Анализ тональности новостей о международной торговле в условиях санкций: подходы NLP

УДК:339.5; ББК:65.428; JEL:F10
DOI: 10.24412/2072-8042-2025-2-77-93

София Алексеевна ОСОКИНА,
Всероссийская академия внешней торговли (117312, Москва, улица Вавилова, 7) - аналитик Центра анализа данных, E-mail: s.osokina@vavt.ru

Виктория Леонидовна АБРАМОВА,
Всероссийская академия внешней торговли (117312, Москва, улица Вавилова, 7) - аналитик Центра анализа данных, E-mail: v.l.abramova@vavt.ru

Дарья Андреевна ЛЮТОВА,
Всероссийская академия внешней торговли (117312, Москва, улица Вавилова, 7) - программист Центра анализа данных, E-mail: d.lyutova@vavt.ru

Аннотация

Статья посвящена исследованию особенностей обработки естественного языка (NLP) в новостных сообщениях о торговых санкциях. Основное внимание уделяется выявлению лексических и структурных характеристик текстов, которые могут повлиять на качество автоматизированного анализа. В работе подчеркивается важность учета контекста и культурных различий при оценке тональности новостей, а также обсуждаются сложности, связанные с интерпретацией экономического и политического контента. Представлен обзор современных методов анализа сентимента, включая подходы на основе машинного обучения и нейронных сетей. Рассматриваются практические аспекты применения этих методов в анализе новостей о санкциях, с учетом их специфики и неоднозначности.

Ключевые слова: обработка естественного языка, NLP, машинная лингвистика, искусственный интеллект, санкции, анализ тональности новостных текстов, торговые новости, международная торговля.

Analyzing the Sentiment of international Trade News in the Context of Sanctions: NLP Approaches

Sofia Alexeevna OSOKINA,
Russian Foreign Trade Academy (117312, Moscow, Vavilova, 7)
Analyst, Center for Data Analysis, E-mail: s.osokina@vavt.ru

Victoria Leonidovna ABRAMOVA,
Russian Foreign Trade Academy (117312, Moscow, Vavilova, 7)
Analyst, Center for Data Analysis, E-mail: v.l.abramova@vavt.ru

Daria Andreevna LYUTOVA,
Russian Foreign Trade Academy (117312, Moscow, Vavilova, 7)
Programmer, Center for Data Analysis, E-mail: d.lyutova@vavt.ru

Abstract

The article focuses on exploring the characteristics of natural language processing (NLP) in trade sanctions-related news. Emphasis is placed on identifying lexical and structural features of texts that can affect the quality of automated analysis. The importance of considering context and cultural differences when evaluating the tone of news is highlighted, along with discussing challenges associated with interpreting economic and political content. An overview of contemporary sentiment analysis methods, including approaches based on machine learning and neural networks, is presented. Practical aspects of applying these methods to analyze sanction-related news, taking into account their specificities and ambiguities, are also discussed.

Keywords: Natural Language Processing, NLP, machine linguistics, artificial intelligence, sanctions, sentiment analysis, trade news, international trade.

ВВЕДЕНИЕ

В условиях усиления глобальной экономической турбулентности и роста числа санкционных ограничений, особую значимость приобретает своевременная обработка и анализ новостей внешнеэкономической тематики. Введение санкций оказывает существенное влияние на международные торгово-экономические связи и кардинально меняет ситуацию на затронутых ограничениями рынках. Оперативный и качественный анализ новостных текстов позволяет не только прогнозировать последствия введенных мер для различных секторов экономики, но и формировать стратегии адаптации для минимизации возможных рисков.

Исследования новостных сообщений остаются актуальным академическим полем в социальных науках с первой половины XX века. Такое продолжительное внимание к новостным текстам связано с центральной ролью, которую средства массовой информации играют в трансляции идей и знаний. Социолог Н. Луман отмечает, что “то, что мы знаем о нашем обществе и даже о мире, в котором живем, мы знаем благодаря массмедиа”[1]. В экономике активная работа с медиа-контентом началась несколько позже – на рубеже XX и XXI веков – так как информация, распространяемая новостными ресурсами долгое время не входила в сферу коли-

чественного анализа. Однако на сегодняшний день индикаторы на основе СМИ используются для прогнозирования инфляции, бизнес-циклов и развития финансовых рынков[2].

Существенная часть экономистов обращается к новостным ресурсам, чтобы устранить временные пробелы, связанные с задержкой в получении данных. Сведения о важных событиях появляются в СМИ намного раньше, чем в опросах или твердых экономических данных[3]. Анализ сообщений СМИ становится еще более значимым, если мы соглашаемся с теоретической установкой о том, что медиадискурсы не только отражают, но и формулируют экономическую повестку, прямо и косвенно влияют на внешнеторговую ситуацию[4], направляя общественное мнение. В работе "Нарративная экономика"[5] Р. Шиллер указывает на то, что популяризированные медиа идеи о состоянии рынков в определенный момент времени могут задавать контекст для восприятия будущих мер экономической политики, оказывать эффект на деятельность инвесторов и бизнеса. Конкретные примеры влияния тональности пресс-релизов на стоимость акций и поведение инвесторов рассматриваются в статье "Придание содержания настроениям инвесторов: Роль СМИ на фондовом рынке"[6]. Автор приходит к выводу о том, что высокие значения пессимизма СМИ оказывают понижательное давление на рыночные цены, особенно компаний с небольшим объемом выпущенных акций.

