Федеральное государственное автономное образовательное учреждение
высшего образования
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет экономических наук

Образовательная программа «Экономика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Использование моделей машинного обучения для предсказания финансовых показателей на основе новостных данных

Application of Machine Learning to the Prediction of Financial Indicators Based on News Data

Выполнил:
студент группы БЭК186,
Драмбян Давид Ваагнович

Руководитель:
доцент факультета
экономических наук,
Мамедли Мариам Октаевна

Москва, 2022

Содержание

Введение

Финансовый рынок на протяжении многих лет агрегирует колоссальные объемы капитала во всем мире. Домохозяйства, банки, организации и даже государства инвестируют ресурсы с целью сохранить и преумножить свое богатство. В самом деле, конкуренция на финансовом рынке достигает такого уровня, при котором даже малейшая информация, сигнализирующая о потенциальном будущем состоянии этого рынка, представляется чрезвычайно полезным активом, поскольку любое дополнительное знание создает для инвестора конкурентное преимущество. Развитие математики и, в особенности, математической статистики, эконометрики и машинного обучения, в совокупности с эволюцией информационных технологий в существенной степени изменили подходы к анализу рынков. Превосходство компьютера над человеком в объемах запоминаемой и интерпретируемой информации, скорости ее обработки и стабильности стало еще более очевидно в XXI-ом веке. Сложившееся понимание и развитие научной области стали предпосылками активного применения методов машинного обучения для предсказания поведения финансового рынка и дали импульс развития исследованиям и разработке.

Данная работа, в свою очередь, фокусирует внимание на изучении одного из наиболее загадочных и противоречивых, с точки зрения своего влияния на движение показателей финансового рынка, источников информации: текстов новостей. Не секрет, что средства массовой информации и, в частности, огромный новостной поток, производимый ими и доступный в цифровую эпоху миллиардам пользователей, на регулярной основе бомбардируют читателей информацией, так или иначе формирующей мнения и оказывающей влияние на их поведение. В свою очередь, инвесторы, обладающие полной информацией в открытой экономике, согласно гипотезе эффективного рынка (efficient market hypothesis — EMH), потребляют доступную информацию (в

виде новостного фона) и отражают ее в полной мере в стоимости рыночного актива посредством изменения объемов спроса и предложения на основе принимаемых ими решений [13, стр. 1]. Именно поэтому, анализ этого потока информации может проливать свет на некоторые причинно-следственные связи в принятии решений экономическими агентами, и, в конечном итоге, создавать возможность предсказания их поведения, формирующего цены торгуемых активов на финансовом рынке. Учитывая исторический контекст и прикладной характер исследования, его основная цель заключается в оценке предиктивных возможностей моделей машинного обучения в предсказании движения финансового рынка на основе текстов новостей и выделении понятных закономерностей влияния этой текстовой информации на рынок. Предметом изучения являются методы обработки и векторизации текстов, их анализа и построения моделей машинного обучения для предсказания целевой финансовой переменной. Погружение в специфику этих инструментов призвано дать ответ на исследовательский вопрос и определить лучшие подходы к решению задачи предсказания на основе новостных данных.

Оперируя в рамках EMH, главной гипотезой исследования представляется возможность предсказания поведения экономических агентов на основании данных новостей, и вследствие предсказания показателей стоимости рыночных активов, формируемых ожиданиями и действиями этих агентов. Учитывая высокий интерес к практическим приложениям этой гипотезы, некоторое количество работ уже было посвящено данной проблеме, несмотря на ее сравнительно недавнее появление. Проведенные исследования должны стать источником примеров как наилучших практик, так и уже совершенных ошибок, для создания фундамента этой работы, исключения повторений в заключениях и оптимального получения новых выводов и исследовательской ценности в виде улучшения метрик качества моделей, создания эффективного предсказательного инструмента и открытия закономерностей влияния новостей на финансовый рынок.

Обзор литературы

Несмотря на цель продемонстрировать свежие, актуальные результаты, очевидно, что данное исследование не является первым, и тем более единственным, ориентированным на похожую проблему. Количество независимых исследований в этой области стало, в особенности, велико в течение последнего десятилетия в связи с ростом популярности методов машинного обучения и их применения к анализу финансовых рынков. Использование этих исследований в качестве фундамента целесообразно не только для того, чтобы избежать бесполезного повторения в анализе, но, что, пожалуй, более важно, как некоторую отправную точку, пример как лучших практик, так и, наоборот, неэффективных методов и источник научного вдохновения.

Статья Роберта Шумейкера и Хсинчуня Чена хронологически является одной из самых ранних работ, содержащих прототип используемой в данном исследовании гипотезы и алгоритмов машинного обучения в качестве инструмента решения поставленной проблемы. Их «Textual analysis of stock market prediction using breaking financial news: The AZFin text system» преподнесла связь теоретического концепта (гипотезы эффективного рынка) с потреблением новостных данных участниками рынка, и, как следствие, колебанием рыночных цен, вызванных этим [15, стр. 3]. Согласно статье, спрос на акции крупных компаний, используемых для расчета индекса S&P500, можно моделировать с помощью новостей, а наилучший результат в метриках достигается путем комбинирования новостных данных и информации о ценах на момент прогноза [15, стр. 26]. Важно отметить, что исследование в значительной степени сосредоточено на предобработке текста, распознавании паттернов в речи и речевом анализе текстов, и среди прочих результатов, оно демонстрирует, что наиболее эффективным способом получения информации из статьи или новости для прогнозирования цены

является извлечение имен собственных из предложений. Исследователи полагают, что имена собственные, как правило, содержат меньше шума по сравнению с остальным текстом, и следовательно, предоставляют более релевантную информацию для текстового анализа [15, стр. 26]. Несмотря на возраст статьи, такое фундаментальное понимание полезно для будущих исследований: эксперименты со стратегиями выбора слов перед дальнейшей предобработкой могут помочь выжать больше информации из меньшего количества текста, что сделает модель как быстрее, так и точнее. Более того, некоторые современные исследования показывают, что тщательность в выборе слов из текстов оправдана, и в дополнение к этому, полезным методом может оказаться обогащение их данными о так называемых «ad hoc» событиях, поскольку некоторые колебания рынка вызваны, в первую очередь, не сопровождающими новостями, а самими событиями, влияющими на инвестора и его поведение [8, стр. 8]. Однако описанные в статье модели можно считать тривиальными по сегодняшним меркам и даже несколько старомодными: Genetic Algorithm, Naïve Bayesian classifier и Support Vector Machine (SVM) [15, стр. 7]. Модели основаны на вероятностях предпосылках, что, безусловно, повышает их интерпретируемость, но сравнение их со state-of-the-art алгоритмами нашего времени определенно представляет их не в лучшем свете с точки зрения предсказательной способности и гибкости. Следовательно, использование градиентного бустинга и нейронных сетей предпочтительнее в нашем исследовании, поскольку одной из целей по-прежнему остается попытка достичь более высоких результатов с точки зрения целевых метрик.

Еще одним фундаментальным трудом, который может послужить подходящей основой для нашего исследования, является статья Хуины Мао, Скотта Каунтса и Йохана Боллена «Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data». В исследовании представлен более статистический подход к проблеме, а особое внимание уделяется выбору

источников данных, методов предобработки и извлечению признаков из новостей. Как и в статье Роберта Шумейкера и Хсинчуня Чена, в этой работе так же утверждается, что гипотеза эффективного рынка и ее влияние на колебания акций являются основной предпосылкой идеи, лежащей в основе исследования, поскольку поведенческие и эмоциональные факторы играют центральную роль при принятии финансовых решений [13, стр. 1]. Трейдеры и инвесторы, прежде всего, — люди, что делает их уязвимыми к перепадам настроения, иррациональным и импульсивным решениям, страху и жадности, на которые, в особенности в информационную эпоху, влияют настроения, приносимые новостями. Переходя к деталям и использованным в исследовании методам, важно отметить, что существенную ценность ему придает работа, проделанная над извлечением признаков. В работе показано, что сами по себе тексты новостей могут быть не единственной информацией, доступной для извлечения в качестве индикатора общественных намерений на финансовом рынке. Во-первых, парсить данные полезно не только из традиционных поставщиков новостей, но и из некоторых более современных источников, таких как, например, Twitter. Во-вторых, непосредственно социальные обследования и опросы очень точны в получении прямой информации от инвесторов [13, стр. 2]. Однако подобных источников довольно мало, а те, что доступны зачастую недостаточно надежны. В-третьих, целесообразно извлечение большего количества информации из социальных сетей: классификация твитов как бычьих и медвежьих, чтобы уловить среднее настроение в обществе, и количественная оценка встречаемости финансовых терминов в твитах и поисковых запросах в Интернете, чтобы определить уровень активности на рынке [13, стр. 3]. В-четвертых, с помощью «индекса негативных слов», рассчитываемого как частота, с которой слов негативного настроения (такие как «кризис», «провал», «потери», …) встречаются в общем количестве слов, используемых в новостях, статьях, твитах и поисковых запросах в Интернете [13, стр. 2]. Согласно гипотезе исследователей, подобные «негативные» слова влияют на

настроение и, как следствие, на поведение человека больше, чем положительные. Исходя из таких предпосылок, «индекс негативных слов» действительно может служить флагом настроения в средствах массовой информации. В-пятых, нельзя пренебрегать использованием финансовых данных [13, стр. 3]. История цен, индексы волатильности, фундаментальные финансовые индикаторы, инструменты технического анализа и объемы торгов на конкретном рынке в определенный момент времени. Эти данные полезны не только как индикатор ситуации на рынке и настроений, но и как характеристика, которую учитывают сами трейдеры при принятии финансовых решений. Поскольку они влияют на экономических агентов на рынке, полезно использовать их в качестве переменной при прогнозировании изменения цены торгуемых акций, фондов, финансовых индикаторов и других инструментов. Все вышеперечисленные методы стали новаторскими с точки зрения извлечения дополнительной информации и помощи в анализе настроений на рынке. Таким образом, в статье представлен довольно широкий, но при этом не менее тщательный подход к процедуре извлечения признаков. Использование представленных методов позволяет глубже понять ситуацию на рынке и увеличивает шансы на правильное определение настроения агентов на рынке.

Статья Эндрю Ло и А. Крэйга Маккинлея «Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test» представляет подход к исследованию, значительно больше основанный на экономической интуиции. Оперируя в рамках парадигмы гипотезы эффективного рынка, исследователи утверждают, что в отличие от подхода «случайного блуждания», при котором колебания цены на фондовом рынке определены как не поддающиеся предсказанию, в реальной жизни изменения цен зависят от паттернов поведения экономических агентов, действующих на рынке [10, стр. 1-2]. Замечание Кейнса о том, что инвесторы и трейдеры, как правило, действуют не просто неэффективно и непоследовательно, а вовсе спонтанно,

все еще остается в силе. Это означает, что рыночная цена активов находится под постоянным влиянием спроса и предложения на этом рынке, которые, в свою очередь, находятся определяются иррациональным духом инвесторов [10, стр. 1]. Учитывая, что эти эмоциональные колебания зачастую вызваны именно новостями или, в иных случаях, эти колебания сами по себе являются объектом новостей и причиной их появления, более глубокий анализ этих данных должен оказаться действенной стратегией для прогнозирования конечных цен финансовых инструментов в заданный момент времени.

Приняв во внимание экономическую интуицию, стоящую за выбранной гипотезой и методы обработки данных и извлечения признаков для работы над проблемой, следующим логическим шагом станет исследование потенциальных моделей в попытке найти наиболее подходящие архитектуры для данной задачи. «Deep learning for stock market prediction from financial news articles» Мануэля Варгаса, Беатрис де Лимы и А. Г. Евсукова представляет собой полноценное подробное погружение в методы глубинного обучения для решения задачи классификации текстов. Исследование проведено с целью таргетирования колебаний индекса Standard & Poor’s 500 (S&P500) внутри торгового дня — задача, довольно похожая нашу [18, стр. 2]. Для решения поставленной задачи исследователи прибегают к использованию ансамбля из двух моделей: одна с использованием технических индикаторов в качестве признаков, а другая с использованием текстов новостей, соответственно [18, стр. 2]. Предварительно обработанные текстовые данные подаются на вход варианту рекуррентной нейронной сети: Long Short-Term Memory (LSTM) архитектуре. Такой выбор оправдан строением рекуррентных нейронных сетей, позволяющим им обрабатывать список слов как последовательность. Подобный подход в некоторой степени реплицирует прочтение и соответственно восприятие текста человеком. Выходы двух независимых моделей (одна построена на заголовках новостей, другая — на технических индикаторах) объединяются, взвешиваются и используются в качестве

входных данных для финального классификатора на последнем уровне. Результаты применения вариантов модели, продемонстрированные в статье, подтверждают эффективность методов глубинного обучения в задаче предсказания изменения цен на финансовом рынке, подтверждая целесообразность подобного подхода к решению обозначенной проблемы.

Описание данных

Мотивация при выборе данных

При решении задачи предсказания изменения финансовых показателей на основе текстов новостей необходимо, в первую очередь, определить источники данных, наиболее подходящие для этой задачи. Глобально, используемую для предсказания информацию можно разбить на три смысловых блока: данные новостей, технические показатели финансового рынка и непосредственно финансовый показатель, выступающий в качестве целевой переменной. Начинать погружение в специфику и само исследование выбранных данных следует, в первую очередь, с определения мотивации, стоящей за их выбором.

