7935	0,6425	0,8247	0,9134
CNN_E	stemmed	0,7928	0,6144	0,8220	0,9420
MLP_B	stemmed	0,7778	0,6613	0,8182	0,8538
CNN_B	lemmatized	0,7757	0,6996	0,7088	0,9200
LSTM_B	stemmed	0,3334	1,0000	0,0000	0,0000
LSTM_B	lemmatized	0,3334	1,0000	0,0000	0,0000
LSTM_B	simple	0,3334	1,0000	0,0000	0,0000

Полученные данные ниже, чем на датасете с одеждой, однако наблюдается общая тенденция, что LSTM модели не подходят для выявления эмоциональных закономерностей. Так же относительно предыдущего обучения, наблюдается тенденция четкого определения нейтральных отзывов, а также низкий процент определения положительных и отрицательных отзывов (относительно нейтральных). Связь в этом заключается в отзывах, содержащих в основном сухую, техническую информацию, из-за чего сложно уловить эмоциональные взаимосвязи.

3. АНАЛИЗ РЕЗУЛЬТАТОВ

Полученные результаты ярко демонстрируют превосходство сверточных архитектур (CNN) над рекуррентными (LSTM) как по метрикам качества, так и по временной эффективности при анализе тональности коротких русскоязычных отзывов.

Такое распределение эффективности объясняется спецификой исследуемого материала. Интернет-отзывы пользователей представляют собой короткие тексты с высокой концентрацией эмоционально окрашенной лексики и сленга. В таких условиях решающее значение приобретают локальные n-граммы и ключевые слова (например, «отличный», «не понравился», «быстро сломался»), которые успешно извлекаются сверточными фильтрами CNN. В то же время сложные грамматические и долгосрочные синтаксические связи, для фиксации которых предназначены блоки LSTM, в коротких зашумленных отзывах практически отсутствуют или нивелируются спецификой интернет-сленга.

ЗАКЛЮЧЕНИЕ

В ходе работы было проведено комплексное эмпирическое сравнение трех типов нейросетевых архитектур в связке с различными методами предобработки зашумленного русскоязычного контента.

1. Опираясь на результаты исследования можно заключить, что сверточные нейросети в базовой комплектации (CNN_B) в сочетании с приведением слова к основе (stemmed) обеспечивают наилучший баланс точности и скорости обучения на больших массивах данных.

2. Выявлена нестабильность рекуррентных сетей (LSTM) на несбалансированных текстовых выборках.

На основе полученных данных выдвигается гипотеза о том, что превосходство CNN над LSTM обусловлено исключительно малой длиной и слабой синтаксической связанностью текстов отзывов.

СПИСОК ЛИТЕРАТУРЫ

1. Голдберг Й. Методы нейронных сетей для обработки естественного языка [Текст]. – М.: ДМК Пресс, 2019.

2. Журавлёв Ю.И., Ковалёв М.М. Методы машинного обучения и интеллектуального анализа данных [Текст]. – М.: ФИЗМАТЛИТ, 2020.

3. Тихомиров В.П., Ковальчук В.В. Интеллектуальный анализ текстовой информации [Текст]. – М.: Горячая линия – Телеком, 2018.