Комплексный анализ публикаций российских и зарубежных СМИ предполагает работу с большими объемами данных. При работе с таким динамичным, ежеминутно обновляющимся источником, как новости внешнеэкономической тематики, принципиальное значение имеет оперативный сбор, классификация и исследование текстов. В связи с этим, перспективным направлением в области анализа медиапространства является обработка естественного языка (NLP) методами машинной лингвистики, применение которых позволяет извлекать значимую информацию из неструктурированных текстовых данных. Автоматическое извлечение и классификация ключевых тематических сущностей (стран, международных организаций, товаров и др.) позволяет быстро определять основные тенденции на различных внешнеэкономических рынках, а выявление тональности или эмоциональной окраски текстов торговых новостей помогает оценить общее настроение рынка и отношение различных участников экономической деятельности к тому или иному событию, прогнозировать потенциальное влияние новостей на поведение экономических акторов.

Учитывая актуальность и значимость сбора и обработки публикаций СМИ, целью данного исследования является выявление лексических и структурных особенностей языка торговых новостей, которые могут влиять на автоматизированный анализ текстов такого типа. Особое внимание будет уделено новостным публикациям, связанным с санкционными ограничениями и их последствиями. Работа направлена на повышение точности и глубины анализа текстовых данных в сфере внешней торговли.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ NLP

Задача автоматической обработки экономических текстов является частью более широкого направления в машинном обучении – обработки естественного языка (NLP), – решающего задачу по обеспечению понимания и генерации человеческой речи машиной. Эта теоретическая и практическая область получила значительное развитие в последнее десятилетие в связи с началом использования моделей глубокого обучения нейронной сети, в особенности на основе архитектуры трансформеров, представленной исследователями Google в работе "Attention is All You Need"[7]. Специфика данного подхода, ставшего основой для множества современных моделей, таких как BERT и GPT, заключается в использовании механизма внимания, позволяющего обрабатывать входные текстуальные данные не по одному элементу за раз, а одновременно, учитывая их взаимные связи. Это позволяет лучше анализировать контекст и зависимость между словами в предложении, вне зависимости от их удаленности друг от друга.

Несмотря на существенный прогресс в обработке естественного языка, некоторые аспекты понимания человеческой речи машиной остаются проблемными зонами. Это связано не только с объемом и качеством доступных для обучения данных, но и с особенностями функционирования естественных языков. Ф. де Соссюр[8] в начале XX века провел разграничение между понятиями "язык" и "речь", утверждая, что последняя является индивидуальным актом использования языка – системы норм и знаков, разделяемой членами некоторого языкового сообщества. Обе составляющие акта человеческого высказывания связаны с трудностями для NLP. В первую очередь, сложность вызывает множественность языков, каждый из которых обладает разнообразными синтаксическими и морфологическими структурами, что создает препятствие для разработки моделей, способных обобщаться на разные языки[9]. Некоторые из них представлены значительно более низким объемом доступных для обучения данных, что создает неравенство в результатах работы моделей с англоязычными текстами и текстами на "низкоресурсных" языках (например африканских)[10].

Анализ человеческой речи, в свою очередь, сопровождается такими трудностями, как смысловая неоднозначность слов и словосочетаний, неологизмы, метафоры, анафоры, омонимы, пропуск слов или фраз, которые подразумеваются из контекста, ирония и сарказм, наличие опечаток/ошибок. А. Талмор[11] отмечает, что даже передовые модели (GPT и BERT) не очень хорошо справляются с такими задачами, как распознавание скрытого смысла или сарказма. Модели не имеют "здравого смысла" и структурированного понимания мира, что приводит к поверхностному анализу текста. Механизм внимания позволяет моделям понимать контекст на уровне предложений и абзацев отдельного текста, но не на уровне дискурса. Решение данной проблемы оказывается особенно актуальным для обработки новостных текстов о торговле, так как содержащаяся в них информация отсылает читателей к множеству экономических и политических контекстов.

Общая специфика текстов новостей о внешней торговле состоит в ориентации на модель возможного читателя[12], обладающего общими и специально экономическими знаниями, достаточными для того, чтобы восстановить контекст описываемого события и верно интерпретировать воспринимаемые языковые выражения. Ориентация на определенный тип читателя становится особенно заметна при сравнительном анализе публикаций СМИ разных стран. Так как тексты создаются из перспективы конкретного государства и пишутся для локальной аудитории, в них часто бывают опущены определения, раскрывающие информацию, очевидную для данной аудитории. Например, может быть не указана страна происхождения местных компаний, сокращены должности официальных лиц: "Новый министр торговли Пичаи Нариптхапхан пообещал принять срочные меры по решению проблемы стоимости жизни, ускорить переговоры о заключении соглашения о свободной торговле (ССТ) и ужесточить контроль за наплывом китайского импорта."[13]. При автоматической обработке текстов это может приводить к ошибкам при выявлении именованных сущностей.

Кроме того, новости о торговле часто содержат специализированную экономическую и юридическую лексику, связанную с международными торговыми соглашениями, санкциями, тарифами, инвестиционной активностью, а также большое количество имен собственных (персоналии, названия компаний, НКО, фондов, и др.). Это требует точного распознавания терминологии и ее контекстуального значения, так как оно может варьироваться в источниках разных стран. Помимо этого, торговые новости часто имеют множество ссылок на связанные с описываемым событием исторические факты и данные, которые не всегда автоматически извлекаются из текста и требуют дополнительных усилий для анализа.

