МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ОСОБЕННОСТЬ ПРИМЕНЕНИЯ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТЕКСТОВ НА ЭКОНОМИЧЕСКУЮ ТЕМАТИКУ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Сомика Артема Васильевича

Научный руководитель
к. э. н., доцент _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2025

ВВЕДЕНИЕ

Традиционные методы ручной обработки текстов, основанные на экспертной оценке, не справляются с масштабом и скоростью поступления новых данных. В связи с этим, применение моделей глубокого обучения, становится важным инструментом для автоматизированного анализа тональности, классификации тем и выявления скрытых закономерностей в экономических текстах. Однако существующие решения требуют значительной адаптации к особенностям русского языка и экономического контекста.

Традиционная задача анализа тональности текстов требует специальных методов для русскоязычного контекста. Специфика русского языка, включающая сложную морфологию, свободный порядок слов и высокую контекстную зависимость, создает существенные трудности для стандартных NLP-методов (Natural Language Processing). Экономический дискурс на русском языке имеет ряд особенностей: обилие профессиональных терминов, специфические сокращения и особые синтаксические конструкции, характерные для деловой коммуникации.

В современную эпоху цифровой экономики объем текстовых данных, связанных с финансовыми рынками, корпоративной аналитикой, новостными потоками и социальными медиа. Экономические субъекты сталкиваются с необходимостью оперативного анализа огромных массивов неструктурированной текстовой информации для принятия обоснованных решений.

Существующие в настоящее время модели, обученные на англоязычных корпусах, демонстрируют более низкую эффективность при работе с русскоязычными экономическими текстами. Это связано не только с лингвистическими различиями, но и с особенностями подачи информации в русскоязычных средствах массовой информации, где часто присутствует скрытая оценка и имплицитные смыслы.

Целью бакалаврской работы является усовершенствование процессов анализа тональности экономических текстов средствами машинного обучения за счет применения модели глубокого обучения.

Поставленная цель определила следующие задачи:

- сравнительный анализ современных методов Sentiment Analysis;
- формирование специализированного корпуса экономических текстов;
- реализация модели глубокого обучения;

– разработка приложение для анализа тональности экономических текстов.

Основная практическая значимость работы обусловлена актуальной проблемой использования устаревших языковых моделей для анализа современных экономических текстов. В частности, модель RuBERT, выпущенная в 2019 году, изначально обучалась на данных, которые не отражают современную языковую реальность, особенно в быстроразвивающихся областях, таких как финансы и экономика. В работе использованы современные датасеты 2022—2023 г.г., отражающие текущие языковые тенденции. Особую практическую ценность представляет разработанное программное приложение, которое реализует предложенный подход и предоставляет удобный интерфейс для работы с различными моделями глубокого обучения. Приложение позволяет проводить их дополнительное обучение на актуальных данных.

Бакалаврская работа состоит из введения, трех разделов, заключения, списка использованных источников и двух приложений. Общий объем работы — 63 страницы, из них 40 страниц – основное содержание, включая 4 рисунка и 3 таблиц, цифровой носитель в качестве приложения, список использованных источников информации — 22 наименования.

Первый раздел «Особенности подходов к анализу тональности» представляет комплексное исследование методов анализа тональности экономических текстов, рассматривающее как специфические проблемы данной предметной области, так и современные подходы к их решению. Экономические тексты (финансовые отчеты, деловые новости, аналитические обзоры) характеризуются рядом особенностей, существенно осложняющих автоматический анализ их эмоциональной окраски. Основная сложность заключается в формальном и сдержанном стиле изложения, где тональность передается преимущественно косвенно - через контекст и сочетание профессиональных терминов, в отличие от явного выражения эмоций в бытовой коммуникации или художественных текстах.

Для эффективной работы с такими текстами требуется использование продвинутых методов анализа, способных учитывать не только лексические особенности, но и контекстуальные, синтаксические и семантические связи. Именно такую задачу решает Text Mining — междисциплинарная область, объединяющая подходы из обработки естественного языка, машинного обучения и анализа данных. Text Mining представляет собой междисциплинарную область,

находящуюся на пересечении обработки естественного языка, машинного обучения и анализа данных. Основной задачей Text Mining является извлечение значимой информации и скрытых закономерностей из неструктурированных текстов. Поскольку текстовая информация представлена в виде символьной последовательности, она требует применения специализированных методов обработки, направленных на интерпретацию лингвистических структур и преобразование текстов в формат, пригодный для последующего анализа.