Рассуждая внутри парадигмы гипотезы эффективного рынка, можно утверждать, что вся информация, появляющаяся в открытом доступе оказывается в полной мере отражена в изменении стоимости активов, к которым относится. В некоторой степени, такой подход согласуется и с кейнсианской позицией, согласно которой экономические агенты ведут себя при принятии инвестиционных решений скорее не рационально и взвешенно, а спонтанно и импульсивно, как бы подвергаясь «животному духу» («animal spirits») [10, стр. 1]. В конечном итоге, решения экономических агентов и, как следствие, их поведение на финансовом рынке, является продуктом восприятия тех или иных новостей этими агентами, окрашенного, в первую очередь, их эмоциональным фоном. Таким образом, сумев оценить кумулятивный эмоциональный и смысловой окрас новостного потока в моменте времени, станет возможно предсказание средней реакции инвесторов, безусловно возникающей согласно гипотезе эффективного рынка, на этот поток, что предоставит возможность оценить изменение их поведения, объемов спроса и предложения и, в конечном итоге, колебания финансового

показателя. Важно заметить, что новостной поток не только непосредственно влияет на восприятие трейдером или инвестором потенциальной доходности актива, но и воздействует на его ожидания относительно поведения остальных участников рынка, усиливая изначальный эффект, обогащая его страхом промедления и опасностью пропустить наиболее выгодную возможность. В результате подобной механики, такой процесс, запущенный одной новостью, оказывает не только влияние прямо на поведение экономического агента, но и становится, по своей сути, саморазгоняющимся до тех пор, пока значительная масса инвесторов снова не переосмыслит собственное восприятие стоимости конкретного актива, оценив его как переоцененный, и, как следствие, изменит как свои ожидания относительно его доходности, так и собственную инвестиционную стратегию касательного актива [16, стр. 25]. Само значение цены, в то же время, на протяжении всего описанного процесса является отражением ожиданий и инвестиционных стратегий, выбранных на основании этих ожиданий торгующими агентами в моменте времени. Отсюда, необходимо добиться того, чтобы финансовый показатель, используемый в качестве таргета (или целевой переменной) был логически согласован с выбранными предпосылками работы: прозрачно и в полной мере освещался средствами массовой информации, открыто торговался на свободном рынке и, предпочтительно, отражал ожидания экономических агентов. Тогда и только тогда таргет окажется репрезентативным относительно реакции трейдеров на возникающий эмоциональный фон новостей (news sentiment), и исследование будет валидным.

Финансовый показатель

Итак, выбранная для анализа целевая переменная должна соответствовать базовым предпосылкам исследования, описанным выше: информация, касающаяся ее как непосредственно, так и косвенно, должна распространяться

непрерывно и открыто для всех участников рынка, а сам актив открыто торговаться на свободном рынке. Этим условиям без особого труда удовлетворяют большинство ценных бумаг, деривативов, фондов и индексов, торгуемых на финансовом рынке. Помимо прочего, наблюдаемый показатель должен торговаться в сравнительно больших и устойчивых объемах. Это позволит обеспечить значимость используемых показателей и избежать «выбросов» в данных: всплесков или резких падений, необоснованных с точки зрения общедоступной информации на рынке. Более того, крупные объемы будут свидетельствовать о вовлеченности в торги большого количества агентов — применимый с точки зрения гипотезы эффективного рынка критерий, который в свою очередь «страхует» от чрезмерного влияния на цену со стороны одного инвестора. Однако еще одним важным критерием при выборе целевой переменной, о применении которого прежде не заходила речь, может являться восприятие актива инвесторами. Для того, чтобы логическая связь была максимально прозрачной и объяснимой, данное исследование сфокусируется на изучении влияния новостей на настроение экономических агентов относительно движения всего рынка в целом. Подобный подход позволит освободиться от накладывающихся друг на друга эффектов от изменения спроса и предложения на разные компании, отрасли или регионы, и даст возможность рассматривать более широкий вопрос, охватывающий все ожидания сразу — вопрос движения всей экономики сразу: роста или падения. Таким образом, необходимо, чтобы используемый таргет (целевая переменная) воспринимался инвесторами как комплексный показатель состояния рынка в конкретный момент времени. В этом случае, подобный показатель будет выступать в качестве отражения ожиданий участников рынка от экономики, в целом. Инвесторы покупают такой актив, когда верят в то, что он недооценен, или что экономика попросту продолжит расти, и продают его, когда ожидают падения, кризиса, регрессии. Несмотря на то, что на первый взгляд подобное изобилие условий может показаться чрезмерно большим при выборе показателя, на самом деле, в нашем распоряжении

остается сразу несколько удовлетворяющих им композитных индексов, агрегирующих поведения большого количества наиболее крупных компаний рынка. В первую очередь, как для Соединенных Штатов Америки, так и для всей западной цивилизации, такими показателями безусловно выступают индекс *NASDAQ Composite* для высокотехнологичных компаний, производящих программное обеспечение, компьютеры и много другое, *Dow Jones Industrial Average* для компаний промышленного сектора и *S&P 500* для наиболее крупных компаний: тех, что имеют наибольшую капитализацию. Эти показатели достаточно коррелированы с настроениями инвесторов и в полной мере отражают общедоступную информацию в изменениях цены [14, стр. 72]. Несмотря на то, что расчет значений этих индексов происходит, преимущественно, на основе результатов торгов бумагами американских компаний, широкая география их деятельности и большая доля американского производства в мировой экономике делают индексы чувствительными к глобальным трендам и шокам. В добавок, США, в действительности, являются одним из основных агрегаторов финансового капитала, поскольку именно на американских биржах торгуется львиная доля наиболее крупных публичных компаний вне зависимости от их происхождения: от китайской Alibaba Group до немецкого Deutsche Bank [22]. При этом, в рамках исследования представляется возможным сфокусироваться и на более узком рынке: например, выбрав российский аналог композитного индекса с целью моделировать ожидания экономических агентов относительно российского рынка. Рассчитываемый в рублях индекс Московской Биржи или рассчитываемый в долларах США индекс РТС могут выступать в качестве подобных индикаторов. Однако это исследование сфокусируется на предсказании ожиданий от движения всего рынка, поэтому в качестве целевой переменной будут использованы американские индексы. Такой выбор обусловлен, в первую очередь, более стабильным поведением американского фондового рынка, его меньшей подверженности экономическим шокам, и большей независимости от внешних факторов: в то время как кризис на

американском рынке мог повлиять на рост или падение российской экономики (как в случае с ипотечным кризисом 2007-2008-ых годов в Соединенных Штатах Америки), обратное если и имело место быть, то в значительно меньшей степени (по крайней мере, на протяжении последних двадцати лет). Таким образом, именно композитные индексы американского фондового рынка выбраны в качестве используемых таргетов. Эти индексы крайне ликвидны, популярны и, что немаловажно, именно они абсорбируют в изменениях собственной цены ожидания инвесторов относительно состояния экономики, позволяя, основываясь на гипотезе эффективного рынка, предположить, что непосредственно эти изменения не подвержены случайному блужданию, а могут быть смоделированы [10, стр. 1-2].

Более конкретный выбор одного индекса из трех описанных выше не имеет большого значения на этой стадии исследования. S&P 500 и Dow Jones Industrial Average являются чуть более ортодоксальными вариантами по сравнению с NASDAQ Composite, поэтому с точки зрения исторической репрезентативности всего состояния экономики выбор в пользу одного из них может быть более оправдан, однако на горизонте нескольких последних лет различия несущественны, что подтверждается даже визуальным сравнением динамик показателей [Рисунок 1]. С одной точки зрения, S&P 500 может показаться более объективным выбором целевой переменной, поскольку он объединяет в себе информацию о крупнейших компаниях всего рынка, а не отдельного сегмента. Для инвесторов, композитный индекс компаний-гигантов является релевантным индикатором производственных темпов во всем мире: если их деятельность нарушается, страдают производства и потребители по всей планете. Более того, применимость S&P 500 не только продиктована логическим пониманием стимулов, но и подтверждена статистически: высокая корреляция с темпами роста макроэкономических показателей экономик государств из G-20 [7, стр. 564]. Тем не менее, изменения Dow Jones Industrial Average и S&P 500 имеют среднюю

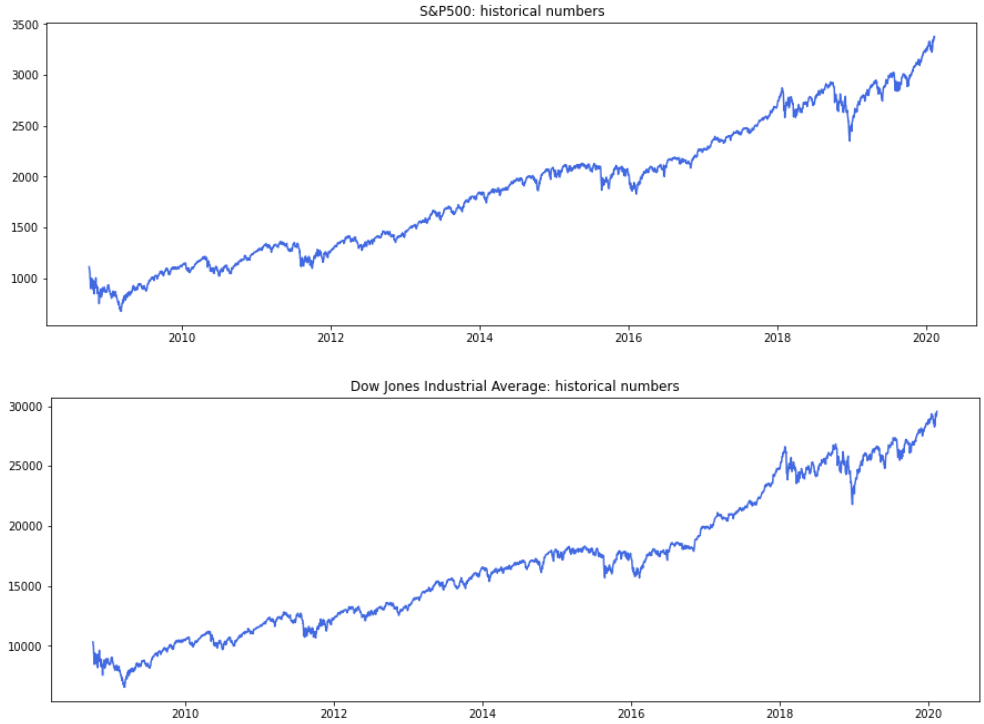


Рисунок 1

[Источник: Приложение 1]

корреляцию порядка 0.95 между собой (со значениями, превышающими средние в последние десятилетия, что немаловажно), фактически означая, что индексы неразрывно связаны друг с другом, по крайней мере, в восприятии большинства трейдеров [21, стр. 2]. В таком ключе, выбор между ними обусловлен, в первую очередь, практическими соображениями и удобством сбора данных, исходя из которых в качестве основного целевого показателя выбран Dow Jones Industrial Average.

Сами данные о значениях индексов есть в открытом доступе и доступны для скачивания [25]. Выбрав финансовый показатель, необходимо определить то, в каком виде он будет представлен для использования в качестве таргета. Потенциальным решением может оказаться предсказание дневного изменения индекса в количестве пунктов. Однако наиболее элементарным, но при этом подходящим под задачи исследования, является бинарный таргет, равный 1, в случае если индекс вырос за день, и 0, если снизился. Поскольку исследование

не несет задачи предсказания того, насколько изменится индекс, подобной целевой переменной, определяющей направление движения по результатам торговой сессии, но не скорость, вполне достаточно. Основная цель заключается непосредственно в определении ожиданий инвесторов относительно рынка: оптимистичных или пессимистичных, — и, как следствие, классификации движения индекса от начала дня до конца. Такой среднесрочный период предсказания обусловлен, в первую очередь, предпосылками исследования, которые основаны на гипотезе эффективного рынка, предполагающей «отыгрывание» новостей участниками рынка. Это «отыгрывание» безусловно требует времени для того, чтобы критическая масса экономических агентов потребила новостной поток и приняла решение относительно корректировки собственной инвестиционной стратегии.

При решении задачи классификации с помощью методов машинного обучения, важно учитывать баланс классов в представленной выборке. В данном случае, классы в целевой переменной достаточно неплохо сбалансированы: порядка 52% торговых сессий за выбранный период, с 2008 по 2016 год, закрылись ростом индекса Dow Jones Industrial Average, и, соответственно, около 48% остальных дней увенчались падением или

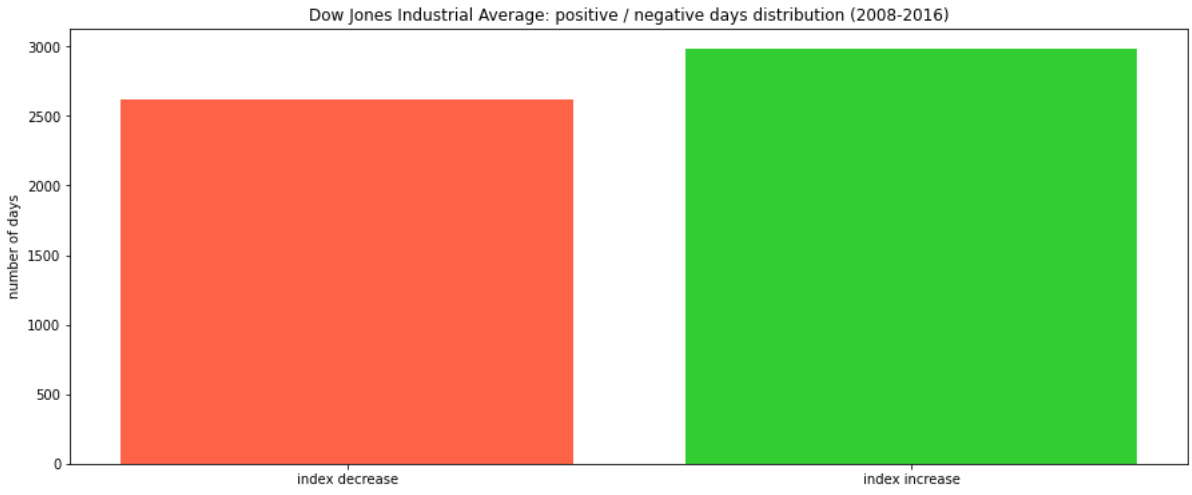


Рисунок 2

[Источник: Приложение 1]

закрытием на отметке равной открытию [Рисунок 2]. Подобная картина удовлетворительна с точки зрения будущего построения модели: примерное равновесие классов позволит избежать смещения модели, ее переобучения в пользу одного из классов и позволит сохранить ее чувствительность как к положительно-, так и отрицательно-размеченным объектам. Более того, с точки зрения проведения комплексного исследования стоит изучить распределение не только самого бинарного таргета, но и другой информации об изменении выбранного индекса. Так, например, распределение абсолютного значения дневного изменения самого индекса Dow Jones Industrial Average за период времени с начала 2000-го года асимптотически близко к нормальному с околонулевым центром в 2.79 и стандартным отклонением в 160 пунктов [Рисунок 3]. В то время как распределение размера дневной свечи (разница между максимальным значением индекса в

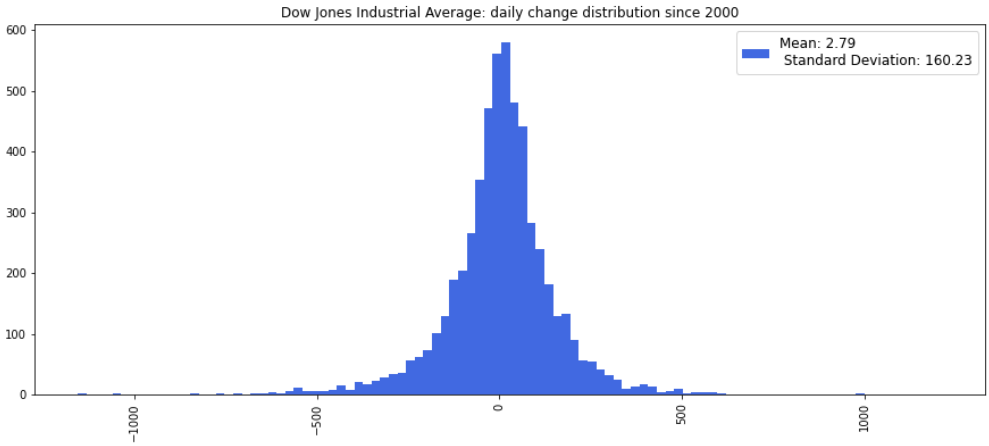


Рисунок 3

[Источник: Приложение 1]

течение дня и минимальным, соответственно) ближе к хи-квадрат со средним в 189 и стандартным отклонением в 156 [Рисунок 4]. Эти данные, имеющие близкие к каноническим распределения, что лишь добавляет им удобства в использовании и обработке, резонно использовать в качестве признаков, характеризующих историческое состояние рынка к моменту времени. Информация о размере свечи в течение нескольких дней, предшествующих

дню предсказания, как и информация об изменениях индекса, может быть полезна в качестве технических показателей для аппроксимации активности на рынке и тренда в движении финансового показателя.