Еще одной проблемой является использование тропов, речевых фигур и образного, эмоционального языка, которое порождает двусмысленности, усложняющие анализ текста. Так, в экономической новости с заголовком "Trade-in campaign bearing fruit"[14] (трейд-ин кампания приносит свои плоды) "fruit" следует понимать не как фрукт, потенциальный товар, а как часть устойчивого выражения "приносить плоды".

Говоря о структурных особенностях текстов торговых новостей, следует выделить несколько типов их организации: формульный и свободный. В первом случае все новостные публикации издания строятся по одной схеме, которая чаще всего включает в себя вводный абзац, в котором суммируется вся основная информация из текста, основную часть и финальный абзац, в котором для контекста приводятся исторические данные или общие сведения об упомянутых в новости персоналиях и организациях. Во втором случае, текст имеет произвольную авторскую структуру, что создает дополнительные трудности для задачи саммаризации, так как модели приходится самой выделять и собирать из разных частей текста наиболее значимую информацию.

Введение контекста, с одной стороны, упрощает интерпретацию текста, а с другой стороны, создает дополнительную неоднозначность, в особенности при анализе сентимента, т.е. общего эмоционального тона статьи по отношению к описываемым экономическим событиям. Даже если большая часть новостного сообщения о торговле между странами содержит информацию о положительном развитии экономических отношений и позитивные статистические данные, новость может быть ошибочно оценена как негативная, так как в ней могут присутствовать отрицательные статистические данные за предыдущие периоды или несколько противоположных мнений политиков и экспертов, по-разному оценивающих ситуацию.

АНАЛИЗ ТОНАЛЬНОСТИ: ПОДХОДЫ И СЛОЖНОСТИ

Анализ сентимента является одной из ключевых задач машинной лингвистики, направленных на определение того, содержит ли рассматриваемый текст негативные, позитивные или нейтральные эмоции. Оценка тональности не ограничивается этими тремя базовыми категориями и может быть кастомизирована под конкретную задачу. В коммерческой сфере методология чаще всего применяется для анализа отзывов клиентов и обзоров продукции. Она также полезна для отслеживания изменений общественного мнения в любых других сферах – от политики до экономики. Изначально процесс оценки тональности автоматизировался с помощью NLP и машинного обучения. В последнее время к решению этой задачи были успешно подключены большие языковые модели (LLM), такие как GPT и LLaMA[15].

Важнейшим элементом анализа сентимента считается классификация полярности. Полярность отражает общее эмоциональное состояние, выраженное некоторым текстом, фразой или словом. Она может быть представлена в виде числового диапазона, например, от -100 до 100, где 0 означает нейтральное настроение. Один из самых простых подходов к определению тональности, разработка которого велась с 2000-х гг., использует заранее составленные правила – лексиконы слов и их полярностей (SentiWordNet, SenticNet и др.)[16]. Алгоритм ищет в тексте ключевые слова и сопоставляет их с данными из тезауруса. Такой метод не требует обучения модели, но плохо справляется со сложными языковыми конструкциями и контекстными значениями.

Классические модели ML обучаются под задачу определения тональности на больших массивах данных, чтобы проводить классификацию, основываясь на закономерностях в тексте. Во время обучения модель находит паттерны и признаки, которые помогают различать сентимент. Такие модели работают поэтапно, начиная с преобразования текста в числовой вид и заканчивая классификацией. Некоторые из наиболее популярных алгоритмов для анализа тональности включают байесовский классификатор, логистическую регрессию, метод опорных векторов и случайный лес.

Для повышения точности определения тональности используются аспектно-ориентированные методы (Aspect-Based Sentiment Analysis, ABSA), в рамках которых сентимент текста определяется по отношению к конкретной выделенной сущности (товар, страна, организация и т.п.)[17].

Несмотря на существенный прогресс в решении задачи автоматизированного анализа сентимента, современные языковые модели продолжают сталкиваться с уже упомянутыми сложностями обработки естественного языка: лингвистической многозначностью, амбивалентностью оценок того или иного объекта/события, контекстуальной зависимостью.

Вопрос о степени влияния контекста на тональность остается ключевым для исследований в области обработки естественного языка, особенно при анализе политических и экономических новостей, таких как новости о санкциях. Современные исследования подтверждают, что контекст способен радикально изменить восприятие тональности текста. Политические и экономические факты, содержащие негативную или позитивную окраску, могут восприниматься иначе в зависимости от того, какая информация их окружает. Например, новости о росте инфляции воспринимаются менее негативно, если в статье приводятся ссылки на позитивные ожидания аналитиков на следующие кварталы. Это подчеркивает, что контекст может смягчать или усиливать тональность, влияя на общее восприятие новости.

Восприятие политических событий происходит, как правило, через призму культурного и исторического контекста. То есть освещение международных событий, таких как санкции, часто интерпретируется по-разному в зависимости от страны, в которой были опубликованы новости. Так, публикации о введении санкций против определенных государств воспринимаются их партнерами более негативно, тогда как в других странах они могут читаться нейтрально или даже позитивно. Этот эффект свидетельствует о необходимости учета культурного контекста в анализе тональности новостей.