Анализ тональности (Sentiment Analysis, SA) представляет собой одну из наиболее активно развивающихся задач в области интеллектуального анализа текстов. Его основная цель заключается в выявлении эмоциональной окраски высказываний, что позволяет отнести их к определённой категории, например, положительной, нейтральной или отрицательной. Ключевые проблемы анализа тональности экономических текстов включают: насыщенность специализированной терминологией, отсутствующей в стандартных словарях тональности; контекстуальную зависимость эмоциональной нагрузки терминов; сложные синтаксические конструкции, характерные для экономического дискурса [1]. При работе с русскоязычными экономическими текстами возникают специфические сложности. Эти трудности связаны не только с особенностями морфологии русского языка, но и со спецификой финансовой терминологии, что требует особых подходов к обработке текстов. Эти факторы существенно ограничивают эффективность традиционных лексиконных методов, основанных на словарных сопоставлениях [2].

Современные подходы к решению задачи можно классифицировать на три группы. Лексиконные методы, например, с использованием Loughran-McDonald Financial Sentiment Dictionary, несмотря на простоту реализации, демонстрируют ограниченную точность из-за неспособности учитывать контекст. Методы машинного обучения показывают лучшие результаты, но требуют тщательного проектирования признаков и значительных объемов размеченных данных. Подходы на основе глубокого обучения, особенно трансформерные архитектуры (BERT, FinBERT, RuBERT), достигают высоких результов благодаря способности анализировать контекст и семантические связи [3].

Для объективной оценки качества моделей анализа тональности экономических текстов используются ключевые метрики: Accuracy, Precision, Recall и F1-мера. Особое значение имеет анализ матрицы ошибок, которая наглядно

показывает распределение верных и ошибочных классификаций по категориям тональности. В контексте экономических текстов наиболее критичны показатели precision, отражающие способность модели точно идентифицировать нюансы профессиональной лексики, и recall, показывающий полноту выявления значимых сигналов в формальных текстах [4].

Таким образом, проведенный анализ свидетельствует, что для эффективного анализа тональности экономических текстов необходимо применение современных моделей глубокого обучения с адаптацией к специфике профессиональной терминологии и особенностям русскоязычного экономического дискурса, при этом сохраняя баланс между точностью и интерпретируемостью результатов.

Во втором разде «Инструментальные средства для построения модели оценки тональности» представлен детальный анализ архитектурных особенностеймоделей трансформеров, которые стали стандартом в обработке естественного языка. Основное внимание уделено механизмам работы и сравнительным характеристикам ключевых моделей семейства BERT. Архитектура BERT основана на 12 слоях трансформеров с 12 головами внимания и скрытым размером 768, что обеспечивает глубокое понимание контекстных зависимостей [5]. Модель RoBERTa, являясь оптимизированной версией BERT. Она использует те же 12 слоёв, 12 голов и скрытый размер 768. Главное отличие в стратегии предобучения: RoBERTa обучается на значительно большем объёме данных, использует динамическое маскирование, то есть маски создаются при каждой эпохе заново, и исключает задачу NSP (Next Sentence Prediction). Для практического применения особый интерес представляет DistilBERT — облегченная 6-слойная версия, сохраняющая около 95% точности оригинальной модели при значительном сокращении вычислительных ресурсов [6]. В контексте экономических текстов наиболее эффективными оказываются специализированные модели: FinBERT, которая сотсоит из 12 трансформерных слоёв, 12 голов внимания и скрытый размер 768, и RuBERT, которая по архитектуре также повторяет BERT-base: 12 слоёв, 12 голов внимания и 768-мерные скрытые векторы. Отличие в том, что она была предобучена на русскоязычных корпусах [7].

Процесс дообучения (fine-tuning) трансформерных моделей требует тщательного подхода к настройке гиперпараметров. Оптимальные значения скорости обучения (learning rate) находятся в диапазоне $2E-5$ до $5E-5$, что обеспе-

чивает баланс между скоростью сходимости и стабильностью обучения. Размер пакета (batch size) рекомендуется устанавливать в 16 или 32 примера, в зависимости от доступных вычислительных ресурсов. Критически важным является контроль переобучения через механизмы ранней остановки (early stopping) и регуляризации [8].