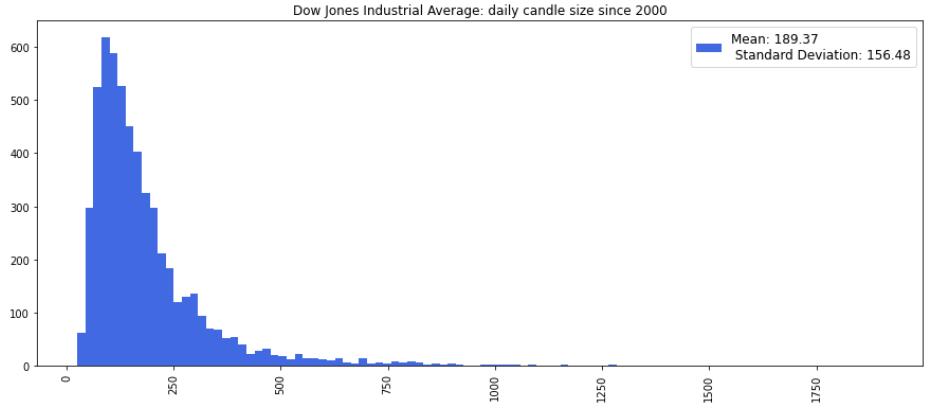


Рисунок 4

[Источник: Приложение 1]

Новостные данные

Итак, разобравшись с целевой переменной, необходимо перейти к сбору данных новостей для дальнейшего формирования признаков. Помня об описанных предпосылках исследования, необходимо удостовериться в том, что используемые новости общедоступны и используемы большим количеством экономических агентов, а средства массовой информации, публикующие их, авторитетны. Для этого будем использовать новости, публикуемые крупными мировыми изданиями. Площадкой консолидации большого количества подобного рода данных могут выступать социальные сети, регулярно агрегирующие новостной поток. Используем площадку Reddit и, в частности, ветку r/worldnews, публикующую в точности названия оригинальных новостных статей со ссылками на источник [29]. По состоянию на девятое апреля 2022-го года на нее подписано порядка тридцати миллионов участников, а новости, в свою очередь публикуются из авторитетных изданий:

Forbes, Business Insider, Reuters, The New York Times и других. В добавок, помимо того, что ветка агрегирует большое количество релевантной информации, сама платформа Reddit предоставляет потрясающе полезный функционал в виде голосов участников, оценивающих «горячесть» или полезность той или иной записи. Очевидно, использовать все новости из всех источников, созданные в течение одного дня, было бы практически функционально неосуществимо: длина текстового признака для одного таргет-дня была бы колоссальной (объединение всех новостей в один текст). К тому же, в этом нет существенной необходимости, поскольку многие средства массовой информации, находящиеся в одном информационном поле, так или иначе, дублируют новости друг друга, по крайней мере точки зрения смысловой нагрузки. В этом смысле, использование ветки Reddit решает обе проблемы: новости, дублирующие предыдущие, не публикуются, за редким исключением, а система голосов пользователей позволяет ранжировать новости между собой, делая выбор в пользу наиболее «горячих». Подобное ранжирование особенно релевантно, поскольку пользователи, голосующие на платформе, во многом и «являются голосом» экономических агентов, и новость, заинтересовавшая их, с большой долей вероятности повлияет и на инвестиционные решения агентов на финансовом рынке. Для начала, воспользуемся готовым набором данных из описанных источников, собранном и размещенном пользователем Kaggle в открытом доступе [27]. Тексты заголовков новостей, в привязке к дню публикации и изменению индекса Dow Jones Industrial Average по итогам торговой сессии, уже отсортированы по популярности, и среди них выделено 25 наиболее «горячих» для каждого дня.

Важный вопрос, возникающий на стадии сбора новостных данных, заключается в том, какую часть новости использовать для предсказания: заголовок или весь текст. Забегая вперед, попытка сравнить два подхода эмпирически демонстрирует, что модель, обученная на текстах заголовков,

показывает результат (с точки зрения основных метрик), не уступающий модели, построенной на полных текстах [Рисунок 5]. Подобное наблюдение

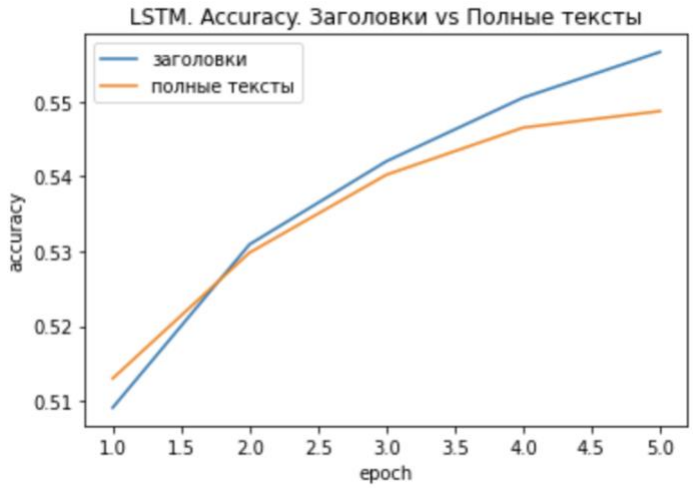


Рисунок 5

[Источник: Приложение 2]

объяснимо, в первую очередь, тем, что заголовки в подавляющем большинстве случаев содержат в себе суть всей статьи, сжатую до одного-двух предложений. В сущности, они отражают посыл всего текста, но при этом избавлены от большого количества излишней информации, содержащейся в тексте статьи, и значительно очищены от речевых конструкций, не несущих большой смысловой нагрузки: например, союзов и предлогов. Эта аргументация является причиной использования именно заголовков новостей в противовес целым текстам в релевантных исследованиях, решающих похожую задачу: в том числе, «Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data» и «Deep learning for stock market prediction from financial news articles», упомянутые в обзоре литературы [13][18]. Основываясь на комбинации best-practice решений отрасли и эмпирическом доказательстве, в качестве основного источника текстовых признаков будут использованы именно заголовки новостей.

Однако наличия самих текстов недостаточно для того, чтобы можно было предсказывать движение финансового индикатора с помощью них. Поскольку целевой переменной в задаче является флаг роста значения индекса по итогам торгового дня, объектом для предсказания должна быть некоторая комбинация новостей, относящихся к этому таргет-дню. Эта комбинация призвана отражать настроение (sentiment) всего новостного потока, на основе которого можно определить средние созданные ожидания у экономических агентов, их реакцию и, как следствие, само изменение значения финансового показателя. В качестве способа объединения набора наиболее популярных согласно ранжированию новостей в течение дня будет использована конкатенация текстов внутри этого дня для создания единого новостного потока, который и выступит в качестве текстового признака.

Для того, чтобы использовать полученные конкатенации для предсказания изменения индекса, их необходимо привязать непосредственно к самим таргет-дням. Сделать это можно несколькими способами. Первый и наиболее прямолинейный заключается в то, чтобы связывать все новости, пришедшие в течение дня, напрямую с изменением индекса по итогам торговой сессии того же дня. Однако данный подход вызывает некоторые вопросы по части справедливости такой связи. Несмотря на то, что большая часть новостей действительно производится в первой половине дня, во время работы биржи, какая-то их доля появляется в вечернее время, и даже гипотетически неспособна повлиять на котировки во время уже закрытой на тот момент торговой сессии. Более того, даже новости, пришедшие непосредственно в часы работы биржи, могут не быть полностью отыграны участниками рынка за отведенное время, оказывая дополнительный эффект на колебания цены актива на протяжении следующего торгового дня. Элегантным решением в такой ситуации может оказаться стратегия связывания дневной конкатенации новостей с таргетом следующей торговой сессии. Подобная тактика была применена в некоторых предшествующих работах, в которых и доказала свою

состоятельность [4, стр. 1419] и [18, стр. 2]. Интуиция, стоящая за этим подходом, заключается в том, что инвесторам на рынке требуется некоторое время для того, чтобы увидеть новость в колоссальном потоке информации, отреагировать на нее и скорректировать собственные ожидания относительно финансового актива, подкрепив их действиями. Применительно к большому количеству экономических агентов, этот принцип времени реакции особенно актуален: критическая доля инвесторов реагирует на новую информацию, в среднем за некоторое время, осуществляя корректировку стоимости актива согласно гипотезе эффективного рынка.

С целью убедиться в равномерности распределения таргет-дней во времени (это необходимо для избежания смещения предсказаний модели и уверенности в репрезентативности выборки), взглянем на него в разрезе нескольких временных сущностей. В первую очередь, для построения модели будут использованы данные за период с августа 2008 года по июль 2016 года включительно. Такая выборка обрезает крайние года, однако в остальном количество дней, используемых для предсказания, распределено равномерно на уровне порядка 250 таргет-дней в год, что вполне соответствует ожиданию

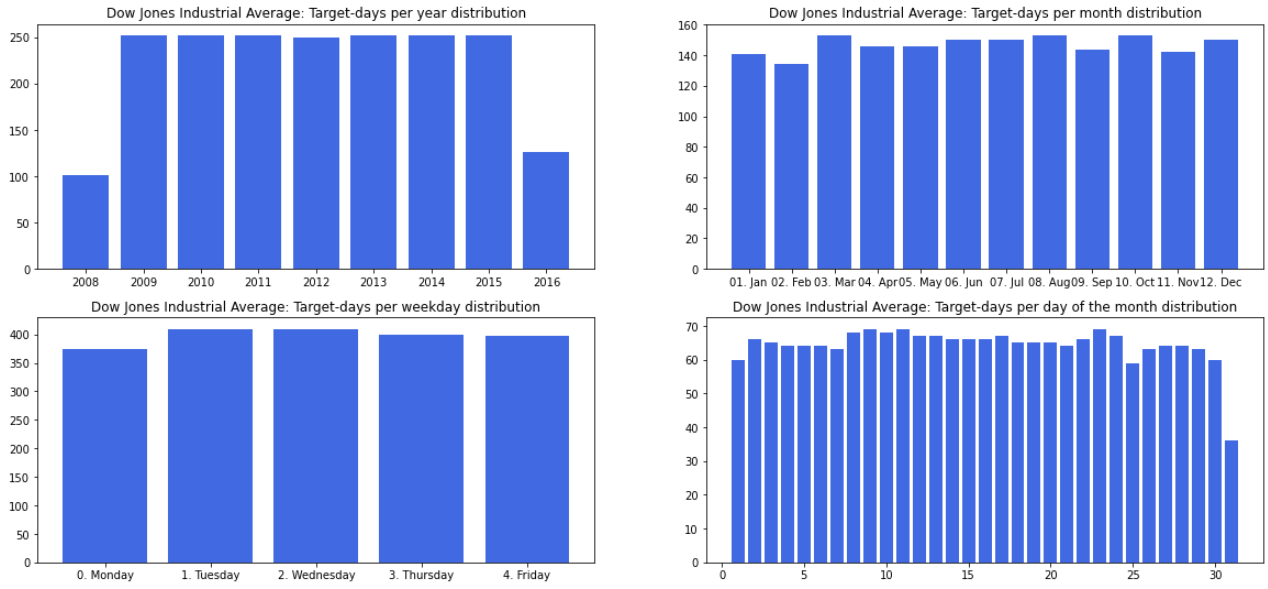


Рисунок 6

[Источник: Приложение 1]

в 5/7 (5 рабочих дней в неделю) от 365 дней в году, с учетом праздников, во время которых биржа закрыта [Рисунок 6].

Опускаясь на уровень ниже, отмечаем, что распределение по месяцам внутри года, как и распределение по дням недели за выбранный период сравнительно близки к равномерному [Рисунок 6]. Некоторые выбросы, особенно в разрезе месяцев, спровоцированы тем, что 2008-ой и 2016-ый годы представлены в данных в ограниченных вариантах. Однако взглянув на количество дней, сгруппированных по месяцам, используя года с 2009-го по 2015-ый, заметной особенностью так же является небольшое снижение в январе, вызванное праздниками, и феврале, вызванное коротким месяцем [Приложение 1]. Таким образом, используемые в качестве целевой переменной данные имеют ожидаемое распределение, распределение, которое, к тому же, близко к равномерному в большинстве временных разрезов.