Особую важность контекст приобретает в политических новостях, где одно и то же событие может трактоваться диаметрально противоположно в зависимости от политической позиции СМИ. В работе “Что важнее, контекст или сентимент?: Анализ влияния новостей на выборы в США с использованием обработки естественного языка”[18] авторы ставят перед собой задачу изучить воздействие политических новостей на общественное мнение и предпочтения избирателей в США во время президентских выборов 2016 года. Центральный вопрос исследования – оценить, какое влияние оказывает тональность новостей по сравнению с их контекстом и тематикой. Для анализа применены методы обработки естественного языка (NLP), позволяющие измерить, как конкретные темы и эмоциональная окраска сообщений воздействуют на соотношение рейтингов кандидатов Хиллари Клинтон и Дональда Трампа. Внимание уделяется не только прямой корреляции, но и причинно-следственным связям, что является ценным вкладом в анализ политического влияния СМИ.

Исследование опирается на новости из двух крупных американских СМИ – *The New York Times* и *Fox News*, которые имеют разные политические позиции. В ходе исследования были выделены ключевые темы, связанные с вопросами экономики, внешней политики, миграции и внутренних политических скандалов, которые получали освещение в обоих СМИ.

Эти темы широко обсуждались в новостях и формировали повестку, что, как предполагают авторы, влияло на изменение общественного мнения. Анализ показывает, что у каждого издания наблюдаются свои акценты: *The New York Times* чаще освещала экономику и внешнюю политику, в то время как *Fox News* акцентировалась на скандалах вокруг Клинтон и иммиграционной политике.

Одним из первых этапов исследования было выявление корреляции между количеством упоминаний кандидатов и разрывом в их рейтингах в опросах. Увеличение упоминаний Клинтон в *The New York Times* коррелировало с увеличением её рейтинга относительно Трампа, а аналогичные упоминания в *Fox News* давали обратный результат, что можно объяснить политическими предпочтениями изданий. Авторы приходят к выводу, что простая частота упоминаний кандидата положительно коррелирует с его рейтингом в изданиях, которые оказывают ему поддержку, и наоборот.

Чтобы углубить анализ, исследователи выделили тональность упоминаний кандидатов по дням. С помощью алгоритмов на основе рекурсивных моделей они оценивали количество позитивных, негативных и нейтральных фраз. Интересным результатом стало то, что для *The New York Times* средний коэффициент тональности был всегда положительным для Клинтон, тогда как для *Fox News* этот коэффициент варьировался, и часто был отрицательным в отношении Клинтон.

Однако тональность сама по себе не смогла объяснить изменения в общественном мнении: она была важна, но не была определяющей. Это привело к выводу о значимости контекста – или тем, в рамках которых упоминались кандидаты, – в изменении общественного мнения.

Подходы к анализу тональности, такие как модели Word2Vec, BERT и трансформеры, за последние годы значительно усовершенствовались, научившись не только идентифицировать эмоциональную окраску текста, но и учитывать контекст, в котором она возникает. Если раньше тональность текста определялась преимущественно на основе присутствия отдельных слов и фраз с позитивной или негативной окраской, что зачастую приводило к поверхностному анализу, то современные модели способны «понимать» значение слов в контексте более широкого смыслового поля текста. Этот подход позволяет моделям учитывать не только прямую эмоциональную окраску слов, но и нюансы, возникающие в зависимости от их окружения и значений в различных сочетаниях. BERT и другие трансформерные модели позволяют оценивать взаимоотношения слов в рамках всего предложения или даже текста целиком, что способствует более глубокой интерпретации.

Кроме того, современные модели способны учитывать культурные и социальные аспекты, которые также влияют на тональность. Модели вроде BERT, обученные на мультикультурных и многоязычных данных, способны адаптироваться к контекстным различиям, обнаруживая закономерности в восприятии схожих событий в разных странах или социальных группах.

Таким образом, сочетание анализа тональности и контекста позволяет моделям достигать более точных и интерпретируемых результатов, понимая тональность не как фиксированное свойство текста, а как динамическое, изменяющееся в зависимости от окружения и контекста. Это стало возможным благодаря архитектуре трансформеров, которые могут сохранять и передавать контекстную информацию на протяжении всего текста, делая анализ тональности многоуровневым и гибким. Такой подход позволяет учитывать тонкие грани эмоциональной окраски и создавать более точные модели восприятия информации, особенно в случаях, когда нейтральные или позитивные слова могут приобретать противоположные значения в зависимости от культурного, социального или исторического фона текста.

Авторы статьи "Расширенный NLP-анализ трансграничных настроений в СМИ Китая и Южной Кореи"[19] используют языковую модель BERT, специально адаптированную и усовершенствованную для анализа текстов экономической тематики. З. Лю и коллеги[20] дообучили BERT на широком корпусе финансовых данных, включая корпоративные документы, аналитические отчеты, новости и др. Полученная в результате этого процесса модель FinBERT показала лучшие результаты в понимании языковых особенностей, терминологии и контекста финансовых документов. Одним из главных применений модели стал анализ тональности в финансовой сфере, необходимый для оценки рыночных настроений и прогнозирования движения акций.

Позже FinBERT была дополнительно дообучена на документах Федерального комитета США по открытым рынкам (FOMC) – подразделения Федеральной резервной системы, отвечающего за операции на открытом рынке и ключевые решения по денежно-кредитной политике, такие как установка процентных ставок. Это привело к созданию модели FinBERT-FOMC[21]. В материалах FOMC используется богатый и сложный экономический язык с большим количеством терминологии, связанной с экономической политикой, что отличает их от других финансовых текстов. Хотя модель была обучена специально для анализа тональности документов FOMC, она может быть полезна и для других задач, связанных с выявлением сентимента текстов политико-экономической направленности.

