Отчёт по заданию:

Классификация текстовых данных с использованием нейронных сетей

1. Описание датасета

- Параметры данных

Датасет содержит текстовые комментарии на английском и русском языках с метками токсичности. Английская часть включает 159 571 запись, русская часть - 248 290 записей. Каждая запись содержит текст комментария и три метки: нейтральный, токсичный без угрозы и токсичный с угрозой.

- Качество данных

Данные обладают полным покрытием без пропущенных значений, все записи уникальны и не содержат дубликатов. Тексты прошли базовую предобработку, включающую приведение к нижнему регистру и удаление специальных символов.

Технические параметры показывают, что средняя длина текста составляет 394 символа (67 слов), с медианной длиной 205 символов. Около 9% записей являются выбросами по длине. Качество сегментов высокое: 96% данных классифицируются как чистые, лишь 4% содержат шум.

Лингвистический анализ выявил среднюю длину предложения 13 слов, уровень читаемости 61 из 100 баллов, богатство словаря на уровне 0.815. Доля специальных символов составляет 5%, цифр - 1%. Общая оценка качества данных - 79.5 из 100 баллов, что соответствует среднему уровню.

Исходное распределение меток демонстрировало значительный дисбаланс, особенно для класса токсичных комментариев с угрозами. Для обеспечения качественного обучения была проведена балансировка классов путем выборки равного количества примеров для каждой категории.

2. Метод предобработки данных

Для подготовки текстовых данных к обучению использовался комплексный метод предобработки. Тексты приводились к нижнему регистру, очищались от специальных символов и лишних пробелов. После токенизации выполнялась обрезка до 50 слов с последующей векторизацией с использованием предобученных моделей FastText. Для унификации размерности применялось дополнение нулями до фиксированной длины.

3. Методы классификации

- Реализованные архитектуры

Были реализованы три типа нейронных сетей: сверточная сеть (CNN), двунаправленная LSTM и двунаправленная GRU с механизмом внимания. CNN архитектура использовала многоканальный подход с различными размерами ядер свертки, в то время как рекуррентные сети применяли многослойную структуру с механизмами регуляризации.

- Обоснование выбора архитектур

В CNN-архитектуре была применена комбинация max-pooling и average-pooling. Max-pooling эффективно выделяет наиболее значимые признаки, такие как ключевые слова и эмоционально окрашенные выражения, в то время как average-pooling сохраняет общий контекст и тональность текста. Синергетический эффект от комбинации этих подходов позволил улучшить точность классификации на 30% по сравнению с использованием только одного типа пулинга.

Выбор между LSTM и GRU обусловлен их комплементарными преимуществами. LSTM демонстрирует высокую эффективность в работе с длинными зависимостями в текстах и устойчивость к проблеме исчезающего градиента. GRU предлагает более простую архитектуру с меньшим количеством параметров, что обеспечивает более быстрое обучение при сопоставимом качестве и снижает риск переобучения.

Все рекуррентные сети реализованы как двунаправленные, что критически важно для задач анализа текста. Такой подход позволяет модели анализировать контекст в обоих направлениях, значительно улучшая понимание смысла. Это особенно важно для detection токсичности, где отрицательные частицы могут полностью менять смысл последующих слов.

В отличие от стандартного подхода использования только последнего скрытого состояния, в реализованных моделях задействовались выходы всех скрытых слоев. Это обеспечивает более богатое представление данных, поскольку каждый слой извлекает признаки разного уровня абстракции. Такой подход также повышает устойчивость к шуму и не зависит от позиции ключевой информации в тексте. Экспериментально подтверждено, что использование только последнего состояния дает неудовлетворительные результаты около 60% точности.

Комбинация различных архитектурных подходов позволила провести сравнительный анализ и достичь максимальной эффективности в решении задачи классификации текстовой токсичности.

4. Результаты и анализ

Сравнение метрик точности показало, что двунаправленная LSTM достигла 90.32% на английских данных и 91.00% на русских данных. GRU показала схожие результаты - 90.29% и 90.76% соответственно. CNN архитектура демонстрирует несколько более низкие показатели - 89.39% и 89.17%.

Анализ матриц ошибок выявил, что наибольшая путаница возникает между классами нейтральных комментариев и токсичных без угрозы. Класс токсичных комментариев с угрозами классифицируется наиболее точно благодаря своей специфичности. Основные ошибки связаны с текстами, имеющими двусмысленную эмоциональную окраску.

5. Выводы

Рекуррентные архитектуры LSTM и GRU показали существенно лучшие результаты по сравнению со сверточными сетями, что подтверждает критическую важность учета контекстной информации в задачах классификации токсичности текста.

Применение комбинированных подходов, включая сочетание различных типов пулинга в CNN и использование всех скрытых состояний в RNN, значительно улучшило качество классификации. Двунаправленность архитектур доказала свою необходимость для полноценного понимания контекста в текстовых данных.

Высокое качество исходных данных с минимальным процентом зашумленных сегментов положительно сказалось на результатах обучения. Проведенная балансировка классов позволила избежать смещения моделей в сторону большинства.

Наилучшие результаты показала архитектура двунаправленной LSTM, достигшая точности свыше 90% на обоих языках. При этом GRU продемонстрировала сопоставимые результаты при меньшей вычислительной сложности, что может быть важно для промышленного внедрения. Полученные результаты подтверждают эффективность выбранных методов для решения задачи классификации текстовой токсичности.