Технические показатели финансового рынка

Помимо текстовых данных из новостей, в качестве признаков разумно использовать и некоторые технические индикаторы, несущие информацию о состоянии финансового рынка в моменте времени. Такой подход оправдан тем, что показатели, динамически оценивающие объемы торгов, волатильность актива, тренд в изменении его цены и многие другие характеристики, непосредственно несут информацию как о состоянии рынка, так и о самом активе, в момент, на который совершается предсказание. Более того, инструмент использования технических индикаторов финансового рынка в дополнение к текстовым данным был использован в фундаментальных работах, используемых в качестве источника best-practice методов в исследовании [18, стр. 2]. Использование успешного опыта предыдущих исследований может принести положительный результат.

Помимо тривиальных объемов торгов и размеров торговых свечей, резонно использовать и такие показатели, как тренд изменения стоимости актива за фиксированный промежуток времени до даты предсказания, скорость прироста цены (показатель, концептуально близкий к второй производной некоторого графика графика) и даже отрезок максимального роста цены в заданном временном диапазоне для аппроксимации степени и направления волатильности на рынке [Рисунок 7].

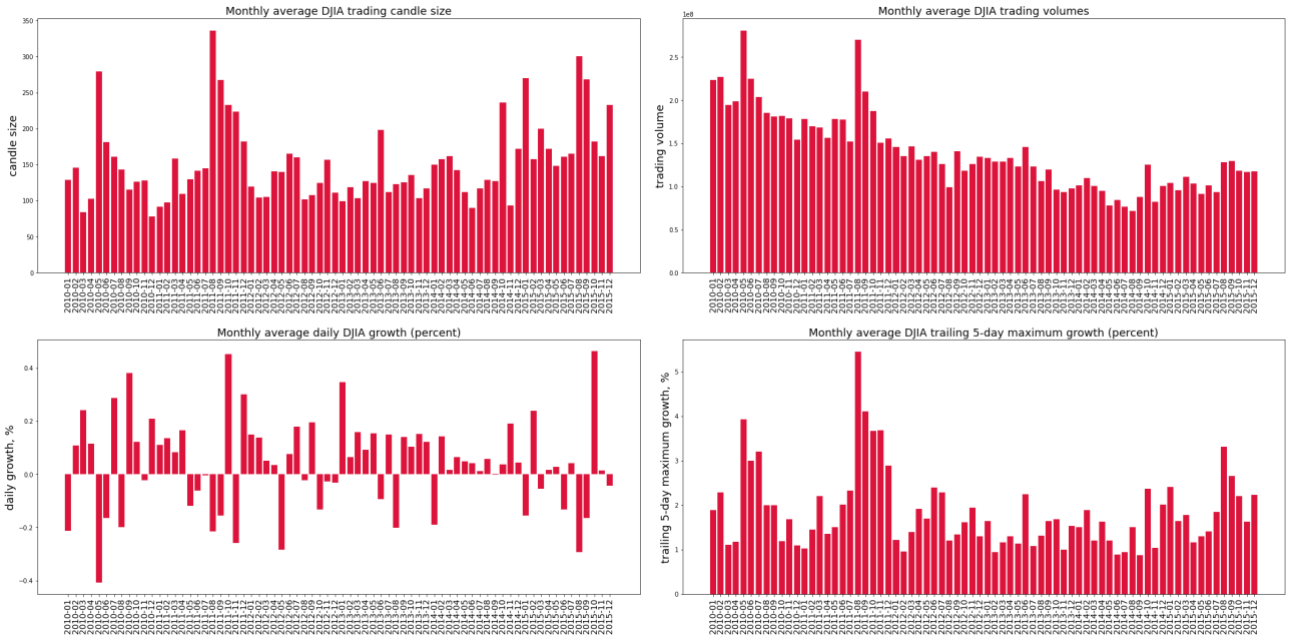


Рисунок 7

[Приложение 1]

Эти показатели, наряду с некоторыми другими добавлены в набор признаков для прогнозирования целевой переменной и комбинации с информацией, полученной из текстов новостей. Исходя из описанного понимания, объединение технических показателей с информацией, полученной из заголовков новостей, будет использовано для достижения большего эффекта и построения более точной модели.

Переход в признаковое пространство

Однако данные, а текстовые данные в особенности, далеко не всегда полезны и интерпретируемы в сыром виде. Большая часть данных так или иначе нуждается в дополнительной предобработке перед использованием в предсказательных моделях. Более того, предобработка и эксперименты с разными ее методами может поспособствовать более глубокому пониманию данных. Учитывая специфику задачи, резонно, в первую очередь, сконцентрироваться на обработке текстовых данных и генерации признаков из них. Очевидно, что слова из заголовков новостей в привычном для человека виде невозможно передать в математическую модель в качестве признака, однако, создав векторное представление для слов, это станет возможным. Правда, прежде чем переходить к векторизации, необходимо провести некоторые операции над самими словами. Для начала, предложения нужно токенизировать: то есть перевести строковое представление предложения в набор токенов. Токеном будем называть единицу письменной речи: слово, знак препинания, артикль и тому подобное. Такой шаг нужен для того, чтобы вести работу на более элементарном уровне, обрабатывая предложения не целиком, а по токенам. Для этого, будем использовать метод *build_analyzer()* из класса *feature_extraction.text.CountVectorizer* библиотеки *sklearn* [23]. Альтернативным вариантом для токенизации может быть метод *TweetTokenizer()* из библиотеки *nltk*, поскольку заголовки новостей по своей сути могут иметь схожесть с твитами, для которых адаптирован *TweetTokenizer()*: короткие, отрывистые фразы, подчас содержащие сленг, отдающие большое значение знакам препинания в вопросе передачи эмоций [28]. Далее, различные формы одних и тех же слов стоит привести к общему виду. Помимо того, что с точки зрения вклада в смысл предложения слова, например, ‘акция’ и ‘акции’, практически неразличимы, унификация так же позволит значительно сократить словарь для будущей модели (словарем будем называть все многообразие токенов, которое есть в тренировочных

данных). С этой целью более всего подходят методы лемматизации и стемминга, разница между которыми заключается в том, что лемматизация призвана привести однокоренные слова к общему корню, в то время как стемминг унифицирует слова более грубым обрезанием окончаний. Опыт применения этих методов свидетельствует о том, что они демонстрируют сравнительно близкий с точки зрения качества обработки результат в большинстве случаев. В рамках экспериментов над дизайном модели будем использовать оба метода по-отдельности, сравнивая качество между собой. Резонным методом предобработки также может считаться очистка текстов от стоп-слов. Стоп-словами называем токены, не несущие значительной смысловой нагрузки: артикли, предлоги, союзы. Мы предполагаем, что такие слова, по большей части, засоряют наш словарь и не содержат никакой полезной информации относительно эмоционального окраса текста.

Итак, первичная предобработка позволила значительно сократить словарь, и сгладила распределение частоты токенов [Рисунок 8]. Использование стеммера и словаря стоп-слов очистило тексты от токенов, не несущих существенной информации, по крайней мере визуально. Следующий шаг заключается в переводе предобработанных токенов в вектора, представляющие их. Одна из задач исследователя в области обработки естественного языка состоит в том, чтобы переход от слов к векторам (эмбеддинг) сохранял как абсолютные значения, так и отношения между словами. Однако, в качестве baseline-модели (базовой модели, далее — бейзлайн) разумно использовать наиболее примитивный метод векторизации текста *Bag-of-words*, несущий в первую очередь частотную информацию о токенах. Бейзлайн позволит наиболее быстро провалидировать модель и использовать полученные метрики в качестве отправной точки по отношению к будущим результатам экспериментов. Концептуально, *Bag-of-words* ставит в соответствие каждому тексту вектор длины размера словаря, где на каждой позиции в этом векторе находится счетчик количества вхождений

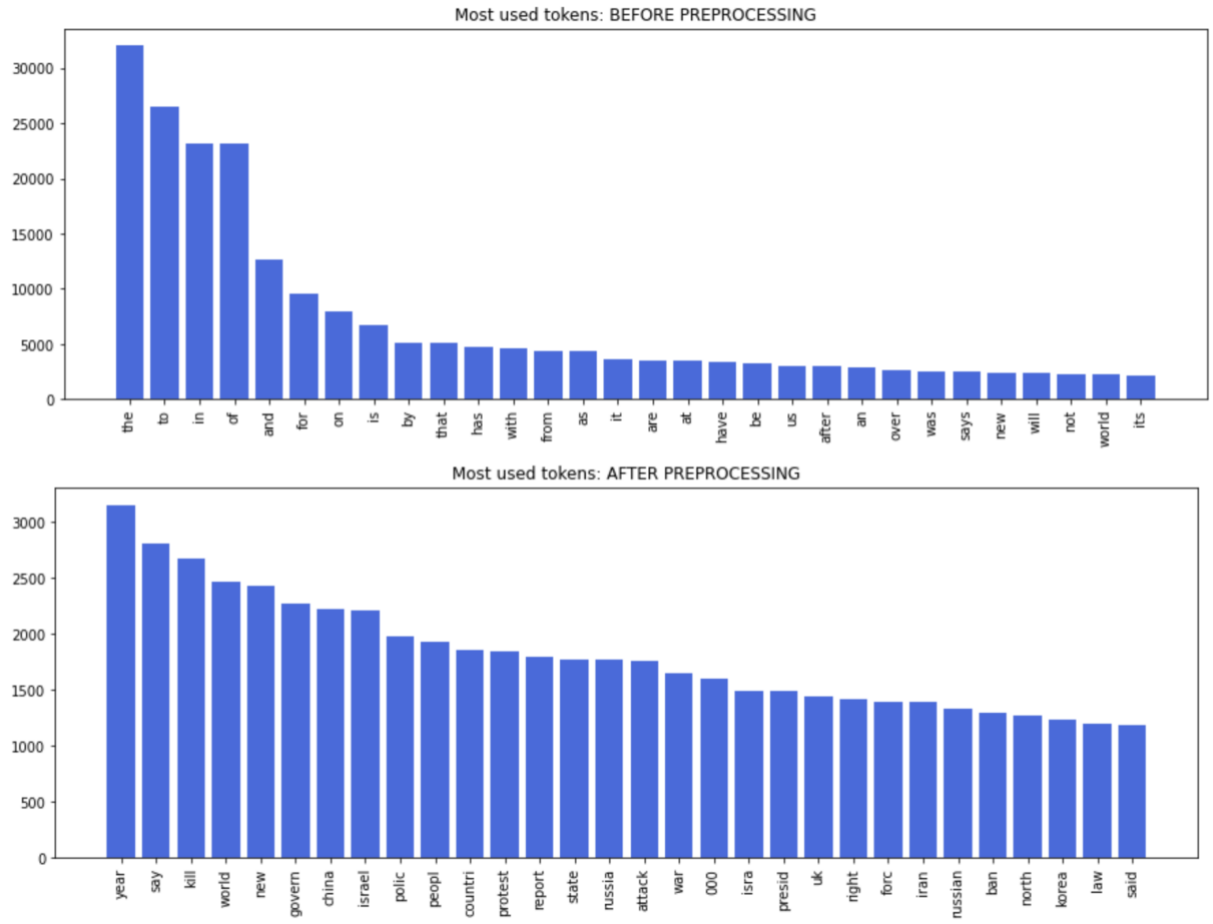


Рисунок 8

[Источник: Приложение 1]

соответствующего слова в этот текст. В качестве метода реализации можно воспользоваться *CountVectorizer* [23]. Одна из проблем, связанных с таким подходом, заключается в том, что *Bag-of-words* не учитывает специфику текстов во всем наборе документов. Своеобразным решением может выступать использование альтернативного метода *TF-IDF* для векторизации. Идея метода заключается в том, чтобы для каждого текста рассчитывать два показателя: *Term Frequency (TF)* — отношение количества вхождений данного токена в текст ко всей длине текста, и *Inverse Document Frequency (IDF)* — логарифм от числа, обратного отношению числа текстов, содержащих этот токен, к общему числу текстов. Произведение этих двух показателей для конкретного токена и представляет собой его *TF-IDF* значение в рамках данного текста. Подобная идея позволяет, в некоторой степени, нормировать

тексты на общую направленность новостного потока для выделения наиболее выдающихся экземпляров. Очевидно, эти методы не представляют собой state-of-the-art практик в обработке естественного языка, однако их интуитивность и интерпретируемость делают их вполне подходящими опциями для построения бейзлайна, в то время как использование более сложных эмбеддингов (векторных представлений), их сущность и приложение к поставленной задаче, будут описаны позднее, в экспериментальной части исследования.