ПОРТАЛ "ТОРГОВЫЕ НОВОСТИ" ЦЕНТРА АНАЛИЗА ДАННЫХ ВАВТ

Одной из исследовательских задач, решаемых Центром анализа данных ВАВТ в настоящий момент, является расширение функционала портала Торговые Новости[22], интеллектуальной базы данных новостей торговой тематики, собранных и проанализированных с применением моделей машинного обучения. На данном

этапе обработка новостей включает выделение в тексте нескольких типов сущностей (стран, международных экономических организаций и товаров, в соответствии с Международной стандартной торговой классификацией), а также классификацию по нескольким типам экономических отношений (Внешняя торговля, Инвестиции, Совместные проекты и программы, Специальные отношения, не классифицированные по типу). Помимо этого, для каждой новости автоматически формируется дайджест, краткий пересказ ее основного содержания. В перспективе Портал будет дополнен информацией о тональности новостей.

В качестве первого эксперимента по определению сентимента новостей о санкциях выбрана многоязычная лингвистическая модель MoritzLaurer/mDeBERTa-v3-base-mnli-xnli[23], работающая по принципу zero-shot обучения, то есть обучения "без примеров", когда ИИ учится выполнять новые задачи, не обучаясь на специфических примерах для этих задач. Модель такого типа была выбрана в связи с отсутствием размеченных данных для обучения и анализа. Она работала с набором предварительно отобранных новостей, содержащих оценки влияния торговых санкций на экономику разных стран, и классифицировала тональность базовыми метками "positive", "neutral", "negative". Результаты автоматизированного анализа оказались не полностью удовлетворяющими, так как, в силу специфики и неоднозначности лексики новостей санкционной тематики, выводы модели нуждались в дальнейшей ручной проверке.

Оценка тональности через стандартные маркеры "позитивная", "негативная", "нейтральная" была затруднена, поскольку восприятие содержания публикаций СМИ субъективно и может быть расценено по-разному по политическим причинам. Стандартные лингвистические модели, предобученные для определения сентимента, делают вывод о настроении текста, обращая внимание на слова, которые, в подавляющем большинстве случаев, имеют положительную или отрицательную окраску. Релевантная для экономических текстов лексика включает такие примеры, как "вырос", "улучшился", "открылся", "расширился" или "упал", "сократился", "расторг", "ухудшился", соответственно. Однако применительно к новостям о санкциях, перечисленные слова-маркеры могут указывать на противоположную тональность.

В связи с этим, с целью выявления оценок влияния торговых ограничений на мировую и российскую экономику, было принято решение определять тональность не через базовые маркеры, а по пяти специально сформулированным классам:

• Санкции вредят России (в тексте новости явно обозначается, что введенные западными странами санкции наносят ущерб российской экономике, внешней торговле, совместным проектам с зарубежными странами);

• Санкции вредят международной торговле (в тексте новости явно обозначается, что введенные западными странами санкции наносят ущерб международной торговле в целом, но не им самим);

• Санкции вредят странам, которые их ввели (в тексте новости явно обозначается, что введенные западными странами санкции наносят ущерб их же экономике, торговле);

• Санкции не вредят России (в тексте новости явно обозначается, что введенные западными странами санкции не оказали никакого эффекта на экономику России, они не наносят ущерб и не приносят пользу);

• Санкции приносят пользу России (в тексте новости явно обозначается, что в результате введения западными странами санкций против России страна оказалась в плюсе. Например: улучшилось внутреннее производство, экспорт увеличился, открылись новые рынки, появились новые партнеры и т.д.);

Для обучения модели на определение перечисленных классов эксперты с экономическим образованием провели разметку новостных текстов, отмечая один или несколько типов тональности, а также части текста, указывающие на эти типы. Ввиду ограниченности человеческих ресурсов, для разметки были поданы предварительно отобранные данные. 5000 публикаций с тегом “Санкции” были взяты из базы данных Торговых Новостей. Параллельно с этим, для каждого из пяти классов сентимента были подобраны ключевые слова. Затем из выборки 5000 новостей были оставлены только те, в которых эти ключевые слова встречаются хотя бы один раз. В результате, для разметки пятью аннотаторами было оставлено около 2000 сообщений, 1200 на русском и 800 на английском языке.

Полученные размеченные данные оказались малопригодны для обучения моделей ввиду высокой степени рассогласованности оценок тональности. В связи с этим, было проведено дополнительное исследование данных, мнения экспертов о тональности которых не совпали.

Таблица 1

Статистика разметки новостей по 5 классам

Всего	***1948***
Есть хотя бы одна метка	1254
Размечено более чем одним аннотатором	694
Одинаково ответили большинство аннотаторов	258
Рассогласованность:	436
в новостях на английском	96
в новостях на русском	340

Источник: составлено авторами.