Практическая реализация моделей осуществляется на языке Python 3.8.0 с использованием современных библиотек. Фреймворк PyTorch обеспечивает гибкость при работе с нейронными сетями, а библиотека Transformers от Hugging Face предоставляет доступ к предобученным моделям. Для обработки данных применяются Pandas и NumPy, визуализация результатов строится на основе Matplotlib. Особое внимание уделяется подготовке данных: очистке текстов, обработке финансовой терминологии, переводу и нормализации [9].

Проведенный анализ позволяет утверждать, что современные трансформерные архитектуры представляют собой мощный инструмент для анализа тональности экономических текстов, однако их эффективность существенно зависит от трех ключевых факторов: правильного выбора базовой модели с учетом языковой специфики; качественной подготовки и предобработки данных, включая очистку и нормализацию текстов, обработку доменно-специфичной терминологии, обеспечение сбалансированности обучающей выборки; грамотной настройки процесса дообучения.

Третий раздел «Применение приложения для анализа тональности» посвящен практической реализации и всестороннему тестированию программного приложения для анализа тональности экономических текстов. Основное внимание уделено самостоятельной разработке автором комплексного решения, включающего как графический интерфейс, так и алгоритмическую часть.

Разработанное приложение представляет собой законченный программный продукт, реализованный на языке Python с использованием библиотеки PyQt5 для создания интуитивно понятного графического интерфейса. Особенностью разработки стало тщательное проектирование архитектуры приложения, обеспечивающее четкое разделение между визуальным представлением (реализованным в Qt Designer) и бизнес-логикой (написаной на Python). Приложение поддерживает полный цикл анализа тональности — от загрузки и предварительной обработки данных до обучения моделей и визуализации результатов.

Функциональность приложения включает следующие возможности:

– загрузка данных (поддержка формата csv для обучающих и тестовых данных, функция предварительного просмотра обучающих данных);
– работа с алгоритмами (выбор из нескольких моделей глубокого обучения (RuBERT, FinBERT, RoBERT, DistilBERT, BERT));
– настройка гиперпараметров (ручная настройка гиперпараметров, использование стандартных гиперпараметров);
– процесс обучения (запуск обучения с визуализацией прогресса);
– анализ результатов (отображение метрик моделей (Accuracy, Precision, Recall, F1-score), визуализация графиков обучения);
– экспорт (сохранение обученных модей и метрик).

Особое внимание было уделено обеспечению удобства работы для пользователей без специальной технической подготовки. Интерфейс приложения организован в виде четырех интуитивно понятных вкладок, последовательно ведущих пользователя через все этапы анализа.

Для апробации приложения была сформирована тестовая выборка из 300 экономических новостей, собранных из средств массовой информации и прошедших ручную разметку.

Для объективности оценки использовались основные метрики оценки:Accuracy, Precision, Recall и F1-score, что обеспечивало всестороннюю оценку качества классификации. В таблице 1 видно, что наибольшую точность продемонстрировала модель RuBERT. Сравнительный анализ, представленный в таблице 2 с тестовыми значениями показал 94% согласованности предсказаний модели с ручной разметкой, что подтверждает высокую эффективность разработанного решения.

Таблица 1 – Сравнение качества моделей на тестовой выборке

Модель	**Accuracy**	**Precision**	**Recall**	**F1-score**
RuBERT	0.91	0.90	0.92	0.91
FinBERT	0.88	0.87	0.89	0.88
RoBERTa	0.86	0.85	0.87	0.86
BERT	0.85	0.84	0.85	0.84
DistilBERT	0.82	0.81	0.83	0.82

ЗАКЛЮЧЕНИЕ

Анализ тональности русскоязычных экономических текстов сталкивается с рядом серьёзных вызовов, обусловленных как особенностями языка, так и спе-

Таблица 2 – Сравнение предсказаний RuBERT и размеченной тестовой выборки

Критерий	**Позитивные**	**Негативные**	**Всего**
Экспертная разметка	166	134	300
Предсказания RuBERT	168	132	300
Совпадения	158	124	282
Согласие	95.2%	92.5%	94.0%