В добавок к техникам векторизации текстов, закономерен и немного иной подход к получению информации об эмоциональном окрасе новости. Исследования свидетельствуют о том, что инвесторы так или иначе склонны сильно реагировать на пессимизм в средствах массовой информации, измеряемый количеством и долей негативных слов в новостях [17, стр. 30-31]. Негативными словами будем называть слова, традиционно имеющие пессимистичный в человеческом понимании семантический окрас. Пессимизм, транслируемый средствами массовой информации в новостях, традиционно провоцирует падения на финансовом рынке. В добавок к этому, отклонения от среднего уровня негативности слов в СМИ связаны с ростом объемов торгов, что, в свою очередь, объяснимо реакцией инвесторов на изменение настроения, происходящее на рынке [17, стр. 30-31]. Валидным показателем уровня пессимизма в новостях может выступать индекс негативных слов для текста, рассчитываемый как отношение количеств слов с негативным окрасом в документе к общему количеству слов в документе [13, стр. 2]. Средний помесячный уровень этого показателя, в целом, достаточно стабилен на уровне, близком к 0.09 пессимистичных слова в новостном заголовке [Рисунок 9]. Однако, подобный показатель требует некоторого рода осторожности при расчете. Потенциальная ловушка может заключаться в том, что одни и те же слова в повседневной речи и в финансовом или новостном жаргоне могут нести разную эмоциональную нагрузку. Так, например,

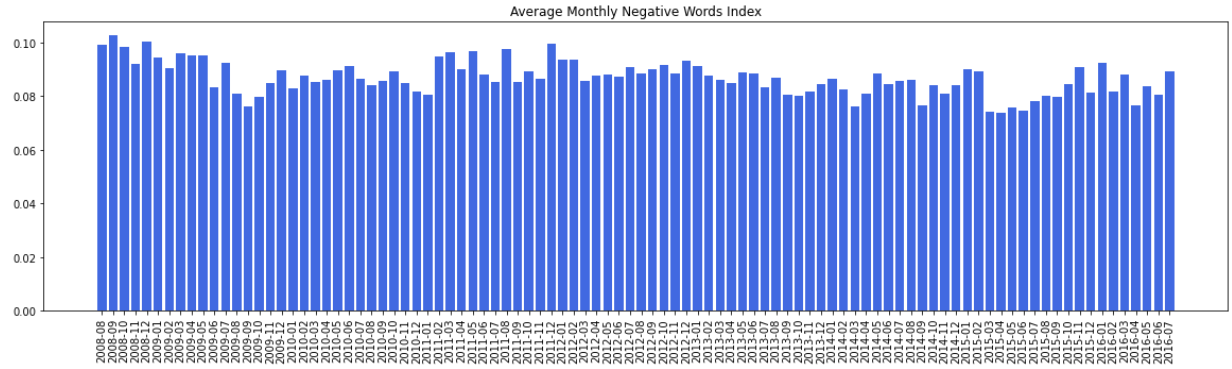


Рисунок 9

[Источник: Приложение 1]

нейтральные в финансовом контексте слова «liability» (обязательство), «foreign» (иностранный) или «depreciation» (амортизация) могут быть классифицированы как пессимистичные [11, стр. 61-62]. Эта проблема имеет несколько решений, наиболее трудоемкое из которых заключается в самостоятельной разметке негативности слов в зависимости от решаемой задачи. Другое же решение основывается на концепте, уже описанном выше: процесс расчета индекса негативных слов можно усовершенствовать, нормируя слова на их частоту относительно всего корпуса текстов (идея, аналогичная TF-IDF) [11, стр. 61-62]. Таким образом, рассчитанный для каждого текста индекс негативных слов можно использовать для обогащения набора признаков в качестве некоторого индикатора негативного настроения в информационном поле.

В конечном итоге, непосредственно перед построением модели мы располагаем векторными представлениями конкатенаций заголовков новостей, сгруппированных по дням, индексом негативных слов, рассчитанным на основе этих конкатенаций и набором технических показателей и характеристик финансового рынка в моменте времени. Комбинация этих данных будет использована с целью построения модели для достижения наилучшего результата относительно выбранных метрик.

Построение моделей

Метрики качества модели

Построение моделей машинного обучения и их интерпретация невозможны без выбора метрик качества для оценки точности этих моделей и сравнения их между собой. В данном случае, стоит обратить внимание на то, что решаемая проблема представляет из себя задачу бинарной классификации: алгоритм должен предсказывать, вырастет индекс по результатам торговой сессии или нет. Наиболее интуитивная из метрик, *accuracy*, будет использована в качестве измерителя точности модели: какую долю (значение лежит между 0 и 1, включительно) объектов модель классифицировала верно. Поскольку баланс классов в данных естественным образом достаточно близок к равномерному, метрика не должна оказаться смещенной [Рисунок 2]. Однако, единственного показателя недостаточно для оценки качества предсказания в разных

confusion matrix / confusion table	Положительное предсказание	Отрицательное предсказание
Положительный класс	True Positive (TP)	False Negative (FN)
Отрицательный класс	False Positive (FP)	True Negative (TN)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive},\ Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(Precision + \beta^2 Recall)}$$

Рисунок 10

[Источник: 6, стр. 346-347]

категориях. Используем матрицу ошибок (*confusion matrix* или *confusion table*) и производные из нее метрики: в том числе, *precision, recall, F-score* [Рисунок 10]. Матрица ошибок ставит предсказанным значениям в задаче классификации в соответствие истинные метки классов. На основе полученных же показателей рассчитываются значения *precision*, отвечающего за долю правильных предсказаний из всех предсказаний положительного класса, и *recall*, означающего долю правильно классифицированных объектов положительно класса. F-score, в свою очередь, поможет комбинировать эти метрики, балансируя их между собой. Помимо прочего, выбор этих метрик оправдан вероятностным подходом к решению задачи [6, стр. 347]. На основании значений *precision* и *recall* для разных трешхолдов вероятности (отметок вероятности, значения выше которых будут определены как положительный класс, а ниже — отрицательный соответственно; в классической модели, это значение равно 0.5), можно построить общепринятую Precision-Recall Curve (PR кривая), площадь под которой так же может быть использована в качестве метрики точности модели. Однако, поскольку сами *precision* и *recall* уже используются, будем строить ROC-кривую и измерять площадь под ней, соответственно, стремясь максимизировать ROC-AUC (Receiver Operating Characteristic Area Under Curve). ROC-кривая строится по аналогичному PR кривой алгоритму, только в качестве осей для нее используются значения *True Positive Rate* и *False Positive Rate*, тоже рассчитываемые на основе матрицы ошибок.

Baseline-модель

Итак, подготовив данные и определив метрики качества модели, пришло время перейти непосредственно к ее построению. Модель на первом этапе будет максимально упрощена для получения базовой версии, которую можно итеративно обновлять и совершенствовать. Бейзлайн должен быть основой

исследования и с точки зрения метрик качества, поскольку последующие итерации модели будут конкурировать, в первую очередь, с его результатами. Концептуально, модель должна совмещать непосредственно информацию из текстов заголовков новостей (выраженную в виде эмбеддингов) и остальные числовые признаки (в том числе, и показатели, рассчитанные на основе самих текстов: индекс негативных новостей). Для достижение подобного результата можно использовать несколько разных подходов. С одной стороны, эмбеддинги и числовые показатели можно объединить конкатенацией в общий признаковый вектор для каждого объекта, но такой подход может нарушать предпосылки использования некоторых моделей (например, рекуррентных нейронных сетей, которые принимают на вход данные в виде именно упорядоченной последовательности). К тому же, перенос TF-IDF вектора для каждого объекта, используемого в бейзлайне исследования, из разреженной матрицы в обыкновенную с целью объединения с финансовыми показателями неэффективен вычислительно с точки зрения использования памяти. Другое решение заключается в построении двух моделей и использовании их предсказаний в качестве скрытого слоя для модели верхнего уровня. В добавок, подобный ансамбль моделей уже встречался в статьях прошлых лет [18, стр. 2]. Разделив полученные данные на непересекающиеся выборки для обучения (тренировки) и валидации, соответственно, векторизуем конкатенацию текстов с помощью TF-IDF алгоритма, обученного на корпусе тренировочных документов. Финансовые же показатели отмасштабируем с помощью нормализации, посредством вычитания выборочного среднего и деления на стандартное отклонение. Подобная мера необходима для того, чтобы привести разные в масштабах данные (например, объемы торгов и скользящий индекс роста целевой переменной) к одному порядку. Важно упомянуть о том, что, поскольку финансовые признаки содержат информацию о хронологически предшествующих значениях целевой переменной, деление на тренировочную и валидационную выборку должно происходить по временному трешхолду. Таким образом, модель будет защищена от ситуации,

в которой обучение происходит на данных «из будущего», что безусловно нарушало бы логические предпосылки применения модели.

На основе каждой из подвыборок обучим логистическую регрессию, и, используя предсказания вероятностей этих моделей для таргета на объектах в качестве скрытых переменных, построим еще одну логистическую регрессию с теми же таргетами. Таким образом, финальная модель представляет из себя ансамбль из двух логистических регрессий, выходы которых взвешены для получения финальной оценки. Выбор логистических регрессий в качестве элементарных моделей, в свою очередь, обусловлен их интерпретируемостью и непосредственно спецификой бейзлайна. Полученная модель содержит ряд гиперпараметров (фиксируемых предварительно, до начала процесса обучения модели, в отличие от обучаемых параметров: например, весов в линейной модели), которые нуждаются в тщательном подборе с целью оптимизации. В их числе, количество используемых заголовков новостей в конкатенации (сами новости упорядочены по популярности), способ препроцессинга токенов (стемминг, лемматизация, отсутствие препроцессинга), использование словаря стоп-слов, диапазон размера n-грамм при векторизации, максимальное количество слов в словаре или минимальный *document frequency* для TF-IDF, *learning rate* логистической регрессии и другие. Перебор их будет осуществлен с помощью поиска по сетке (*Grid Search*): алгоритма циклического перебора всех возможных комбинаций гиперпараметров. Таким образом, итеративно рассчитывая метрики качества модели при использовании того или иного набора гиперпараметров, решается задача выбора оптимального. В данном случае, Наилучшие значения точности предсказания, ROC-AUC, precision и recall достигаются при использовании всех 25-ти заголовков, собранных в датасете, биграмм в комбинации с отсутствием дополнительной обработки токенов и очищения от стоп-слов, а так же ограничении максимального размера словаря тридцатью тысячами токенов [Рисунок 11]. Такой выбор гиперпараметров демонстрирует accuracy

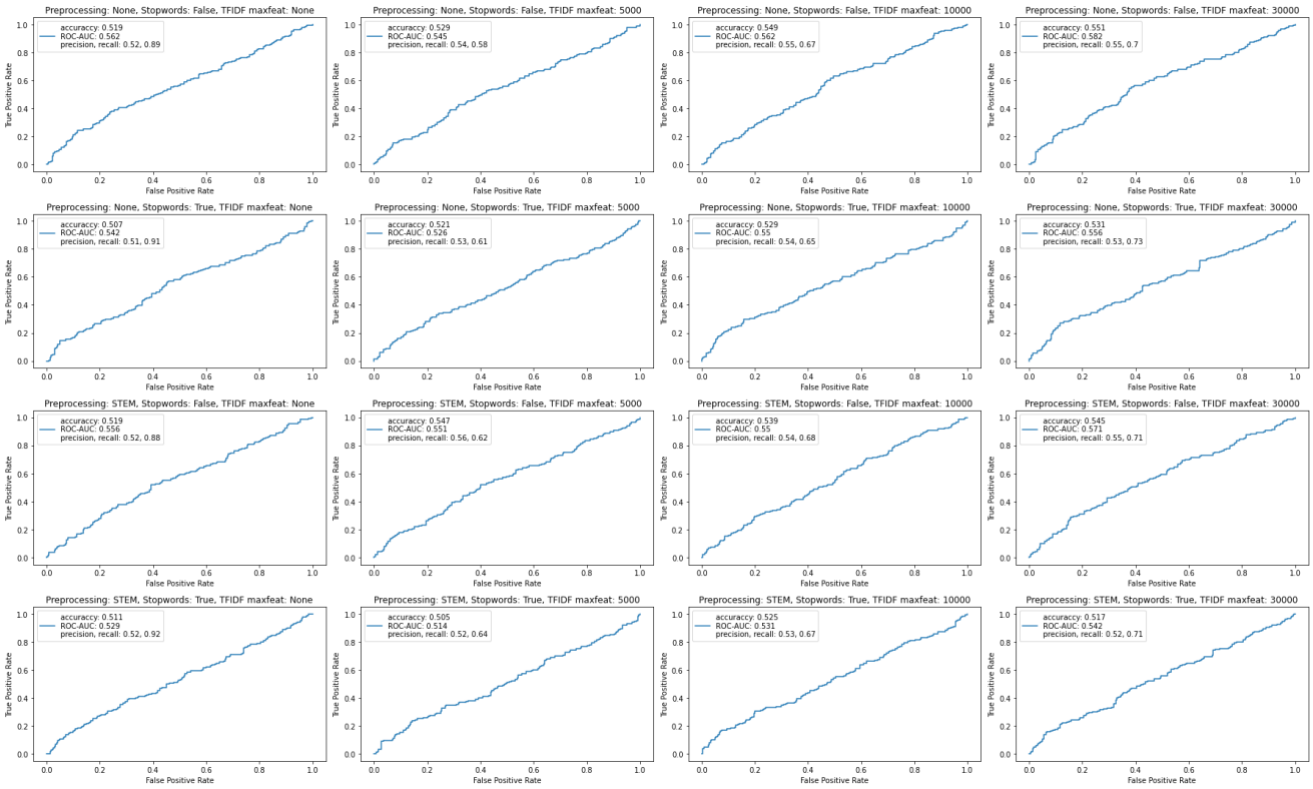


Рисунок 11

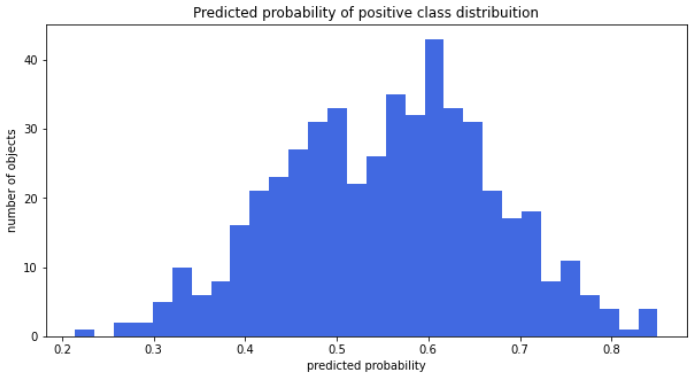
Перебор гиперпараметров: ROC-кривая и основные метрики

[Источник: Приложение 1]

на уровне 0.55 и, что принципиально, ROC-AUC порядка 0.58, в том числе несколько жертвуя precision-ом в пользу более высокого значения recall-а за счет более высокого среднего предсказанных вероятностей положительного класса. Именно эта конфигурация гиперпараметров и будет использована в качестве основной для построенного бейзлайна.