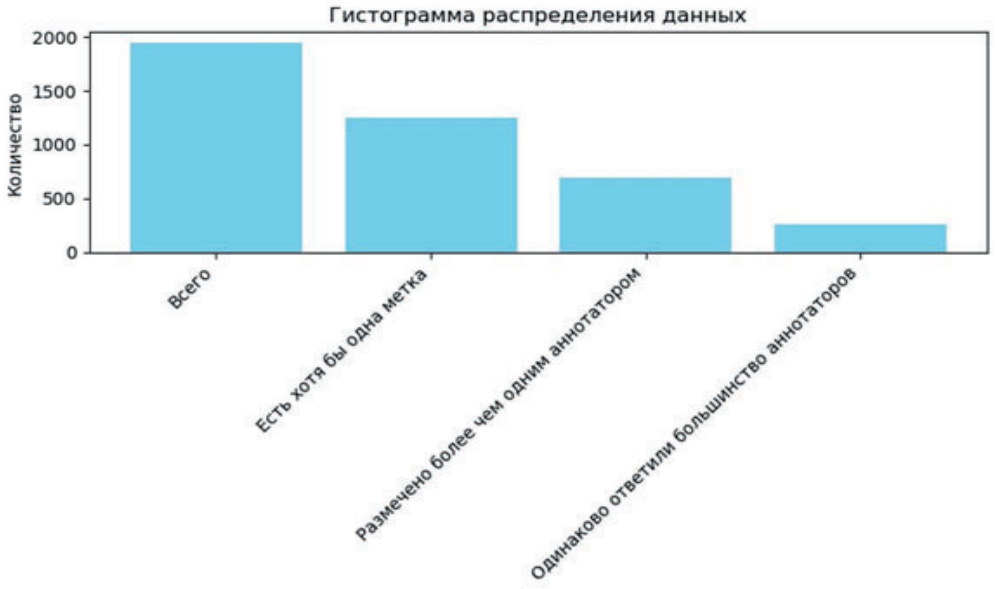

Рис. 1. Гистограмма распределения размеченных данных.
Fig. 1. Histogram of the distribution of the posted data.

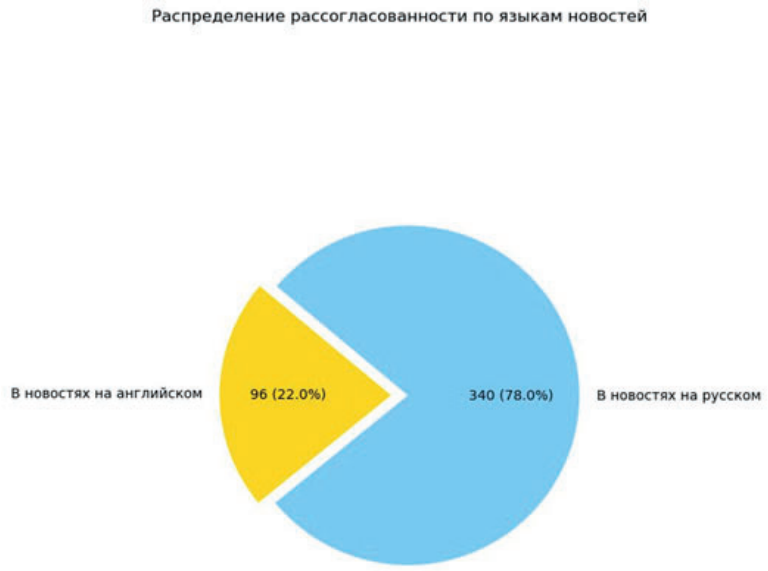

Рис. 2. Распределение рассогласованности разметки по языкам новостей.
Fig. 2. Distribution of markup inconsistency by news languages.

Разница в оценках наблюдалась не только в таких близких по смыслу классах, как ‘не вредят России’ и ‘приносят пользу России’, но и в семантически противоположных ‘не вредят/вредят России’. Тем не менее, новости с типами тональности ‘не вредят России’ и ‘вредят России’ лидируют по количеству согласованной разметки. Все пять аннотаторов отнесли к этим классам новости, в заголовках которых прямо указано, какое влияние санкции оказывают на торговлю России с иностранными партнерами. Например: “Турецкий экспорт в Россию упал на фоне указа США о вторичных санкциях”[24]; “Санкции не препятствуют поставкам российских удобрений, утверждает Байден”[25].

В свою очередь, рассогласованность в большинстве случаев оказалась связана с 1. присутствием нескольких мнений в одной новости; 2. наличием описания не только эффектов санкций, но и целей ограничений (“США и Европа прилагают все усилия, чтобы найти способ ограничить доходы России от продажи нефти (...) Несмотря на санкции, Россия все еще получает доходы, продавая свою нефть по более низким ценам”[26]); 3. различием в понимании вреда и его отсутствия. Так, в процитированной новости получение доходов вопреки санкциям может быть охарактеризовано как отсутствие вреда, но продажа нефти по сниженной стоимости – как ущерб. Еще большую трудность вызвало различение классов ‘не вредят России’ и ‘приносят пользу России’. Эксперты также расходились во мнениях относительно направленности вреда экономических ограничений, особенно сложным оказалось определить границу между ущербом международной торговле и странам-инициаторам санкций и ущербом международной торговле и России.

На данный момент, портал Торговые Новости[27] функционирует как полностью автоматизированная система сбора и обработки новостей экономической и санкционной тематики. Новости ежедневно собираются из более чем двухсот источников десятков стран, в том числе на английском, французском и португальском языках. База данных Торговых Новостей насчитывает более шести миллионов уникальных записей. Пользователи сайта могут получить доступ к интересующей их информации, воспользовавшись фильтрами по стране и региону, товарам, типу экономических отношений и временному периоду.