цификой экономического дискурса. Сложная морфология и свободный порядок слов в русском языке затрудняют автоматическую обработку, а контекстная зависимость значений требует глубокого семантического анализа. Экономические тексты отличаются формальным стилем, обилием профессиональной терминологии и отсутствием явных эмоциональных маркеров, из-за чего тональность часто выражается косвенно. Кроме того, одни и те же слова могут менять эмоциональную окраску в зависимости от контекста. Синтаксическая сложность, включающая длинные предложения с вложенными конструкциями, дополнительно усложняет анализ. Традиционные методы, такие как лексиконные подходы или модели, обученные на англоязычных данных, оказываются малоэффективными. Это создаёт необходимость в разработке специализированных решений на основе современных методов глубокого обучения, способных учитывать языковые и тематические особенности русскоязычных экономических текстов.

Был проведён сравнительный анализ современных методов Sentiment Analysis с акцентом на модели глубокого обучения, адаптированные для русскоязычных экономических текстов.

Для дообучения моделей был сформирован специализированный корпус экономических текстов, включающий 6287 единиц данных. Обучающая выборка объединила открытый датасет финансовых новостей (1840 позитивных и 1888 негативных записей) и 3208 текстов сгенерированных с помощью модели ChatGPT. Для объективной оценки качества работы моделей была создана тестовая выборка, состоящая из 300 экономических новостей (166 позитивных и 134 негативных примера), собранных из средств массовой информации.

Были дообучены модели глубокого обучения RuBERT, FinBERT, RoBERTa, BERT и DistilBERT, проведено их тестирование на подготовленной тестовой выборке.

В результате сравнительного анализа моделей с помощью метрик качества, таких как Accuracy, Precision, Recall и F1-score, было установлено, что наилучшие показатели демонстрирует модель RuBERT с точностью 91%.

Для практической реализации анализа тональности было разработано приложение, позволяющее загружать обучающий и тестовый наборы в формате csv с возможностью предварительного просмотра данных. Пользователям доступен выбор из нескольких предобученных моделей глубокого обучения (RuBERT, FinBERT, RoBERT, DistilBERT, BERT) с возможностью как ручной настройки гиперпараметров, так и использования стандартных конфигураций. Функционал приложения включает запуск процесса обучения , а также комплексный анализ результатов через ключевые метрики качества (Accuracy, Precision, Recall, F1-score). Пользователь может экспортировать обученные модели и полученные метрики.

Результаты бакалаврской работы были апробированы на XV Международной научной конференции студентов и аспирантов «Экономика и управление: проблемы, тенденции, перспективы», которая проходила в Саратовском национальном исследовательском государственном университете имени Н.Г. Чернышевского, экономический факультет, с 3 апреля по 5 апреля 2025 года.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Araci, D. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models [Электронный ресурс]. — 2019. — arXiv:1908.10063. —- URL: https://arxiv.org/abs/1908.10063 (дата обращения: 15.04.2025)

2 Smirnov, I. V., Moloshnikov, I. A. Financial Text Processing in Russian // Proceedings of NAACL. — 2021. — P. 234—245. — Яз. англ.

3 Dmitriev, A. S. Advanced Sentiment Analysis for Russian Financial Texts // Computational Linguistics. — 2021. — Vol. 47(2). — P. 345—362. — Яз. англ.

4 Sboev, A. G., Rybka, R. B. Russian Sentiment Analysis Datasets // Proceedings of LREC. — 2020. — P. 6789—6795. — Яз. англ.

5 Huang, A. H., Wang, H., Yang, Y. FinBERT: A Pretrained Language Model for Financial Communications // Proceedings of COLING. — 2020. — P. 4703-4714. — Яз. англ.

6 Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of

NAACL-HLT. — 2019. — P. 4171—4186. — Яз. англ.

7 Kuratov, Yu. M., Arkhipov, M. Yu. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies. — 2019. — Vol. 18. — P. 333—339. — Яз. англ.

8 Zhang, L., Wang, H., Chen, X. Deep Learning for Sentiment Analysis // AI Review. — 2022. — Vol. 55(3). — P. 1—25. — Яз. англ.

9 Python Software Foundation. Python 3.8.0 Documentation [Электронный ресурс]. — 2019. — URL: https://www.python.org/downloads/release/python-380/ (дата обращения: 10.04.2025)