Обучив модель, используя подобранный набор гиперпараметров, имеет смысл рассмотреть распределение предсказанной вероятности отношения объектов валидационной выборки к положительному классу (обозначающему рост финансового индекса в соответствующий день) [Рисунок 12]. Несмотря на асимптотическую близость выборочного распределения к нормальному, его среднее смещено от 0.5 в сторону 0.55. С этой точки зрения, сдвиг трешхолда

для определения положительного класса к значению выборочного среднего может привести к повышению разделяющей способности модели. В действительности, такое изменение позволяет повысить accuracy до уровня 0.57, приводя, в добавок, к большему балансу между recall метрикой положительного и отрицательного классов — закономерное изменение, учитывая увеличение количества предсказаний



Classification report with probability threshold at 0.5:

	precision	recall	f1-score	support
0	0.55	0.39	0.46	240
1	0.55	0.70	0.62	257
accuracy			0.55	497
macro avg	0.55	0.55	0.54	497
weighted avg	0.55	0.55	0.54	497

Classification report with probability threshold at distribution mean:

	precision	recall	f1-score	support
0	0.56	0.50	0.53	240
1	0.57	0.63	0.60	257
accuracy			0.57	497
macro avg	0.56	0.56	0.56	497
weighted avg	0.56	0.57	0.56	497

Рисунок 12

[Источник: Приложение 1]

отрицательного класса в связи со сдвигом трешхолда вероятности [Рисунок 12]. В сущности, подобный результат успешно противостоит бейзлайнам схожих исследований, точность алгоритмов в которых колеблется в пределах от 0.56 до 0.65, где последнее значение достигается с помощью конволюционных нейронных сетей, надстроенных над сложной предобработкой текстов [3, стр. 2330]. Бейзлайн исследования демонстрирует крепкий результат, используя меньше ресурсов, затрачивая меньше ресурсов и требуя меньше данных для обучения.

Одним из основных преимуществ векторизации текстов с помощью TF-IDF в комбинации с использованием линейной модели, в особенности, в контексте выбранной задачи, представляется интерпретируемость этого метода. С этой точки зрения, веса в логистической регрессии, привязанные к токенам в словаре, отражают влияние того или иного токена на финальное предсказание

модели. Поскольку значения TF-IDF вектора, по определению, неотрицательны, знаки коэффициентов регрессии можно рассматривать как индикатор влияния конкретного слова/n-граммы на изменение предсказания: отрицательный коэффициент приводит к снижению вероятности классификации объекта как положительного, положительный, наоборот, — к росту этой вероятности. При этом, большее абсолютное значение веса соответствует большему влиянию признака на таргет. Анализ наиболее важных n-грамм с точки зрения как положительного, так и отрицательного эффекта освещает некоторые справедливые закономерности [Рисунок 13]. Так, например, фразы, несущие очевидно отрицательную коннотацию (имена террористов, «phone hacking», «sexual abuse», «Haiti earthquake», «nuclear weapons») находятся в начале списка наиболее отрицательно определенных, в

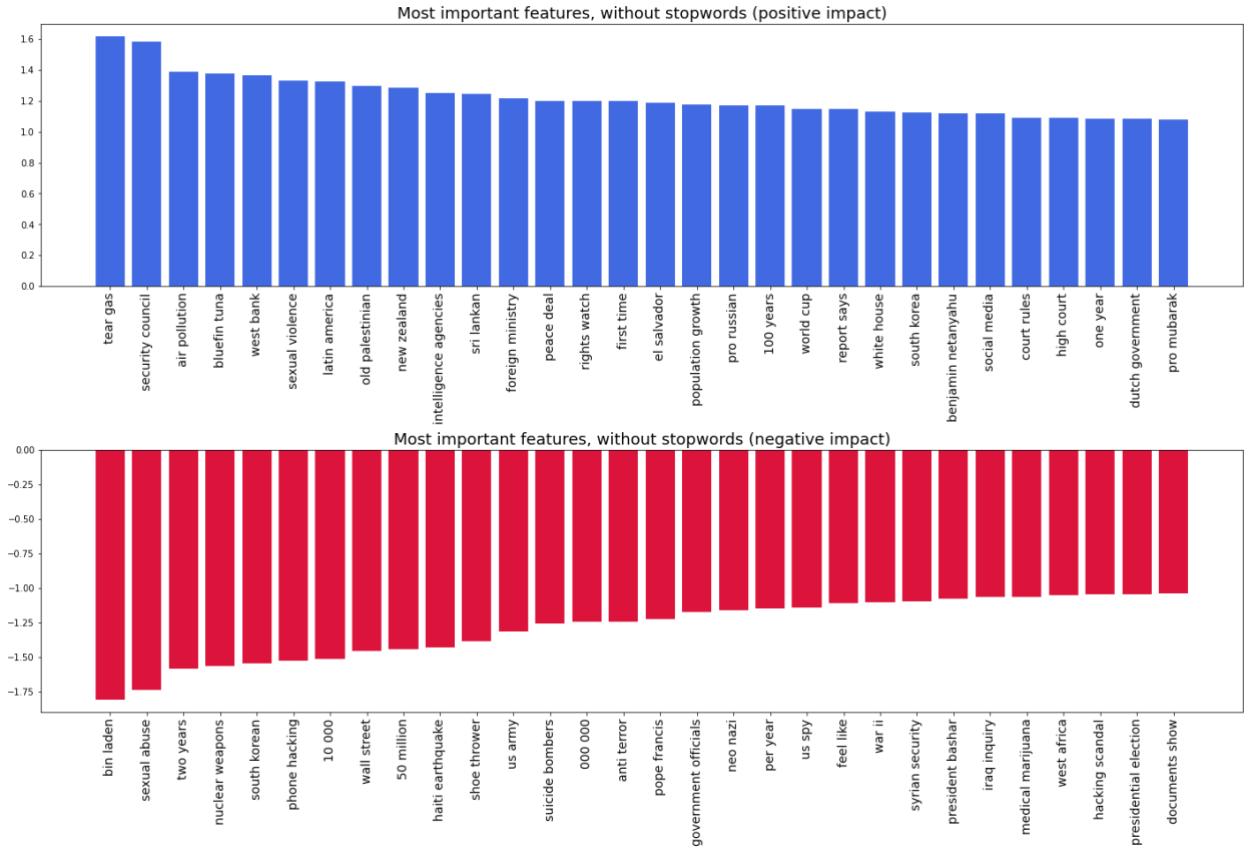


Рисунок 13

[Источник: Приложение 1]

то время как среди положительных признаков наиболее сильное влияние оказывают, в основном, нейтральные слова и выражения, связанные с глобализацией (названия стран, «world cup», «social media»). Однако, как ни странно, в последнем списке содержится небольшое количество фраз негативной семантики — несоответствие, связанное со спецификами TF-IDF векторизации и обучения модели. Правда, часть из них может быть аргументирована при более глубоком изучении. При исследовании примеров использования, например, биграммы «tear gas», просматривается достаточно стабильная тенденция использования фразы в контексте применения слезоточивого газа полицией, приводящего к прекращению беспорядков и нормализации экономической деятельности. В таком контексте выражение более не вызывает вопросов относительно положительности своего влияния на движение финансового индекса. В то же время, среди наиболее влиятельных биграмм встречаются и более абстрактные (эти биграммы были исключены из представленных диаграмм с целью получения как можно большего смыслового контекста из результатов): «and other», «this is», «if he». Подобный результат может быть обусловлен как ложными корреляциями, так и потенциальными скрытыми зависимостями среди токенов, которые невозможно уловить, используя выбранную предобработку и эмбеддинг. Избавлению от подобных токенов в списке наиболее значимых может поспособствовать использование словаря стоп-слов для очистки текстов, однако проведенный подбор гиперпараметров не подтвердил эффективность его использования в такой конфигурации.

Совершенствование векторных представлений

Использование TF-IDF векторизации позволило получить неплохое качество в совокупности с большим количеством информации о зависимостях между упоминаемыми в заголовках новостей словами и движением индекс Dow Jones

Industrial Average. Однако эмбеддинги, основанные на количественных статистиках токенов (такие как Bag-of-words и TF-IDF), несут в себе исключительно информацию о количественном наполнении текста словами в сравнении с остальными текстами корпуса, игнорируя при этом отношения между токенами и абсолютные значения самих слов. Для решения этой проблемы можно использовать более сложные векторные представления, обучаемые на больших корпусах текстов таким образом, чтобы вектора близких по значению слов находились близко друг к другу (с точки зрения той или иной метрики расстояния), и наоборот. Один из state-of-the-art эмбеддингов, *Word2Vec*, является, по своей сути, вероятностным подходом к векторизации слов. Модель векторизации обучается с помощью техник *Skip-gram* и *Continuous Bag of Words*, решая задачу предсказания «соседей» слова (окружающих его слов) на основе его самого и, наоборот, предсказания слова на основе его «соседей» (окружающих его слов), соответственно. В результате чего, эмбеддинги слов в модели Word2Vec сохраняют некоторые особенности взаимодействия слов, перенося их в векторное пространство. Классическим примером такой преемственности является тот факт, что слово «man» («мужчина») относится к слову «woman» («женщина») так же, как слово «king» («король») относится к слову «queen» («королева») с точки зрения векторных представлений, обученных с помощью Word2Vec, сохраняя логическую связь [5, стр. 12321]. Таким образом, именно вектор слова «queen» должен располагаться ближе всего к вектору, полученному при вычитании из вектора «king» вектора «man» и прибавлении «woman». Близость векторов, в свою очередь, оценивается косинусным расстоянием: $cos(\vec{a}, \vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{||\vec{a}|| \cdot ||\vec{b}||}$ [2, стр. 157]. Так, например, слова «bull» и «stock», которые могут быть связаны в финансовом контексте, имеют косинусное расстояние порядка 0.278, в то время как значительно менее контекстуально близкие слова «peace» и «zebra» — 0.059, соответственно [Приложение 1]. Об эффективности метода свидетельствует и его популярность в сообществе: в открытом доступе

доступно большое количество словарей с векторами слов, предварительно обученных с помощью Word2Vec в различных конфигурациях. Исходя из предпосылок исследования, используем вектора размерности 300 из модели «*word2vec-google-news-300*», предварительно обученной на новостях из Google (соответствует специфике задачи) и доступной в библиотеке genism [24]. Для получения векторного представления всего текста используем усредненный вектор для всех слов в предложении.

Выборка, построенная с помощью уже описанной предобработки (токенизации, лемматизации и стоп-слов) и Word2Vec эмбеддингов, содержит 300 признаков для каждого объекта (усредненное векторное представление текста по словам). Потенциальные результаты весов модели после ее обучения более не так хорошо интерпретируемы, как в случае с TF-IDF, поэтому выбор самой модели тоже необходимо изменить: линейная модель может с трудом улавливать закономерности подобного рода данных. В качестве основного алгоритма будет использован градиентный бустинг из библиотеки catboost [20]. Концептуально, выбранная архитектура состоит из большого количества классификаторов (в случае градиентного бустинга из catboost, решающих деревьев), построенных последовательно, «поверх друг друга», для получения финальной оценки. Обучается же алгоритм таким образом, чтобы на каждом следующем шаге минимизировать ошибку предыдущих элементарных моделей.

Для обучения градиентного бустинга может понадобиться большее количество данных, поэтому воспользуемся искусственным расширением датасета. Для каждого таргет-дня будем семплировать пять объектов по пять заголовков новостей без пересечений. Несмотря на то, что новости внутри одного дня не дублируются и переобучения модели не должно произойти, по-прежнему, необходимо убедиться в том, что предсказания при обучении и тестировании совершаются исключительно на исторических данных

относительно дня предсказываемого таргета. Для этого, разбиение на обучающую и валидационную подвыборки будет осуществляться, как и ранее, исключительно с помощью хронологического отсечения. Обучив модель на усредненных Word2Vec эмбеддингах для расширенного набора данных и подобрав оптимальные гиперпараметры, удается достичь значения метрики ROC-AUC порядка 0.58 и accuracy на уровне 0.56 [Рисунок 14]. Как видим,

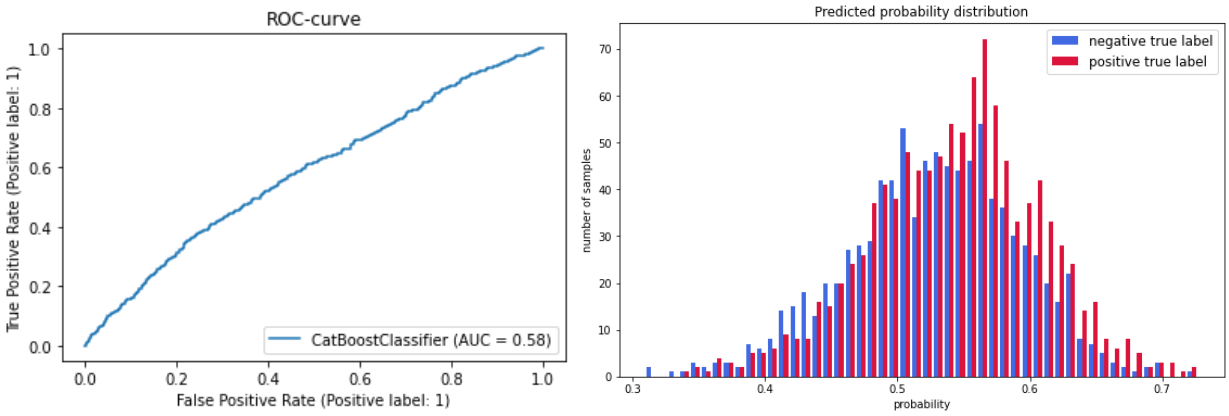


Рисунок 14

[Источник: Приложение 1]

небольшое улучшение в терминах ROC-AUC достижимо, однако оно влечет за собой жертву точности модели. Более подробный взгляд на распределение предсказанных вероятностей классов дает немного больше понимания относительно качества предсказания [Рисунок 14]. Несмотря на то, что среди предсказаний положительного класса (тех объектов, для которых вероятность положительно класса превышает 0.5) наблюдается существенное превосходство истинного класса, действительно четкого разделения распределений не достигается. Визуально значительно меньшее количество объектов истинно положительного класса в отрицательных предсказаниях (FN), чем истинно отрицательного класса в положительных предсказаниях (FP) лишь подтверждает превосходство Recall метрики (0.75) над Precision (0.56) для данной модели.

Подобные результаты, несмотря на большой объем информации, содержащийся в Word2Vec эмбеддингах, объяснимы, поскольку, на самом деле, специфика задачи, предоставляя требования к векторным представлениям, в некоторой степени игнорирует часть этой информации. Необходимость в усреднении эмбеддингов слов внутри одного текста для получения представления всего заголовка новости размывает потенциально важные детали, содержащиеся в векторах отдельных токенов. Поскольку специфика задачи неизменна — необходимо создание инструмента для оценки тональности новости — возможным решением представляется корректировка архитектуры модели для более эффективного использования Word2Vec эмбеддингов.