Расширение функционала портала и добавление возможности определения тональности новостей находится в разработке. Для реализации этой задачи потребуется дальнейшая разметка данных, а также поиск новых технических решений. Так, в 2024 году с целью повышения точности автоматизированного анализа новостей была внедрена большая языковая модель LLaMA 3.1:8b-instruct-q6_K, способная выполнять широкий спектр задач. В будущем, эта модель также может быть использована для выявления сентимента торговых новостей.

Проведенная работа показала, что в случае с новостями о торговых санкциях задача выявления тональности текста сталкивается с множеством неоднозначностей и оказывается сопряжена с анализом множества деталей и контекстов. Обра-

ботка естественного языка представляет собой сложную задачу, требующую учета множества аспектов человеческой речи. Современные лингвистические модели демонстрируют перспективы в обработке новостных текстов, но требуют дополнительного обучения и проверки, особенно в контексте новостей о санкциях. Проведенные исследования показывают необходимость тщательной разметки данных и интеграции экспертных знаний для повышения точности анализа.

Несмотря на существующие сложности, текущие достижения в области обработки естественного языка открывают широкие перспективы для дальнейших исследований и разработок. Внедрение передовых моделей демонстрирует значительный потенциал для улучшения точности анализа и выявления тональности экономических новостей. Работы в этой области продолжают развиваться, и дальнейшие исследования помогут улучшить эффективность автоматизированного анализа.

ПРИМЕЧАНИЯ:

[1] Луман Н. Реальность массмедиа. М., 2005. С. 8.

[2] Например: Lamla, M. J., Lein, S. M., & Sturm, J.-E. Media reporting and business cycles: Empirical evidence based on news data // Empirical Economics. 2020. № 59(3). P. 1085–1105.

[3] Boumans D., Müller H. Sauer S. How media content influences economic expectations: Evidence from a global expert survey // Journal of Forecasting. 2023. URL: https://onlinelibrary.wiley.com/doi/10.1002/for.2961 (дата обращения: 29.08.2024).

[4] Dijk, T. A. van. Power and the news media // Political Communication in Action. NJ. 1995. P. 10.

[5] Шиллер Р. Нарративная экономика. Новая наука о влиянии вирусных историй на экономические события. М., 2024. С. 10.

[6] Tetlock P. Giving Content to Investor Sentiment: The Role of Media in the Stock Market // Journal of Finance, Forthcoming. 2005. 62(3). P. 1139–1168.

[7] Vaswani A., Shazeer N. et al. Attention is All You Need // Neural Information Processing Systems. 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (дата обращения: 26.09.2024).

[8] Соссюр Ф. де. Курс общей лингвистики // Соссюр Ф. де. Труды по языкознанию. М., 1977. С. 31–273.

[9] Ruder S., Peters M. at al. Transfer Learning in Natural Language Processing // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. Minneapolis, 2019. P. 15–18.

[10] Nekoto W. at al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages // Findings of the Association for Computational Linguistics: EMNLP. Stroudsburg, 2020. P. 2144–2160.

[11] Talmor A. et al. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019. Vol. 1. P. 4149–4158.

[12] Эко У. Роль читателя. Исследования по семиотике текста. М., 2005. С. 17.

[13] New minister promises to tackle living costs, Chinese imports // Bangkok Post. 12.09.2024. URL: https://www.bangkokpost.com/business/general/2863898/new-minister-promises-to-tackle-living-costs-chinese-imports. (дата обращения: 26.09.2024).

[14] Trade-in campaign bearing fruit // ChinaDaily. 26.09.2024. URL: https://www.chinadaily.com.cn/a/202409/26/WS66f4a511a310f1265a1c4dc5.html (дата обращения: 27.09.2024).

[15] Хобсон Л., Ханнес Х., Коул Х. Обработка естественного языка в действии. СПб., 2020. С.103–104.

[16] Baccianella S. et. al. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining // Proceedings of LREC. 2010. P. 2200–2204.

[17] Peng H. et al. Knowing What, How and Why: A Near Complete Solution for Aspect-based Sentiment Analysis // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. 34(05), P. 8600-8607.

[18] Albanese, F., Pinto, S. Semeshenko, V., Balenzuela, P. What matters, context or sentiment?: Analysing the influence of news in U.S. elections using Natural Language Processing // ArXiv. 2019. n. page. URL: https://www.researchgate.net/publication/335908711_What_matters_context_or_sentiment_Analysing_the_influence_of_news_in_US_elections_using_Natural_Language_Processing (дата обращения: 29.08.2024).

[19] Kim J., Kim W. Advanced Natural Language Processing Analysis on Cross-Border Media Sentiment from China and South Korea // International Area Studies Review. 2024. Vol. 27, no.1. P. 43–56.

[20] Liu, Z. et. al. Finbert: A pre-trained financial language representation model for financial text mining // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2021. P. 4513–4519.

[21] Gössi, S. et. al. FinBERT-FOMC: Fine-Tuned FinBERT Model with sentiment focus method for enhancing sentiment analysis of FOMC minutes // Proceedings of the Fourth ACM International Conference on AI in Finance, 2023. P. 357–364.

[22] Портал Торговые Новости. ЦАД ВАВТ, Институт Международной экономики и финансов. Москва, 2024. URL: https://trade-news.vavt.ru/ (дата обращения: 30.09.2024).