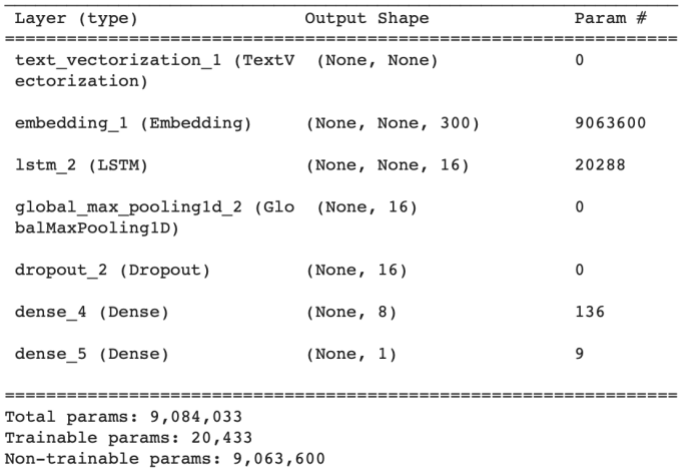
Рекуррентные нейронные сети

В попытке максимально полноценно использовать информацию, которую содержат в себе векторные представления Word2Vec, справедливо использовать модель, принимающую на вход и обрабатывающую все слова в обучающей выборке в отдельности, а не только их усредненное векторное представление. Одним из решений является архитектура *LSTM* (Long short-term memory), представляющая из себя вариацию рекуррентной нейронной сети, передающую между вхождениями не только скрытые состояния, но, что важно для борьбы с проблемой затухания градиента при обучении посредством техники обратного распространения ошибки в нейронной сети, cell state [18, стр. 3]. Это состояние, проходящее через наименьшее количество преобразований из всех, призвано отвечать за сохранение максимального объема информации при переходе от одного вхождения (эмбеддинга) к другому в процессе обучения. LSTM, как и все рекуррентные нейронные сети, получает на вход последовательность элементов (в случае этого исследования,

последовательность векторов, репрезентирующих токены), реплицируя то, как тексты воспринимаются человеком — в виде последовательности слов. Эта особенность архитектуры делает ее особенно предпочтительной в связи с ее совместимостью с предпосылками исследования.

Помимо непосредственно эмбеддингов и LSTM-блока в финальную архитектуру добавлены *Max Pooling* слой, уменьшающий размерность данных после прохода эмбеддингов через рекуррентную сеть за счет подсчета максимумов из подвыборок значений в скрытом слое, *Dropout* слой, обнуляющий в случайном порядке часть вхождений для защиты от переобучения, и полносвязные слои для плавного перехода к единичной размерности, отвечающей за предсказываемую вероятность положительного класса [Рисунок 15]. Финальный слой использует в качестве функции активации сигмоиду для перехода в вероятностное пространство, в то время как остальные полносвязные слои — *Rectified Linear Unit (ReLU)*. Важно отметить, что комбинация Max Pooling и LSTM слоев была прежде использована в исследованиях и продемонстрировала высокое качество при работе с текстовыми данными [18, стр. 2].

Тем не менее, в результате обучения описанной модели, основанной на LSTM архитектуре, не удалось достичь показателей accuracy, F-score и ROC- AUC, превышающих те, что были получены на предыдущих стадиях (при использовании линейной модели и градиентного бустинга). Полученные значения accuracy и ROC-AUC находятся в окрестности 0.55 и 0.54, соответственно. Ротация в использовании эмбеддингов также не приводит к росту качества предсказания нейронной сети: одной из возникших гипотез было существование потенциальной проблемы в качестве векторных представлений Word2Vec. Однако обучение LSTM-архитектуры на эмбеддингах FinBERT привело к получению схожих значений метрик качества, при этом повысив вычислительную сложность алгоритма. FinBERT



Layer (type)	Output Shape	Param #
text_vectorization_1 (TextVectorization)	(None, None)	0
embedding_1 (Embedding)	(None, None, 300)	9063600
lstm_2 (LSTM)	(None, None, 16)	20288
global_max_pooling1d_2 (GlobalMaxPooling1D)	(None, 16)	0
dropout_2 (Dropout)	(None, 16)	0
dense_4 (Dense)	(None, 8)	136
dense_5 (Dense)	(None, 1)	9

Total params: 9,084,033
Trainable params: 20,433
Non-trainable params: 9,063,600

Рисунок 15

[Источник: Приложение 1]

представляет из себя результат обучения state-of-the-art модели *BERT (Bidirectional Encoder Representations from Transformers)* на данных финансовых текстов для задач sentiment-анализа (анализа «настроения», эмоционального окраса текстов) [1, стр. 9]. Сама архитектура BERT основана на языковой модели Transformer, использующей технику self-attention как в encoder, так и в decoder части, для наиболее эффективной обработки последовательностей [19, стр. 2-3]. Таким образом, эмбеддинги FinBERT должны быть более чувствительны к информации, содержащейся в финансовых текстах, точнее перенося ее в векторное пространство, но, исходя из стабильности в результатах между Word2Vec и FinBERT представлениями, данное преимущество последних нивелируется в рамках данной задачи широтой тематического спектра новостей, в действительности влияющих на инвестиционные решения экономических агентов. Отсутствие прогресса в метриках по сравнению с предыдущими стадиями исследования, несмотря на использование best-practice эмбеддингов и модели, может иметь сразу несколько объяснений.

В первую очередь, одна из проблем при использовании предварительно обученных Word2Vec векторных представлений, вполне парадоксально, неразрывна связана с одной из их сильнейших сторон. Языковая модель, обученная на миллионах текстов, способна генерировать векторные представления для огромного количества слов, но, в то же время, сами эмбеддинги из-за этого становятся менее чувствительными к специфике текстов, чем в случае их обучения на тематически специфической выборке данных. Потенциальным решением может являться обучение собственных векторных представлений для конкретной задачи с помощью Word2Vec, однако подобный шаг требует большого объема как данных для обучения, так и вычислительных ресурсов. Именно поэтому его стоит воспринимать в качестве потенциального направления дальнейшего развития исследования. Помимо этого, проблемой при использовании подобной архитектуры для решения поставленной задачи является и структура входных данных. Ключевой предпосылкой использования LSTM блоков при обработке естественного языка является восприятие входных данных в качестве упорядоченной последовательности слов. В то же время, необходимость предсказывать изменение финансового индекса на основе всего новостного фона, а не отдельной новости, накладывает на данные условие наличия множества новостей, привязанных к одному целевому дню. Исходя из этих предпосылок (предпосылок использования модели и предпосылок исследования) и возникает потенциальное несоответствие, не позволяющее значительно повысить метрики качества за счет использования LSTM архитектуры: конкатенация нескольких заголовков новостей в течение дня (к тому же, упорядоченных по популярности) не является упорядоченным массивом в смысловом понимании, поскольку окончание одного заголовка не является логическим предшественником начала другого. В сущности, совокупность этих особенностей и является наиболее значительной причиной отсутствия существенного улучшения качества модели при переходе к

использованию рекуррентных нейронных сетей в комбинации с эмбеддингами Word2Vec.

TF-IDF и градиентный бустинг

Поскольку, по описанным причинам, использование нейронных сетей в комбинации с эмбеддингам Word2Vec не продемонстрировало результата, доминирующего над бейзлайном в разрезе всего набора метрик, справедливым шагом станет возвращение к количественным векторным представлениям. TF-IDF зарекомендовал себя в качестве надежного и прозрачного инструмента, однако применение линейной модели к полученным векторам текстов нельзя назвать лучшим из возможных решений. Имея это в виду, протестируем эмбеддинги TF-IDF совместно с градиентным бустингом. В данном случае, так же используем для обучения модель из библиотеки catboost. Идейно, алгоритм бустинга, в противовес логистической регрессии, способен значительно лучше справляться с моделированием нелинейных зависимостей, которые, в том числе, могут иметь место в случае с векторными представлениями текстов. Сам процесс обучения имеет отдаленно схожие концептуальные черты для двух моделей, однако, линейная модель использует градиентный спуск в пространстве весов для нахождения оптимальных значений, в то время как градиентный бустинг тоже совершает градиентный спуск, но в пространстве алгоритмов (функций от признаков). Используя данную интуицию и применяя подбор гиперпараметров модели, среди которых, в том числе, размер словаря, количество элементарных моделей, максимальная глубина деревьев, скорость обучения и другие, построим наиболее сильную с точки зрения предсказательной силы модель. Построенная модель демонстрирует точность на валидационной выборке на уровне 0.55, при этом обновляя лучшее значение метрики ROC-AUC до 0.593 [Рисунок 16]. Однако более подробный взгляд на распределение предсказанных вероятностей для объектов

положительного и отрицательного классов демонстрирует два близких к нормальному распределения со средними в 0.54 и 0.57, соответственно. Отсюда, возникает гипотеза о том, что разделение предсказаний на основании трешхолда вероятности, отличного от 0.5 и находящегося в интервале от 0.54 до 0.57 позволит достичь большей точности. Действительно, при использовании среднего предсказания на обучающей выборке в качестве разделителя, удается повысить accuracy модели до 0.57 [Рисунок 16]. Подобные значения метрик качества успешно конкурируют с результатами предыдущих исследований: accuracy порядка 0.56-0.59 в задаче предсказания

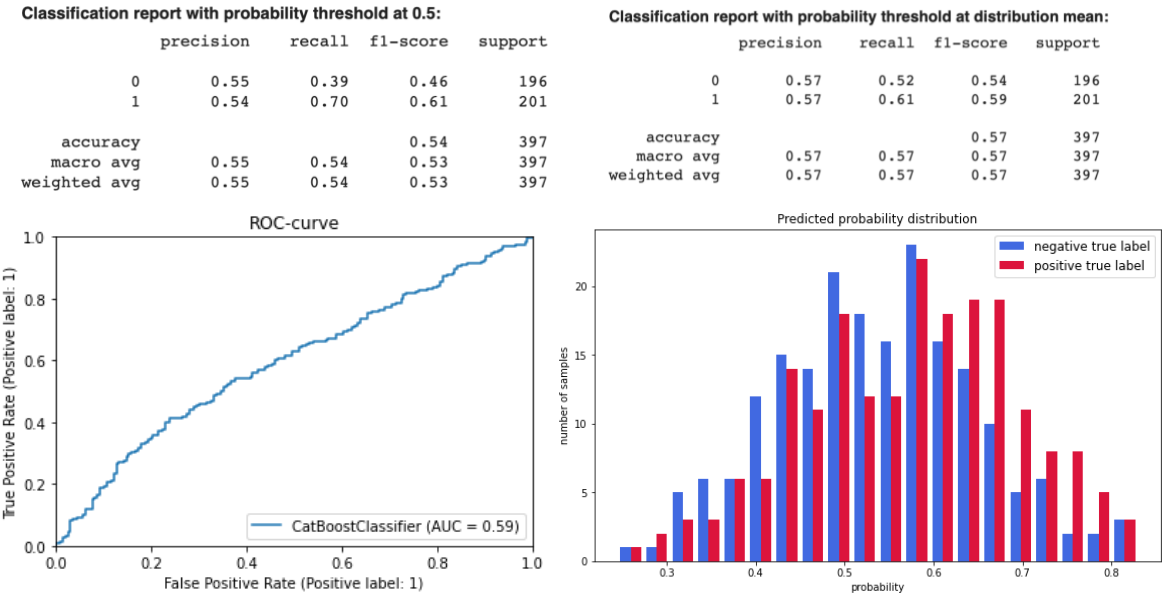


Рисунок 16

[Источник: Приложение 1]

колебаний индекса S&P 500 на основе новостей о событиях является надежным бенчмарком, к которому полученной модели удалось приблизиться [4, стр. 1420-1422]. Таким образом, используя TF-IDF для векторизации текстов, но применяя к полученным эмбеддингам более сильную модель, нежели логистическая регрессия, удается достичь немного более высоких значений основных метрик.

Аналогично изучению весов обученной линейной модели, важность компонент TF-IDF вектора может быть интерпретирована и в градиентном бустинге. Однако, поскольку градиентный бустинг не ставит в соответствие каждому признаку определенный вес, настраивая его в процессе обучения, а делает предсказания с помощью композиции решающих деревьев, выбрать единый параметр в качестве оценки важности признака для градиентного бустинга не удастся. Тем не менее, существуют иные способы, позволяющие определить влияние значений каждого признака на выход модели: одним из наиболее удобных является *SHAP (SHapley Additive exPlanations)*, использующий в качестве показателя важности значения вектора Шепли (Shapley values) для условного математического ожидания исходной модели [12, стр. 4-5]. В добавок к эффективному инструменту интерпретации важности признаков в построенной модели, библиотека shap так же предоставляет отличные возможности для визуализации полученных оценок [26]. Изучение наиболее важных n-грамм в разрезе двух моделей: обученной

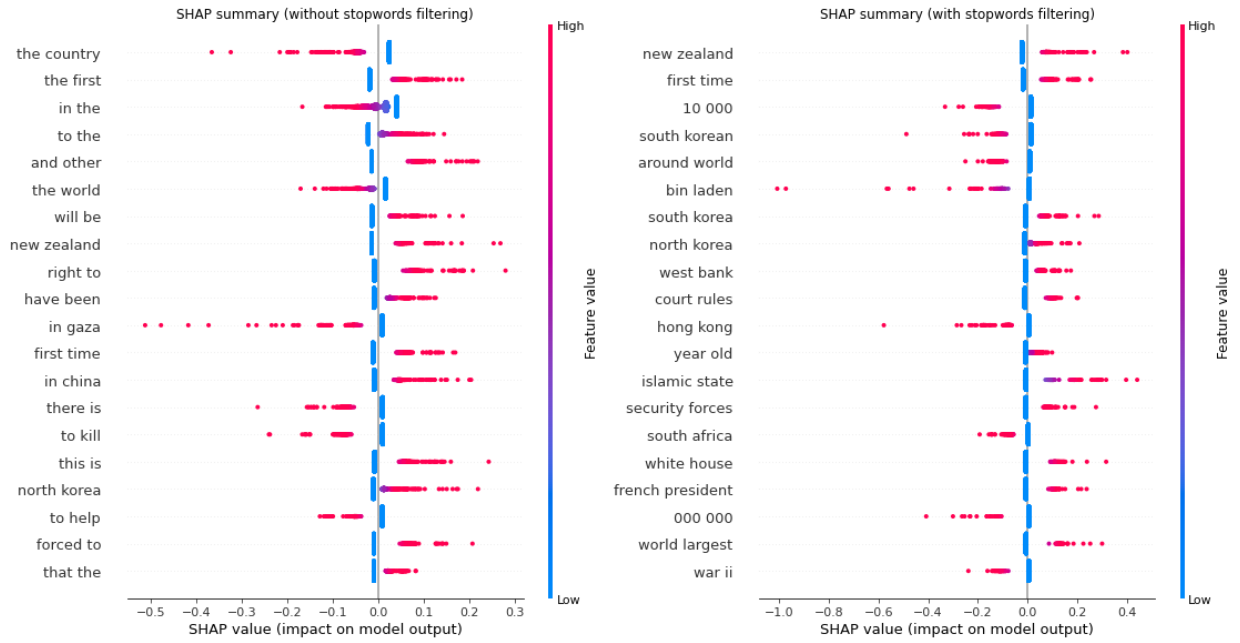


Рисунок 17

[Источник: Приложение 1]