[23] He P., Gao J., Chen W. DeBERTav3: Improving DeBERTa Using Electra-Style Pre-Training With Gradient-Disentangled Embedding Sharing // Arxiv. 2023. URL: https://arxiv.org/abs/2111.09543 (дата обращения: 15.09.2024).

[24] Турецкий экспорт в Россию упал на фоне указа США о вторичных санкциях // РБК. 02.03.2024. URL: https://www.rbc.ru/business/02/03/2024/65e32a199a79477d9581f093 (дата обращения: 29.08.2024).

[25] Санкции не препятствуют поставкам российских удобрений, утверждает Байден // РИА Новости. 21.09.2022. URL: https://ria.ru/20220921/udobreniya-1818509563.html (дата обращения: 29.08.2024).

[26] The US and Europe are searching for a way to limit Russia's oil revenues without driving a price spike, Janet Yellen says // Business Insider. 09.06.2022. URL: https://www.businessinsider.nl/the-us-and-europe-are-searching-for-a-way-to-limit-russias-oil-revenues-without-driving-a-price-spike-janet-yellen-says/ (дата обращения: 29.08.2024).

[27] Портал Торговые Новости. ЦАД ВАВТ, Институт Международной экономики и финансов. Москва, 2024. URL: https://trade-news.vavt.ru/ (дата обращения: 30.09.2024).

БИБЛИОГРАФИЯ:

Луман Н. Реальность массмедиа / Н. Луман, пер. с нем. А. Ю. Антоновский. – М. : Праксис, 2005. – 256 с. @@ Luman N. Real`nost` massmedia / N. Luman, per. s nem. A. Yu. Antonovskij. – M.: Praksis, 2005. – 256 s.

Хобсон Л., Ханнес Х., Коул Х. Обработка естественного языка в действии / Хобсон Л., Ханнес Х., Коул Х. – Санкт-Петербург.: Питер, 2020. – 576 с.: ил. @@ Xobson L., Xannes X., Koul X. Obrabotka estestvennogo yazy`ka v dejstvii / Xobson L., Xannes X., Koul X. – Sankt-Peterburg.: Piter, 2020. – 576 s.: il.

Шиллер Р. Нарративная экономика. Новая наука о влиянии вирусных историй на экономические события / Р. Шиллер, пер. с англ. Е. Калугин. – М.: Бомбора, 2024. – 416 с. @@ Shiller R. Narrativnaya e`konomika. Novaya nauka o vliyanii virusny`x istorij na e`konomicheskie soby`tiya / R. Shiller, per. s angl. E. Kalugin. – M.: Bombora, 2024. – 416 s.

Эко У. Роль читателя. Исследования по семиотике текста / У. Эко. – Москва : Corpus, 2005. – 640 с. @@ E`ko U. Rol` chitatelya. Issledovaniya po semiotike teksta / U. E`ko. – Moskva : Corpus, 2005. – 640 c.

Albanese F., Pinto S., Semeshenko V., Balenzuela P. What matters, context or sentiment?: Analysing the influence of news in U.S. elections using Natural Language Processing // ArXiv. 2019. URL : https://www.researchgate.net/publication/335908711_What_matters_context_or_sentiment_Analysing_the_influence_of_news_in_US_elections_using_Natural_Language_Processing

Baccianella S. et. al. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining // Proceedings of LREC. – 2010. – P. 2200–2204.

Boumans D., Müller H., Sauer S. How media content influences economic expectations: Evidence from a global expert survey // Journal of Forecasting. 2023. URL: https://onlinelibrary.wiley.com/doi/10.1002/for.2961

Gössi S. et. al. FinBERT-FOMC: Fine-Tuned FinBERT Model with sentiment focus method for enhancing sentiment analysis of FOMC minutes // Proceedings of the Fourth ACM International Conference on AI in Finance. – 2023. – P. 357–364.

He P., Gao J., Chen W. DeBERTav3: Improving DeBERTa Using Electra-Style Pre-Training With Gradient-Disentangled Embedding Sharing // Arxiv. – 2023. – URL: https://arxiv.org/abs/2111.09543

Kim J., Kim W. Advanced Natural Language Processing Analysis on Cross-Border Media Sentiment from China and South Korea // International Area Studies Review. – 2024. – Vol. 27, no.1. – P. 43–56.

Liu Z. et. al. Finbert: A pre-trained financial language representation model for financial text mining // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. – 2021. – P. 4513–4519.

Nekoto W. at al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages // Findings of the Association for Computational Linguistics: EMNLP. – Stroudsburg : Association for Computational Linguistics, 2020. – P. 2144–2160.

Peng H. et al. Knowing What, How and Why: A Near Complete Solution for Aspect-based Sentiment Analysis // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – 34(05). – P. 8600-8607.

Ruder S., Peters M. at al. Transfer Learning in Natural Language Processing // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. – Minneapolis : Association for Computational Linguistics, 2019. – P. 15–18.

Talmor A. et al. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Minneapolis : Association for Computational Linguistics, 2019. – P. 4149–4158.

Tetlock P. Giving Content to Investor Sentiment: The Role of Media in the Stock Market // Journal of Finance, Forthcoming. – 2005. – 62(3). – P. 1139–1168.

Van Dijk T.A. Power and the news media // Political communication and action / Ed. by D. Paletz. – Cresskill, NJ: Hampton Press, 1995. – P. 9–36.

Vaswani A., Shazeer N. et al. Attention is All You Need // Neural Information Processing Systems. 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html