без фильтрации стоп-слов в текстах (слева) и обученной с фильтрацией (справа), — демонстрирует множество пересечений с показателями важности признаков при использовании линейной модели (например, обе модели положительно чувствительны к появлению биграмм «new zealand»), но также и открывает новые закономерности [Рисунок 17]. В частности, подчеркивает принципиальные различия, возникающие при очистке от стоп-слов: большая часть наиболее важных биграмм в левой части графика состоит именно из комбинации слов с союзами, артиклями и предлогами («in china», «the world», «the country»). К тому же, погружение в SHAP демонстрирует важность фактора упоминания национальных элементов в новостях на реакцию экономических агентов (большое количество названий стран, языков, национальностей в списке наиболее важных признаков). Упоминание государств в заголовках новостей, по всей видимости, создает сильные ассоциативные сигналы для инвесторов и, пусть непосредственно названия стран не носят эмоционального окраса, сам факт их появления в новостном потоке и контекст этого появления в значительной степени оказывают влияние на ожидания и поведение агентов. В этом ключе, одной из возможностей для продолжения исследования является, в том числе, фокус на исследование имен собственных в текстах при создании векторных представлений новостей.

Target encoding токенов

Изучив эффективность использования количественных методов векторизации, подобных Bag-of-words и TF-IDF, в экспериментальной части исследования разумно протестировать использование кодировки, более привязанной к целевой переменной. Создадим дамми-переменные в виде флага присутствия токена в тексте для нескольких наиболее популярных токенов из корпуса документов (это количество новых дамми-переменных будем подбирать во время общего подбора гиперпараметров) и, используя

технику *target encoding*, рассчитаем статистики для каждой из них. В сущности, флаги вхождения наиболее частых токенов будут заменены на средние значения целевой переменной для объектов обучающей выборки, в зависимости от значения флага. Объединив финансовые технические показатели, индекс негативных слов для конкатенации текстов за день, предшествующий целевому, и target encoding 500 наиболее часто упоминаемых токенов (без учета стоп-слов), удается, в результате тщательного подбора гиперпараметров, достичь значений как accuracy, так и ROC-AUC на уровне 0.54 [Рисунок 18]. К тому же, выбранный подход открывает возможность изучения наиболее важных признаков для принятия решения алгоритмом градиентного бустинга. Вполне закономерно, дневные (*Close_rate2*), как и месячные (*Close_rate30*) темпы роста целевого финансового индекса, совместно с некоторыми другими техническими показателями (в особенности, значениями волатильности и статистиками,

Classification report with probability threshold at 0.5:

	precision	recall	f1-score	support
0	0.52	0.77	0.62	196
1	0.58	0.31	0.40	201
accuracy			0.54	397
macro avg	0.55	0.54	0.51	397
weighted avg	0.55	0.54	0.51	397

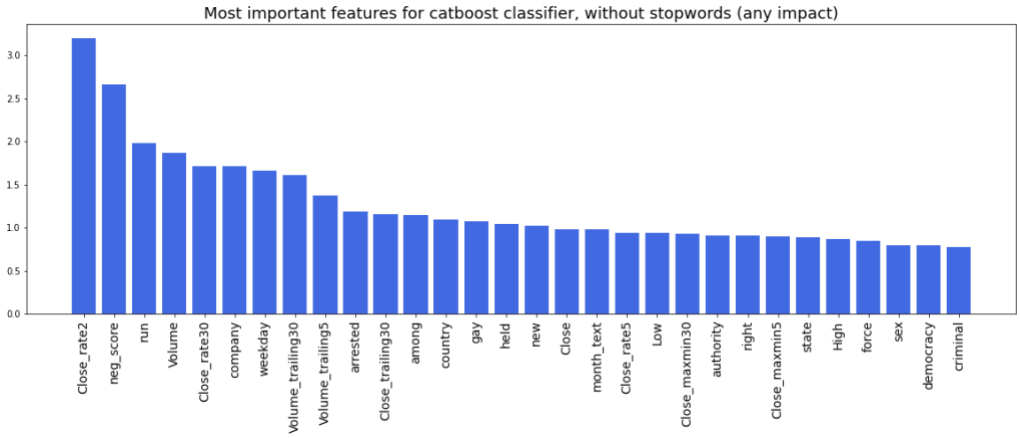


Рисунок 18

[Источник: Приложение 1]

основанными на исторических значениях цен индекса), обладают существенной предсказательной способностью относительно будущих значений Dow Jones Industrial Average [Рисунок 18]. Индекс негативных слов, в свою очередь, так же в большой степени важен для предсказания изменения индекса. Однако более любопытная эмпирическая деталь заключается в важности именно «таргет-энкодингов» для модели: наличие или, наоборот, отсутствие слов «run», «company», «democracy», «arrested» и некоторых других демонстрирует определенную разделительную силу при использовании статистик, основанных на целевой переменной. Набор слов, не встречавшийся в списке наиболее важных при использовании TF-IDF векторизации. Таким образом, несмотря на то что использование метода target encoding не приводит к улучшению метрик качества модели, оно, несомненно, демонстрирует способность раскрывать скрытые прежде, неочевидные зависимости между использованием слов в заголовках новостей и изменение финансового индекса [Приложение 3]. В связи с этим, использование этого подхода, в качестве, в первую очередь, не метода для построения наилучшей предсказательной модели, а экспериментального инструмента анализа закономерностей, важного в рамках исследования, более чем оправдано.

Вывод и дальнейшее развитие исследования

Итак, полученные в ходе обширного исследования результаты, в совокупности с примененными при их получении методами и теоретическими предпосылками, позволяют сформулировать некоторые заключения относительно оригинальных целей работы. В первую очередь, основная гипотеза исследования подтверждена: предсказание средней реакции экономических агентов на новостной фон и, соответственно, следующего за ней, согласно гипотезе эффективного рынка, изменения цены индекса Dow Jones Industrial Average, представляется возможным. Более того, использование методов машинного обучения для предсказания изменения значений финансового показателя на основе текстовых данных позволяет добиться высоких, с учетом специфики задачи и существующих исследований, метрик ROC-AUC на уровне 0.6 и точности (accuracy) порядка 0.57. Второй, но не менее важный, результат заключается в закономерностях и связях между текстами заголовков новостей и изменениями целевой переменной, открытых в процессе исследования. Использование некоторых из продемонстрированных в исследовании слов, в особенности, слов имеющих существенно отрицательный или положительный эмоциональный окрас, в текстах заголовков может являться достаточно сильным информационным сигналом, на который экономические агенты склонны реагировать изменением инвестиционного поведения: продажей или покупкой актива, соответственно. В свою очередь, индекс негативных слов, используемый исходя из тех же предпосылок, доказал свою способность аппроксимировать негативный sentiment текста, а большая важность названий государств, локаций и человеческих имен в предсказаниях построенных моделей продемонстрировала перспективу фокусировки на именах собственных в анализе текстовых данных. В-третьих, исследование подчеркнуло выводы относительно методологии решения задачи предсказания финансовых показателей на основе текстов новостей. Специфика задачи, предполагающая

использование множества новостей для создания оценки всего новостного фона, становится причиной отсутствия доминирования state-of-the-art моделей в обработке естественного языка с точки зрения основных метрик. Количественные методы векторизации текстовой информации (в первую очередь, TF-IDF), в то же время, оказываются достаточно репрезентативными, поскольку представляют из себя оценку всего текста сразу, а не функцию над эмбеддингами отдельных слов, лишь небольшая доля которых, в действительности, определяет sentiment новости. Менее традиционные способы перевода текстов в численные представления, такие как target encoding флагов наличия наиболее популярных токенов в текстах, пусть и не приводят к построению более сильной модели, но позволяют глубже погрузиться в проблему и исследовать скрытые закономерности, возникающие между словами в заголовках новостей и движением финансового индекса. Использование же биграмм в качестве единиц векторизации способствует дополнительному приросту метрик качества во всех моделях, свидетельствуя о важности восприятия словосочетаний в качестве цельных единиц при анализе текстовой информации.

Помимо прямых результатов, достигнутых в ходе исследования, важным является вклад в дальнейшее развитие решения задачи и изучение наиболее перспективных его траекторий. Обучение эмбеддингов на основе собственных текстов является одним из наиболее ресурсоемких направлений, но обладает потенциалом существенного повышения качества модели. Наиболее полный перенос относительного смысла, содержащегося в текстах, в векторное пространство создаст возможность для более точного построения модели за счет наличия большего количества скрытой информации в векторах. Другой подход к совершенствованию полученных результатов предполагает изменение дизайна выборки с целью достижения наиболее точной взаимосвязи между появлением новости и реакцией цены актива в рамках гипотезы эффективного рынка. Использование в качестве целевой переменной

не дневного изменения финансового индекса, а более краткосрочных колебаний, привязанных к точному времени публикации новостей, позволит очистить данные от лишней информации, сократив шум в модели и повысив точность. Наиболее экспериментальное направление дальнейшего исследования заключается в поиске оптимальной стратегии выбора слов из новостей перед переходом в векторное пространство. Одной из найденных в процессе исследования закономерностей является наличие значительной разницы во влиянии разных, в семантическом плане, слов на целевую переменную. Исходя из этого понимания, использование подвыборок имен собственных, негативно окрашенных слов, знаков препинания или иных групп токенов для получения максимальной информации из текстов может очистить заголовки новостей от нерелевантной информации, сохранив лишь ту, которая в большей степени определяет смысловой и эмоциональный окрас текста и, соответственно, наиболее важна для определения реакции экономических агентов и их поведения относительно целевого финансового индекса.

Таким образом, проведенное исследование свидетельствует о наличии определенных взаимосвязей между текстовыми данными новостей и изменением композитного финансового индекса, которое, рассуждая в парадигме гипотезы эффективного рынка, отражает в себе всю существенную информацию, находящуюся в открытом доступе. Использование методов машинного обучения, в свою очередь, позволяет успешно моделировать эти взаимосвязи для предсказания целевой переменной и, в конечном итоге, построения конкурентоспособной модели на уровне best-practice решений в отрасли на данный момент. Помимо того, исследование открывает неочевидные паттерны, в которых текстовые данные из заголовков новостей влияют непосредственно на движение индекса, и предлагает перспективные направления дальнейшего развития на основе полученных результатов. Проведенная работа предоставляет как, своего рода, эмпирическое доказательство концепции, подкрепленное изученными деталями и

закономерностями, так и дополнительное знание, призванное стать частью большого исследовательского процесса решения задачи предсказания финансовых показателей на основе новостных данных.

Список литературы:

1. Araci, Dogu. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.https://doi.org/10.48550/arXiv.1908.10063
2. Church, K.W.. (2017). Emerging Trends: Word2Vec. 23. 155-162. 10.1017/S1351324916000334.
3. Ding, Xiao & Zhang, Yue & Liu, Ting & Duan, Junwen. (2015). Deep Learning for Event-Driven Stock Prediction. IJCAI. 2327-2333
4. Ding, Xiao & Zhang, Yue & Liu, Ting & Duan, Junwen. (2014). Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. 1415-1425. 10.3115/v1/D14-1148.
5. Gennaro, Giovanni & Buonanno, Amedeo & Palmieri, Francesco. (2021). Considerations about learning Word2Vec. The Journal of Supercomputing. 77. 10.1007/s11227-021-03743-2.
6. Goutte, Cyril & Gaussier, Eric. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25.
7. Haque, Mahfuzul & Shamsub, Hannarong. (2016). International Journal of Financial Studies Do Markets Cointegrate after Financial Crises? Evidence from G-20 Stock Markets. International Journal of Financial Studies. 2015. 557-586. 10.3390/ijfs3040557.
8. Kerssenfischer, Mark & Schmeling, Maik. (2021). What Moves Markets?. SSRN Electronic Journal. 10.2139/ssrn.3933777.
9. Khan, Wasiat & Ghazanfar, Mustansar ali & Azam, Muhammad Awais & Karami, Amin & Alyoubi, Khaled & Alfakeeh, Ahmed. (2020). Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing. 10.1007/s12652-020-01839-w.
10. Lo, Andrew, & Mackinlay, Craig. (1987). Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. NBER Working Paper Series. 10.3386/w2168

11. Loughran, Tim & Mcdonald, Bill. (2011). When Is a Liability NOT a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance. 66. 35 - 65. 10.1111/j.1540-6261.2010.01625.x.
12. Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.
13. Mao, Huina & Counts, Scott & Bollen, Johan. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. https://doi.org/10.48550/arXiv.1112.1051
14. Mao, Yuexin & Wei, Wei & Liu, Benyuan. (2012). Correlating S&P 500 stocks with Twitter data. 69-72. 10.1145/2392622.2392634.
15. Schumaker, Rob & Chen, Hsiu-chin. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst.. 27. 10.1145/1462198.1462204.
16. Soros, George. (1987). The Alchemy of Finance: Reading the Mind of the Market.
17. Tetlock, Paul. (2005). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance. 62. 10.2139/ssrn.685145.
18. Vargas, Manuel & Lima, Beatriz & Evsukoff, Alexandre. (2017). Deep learning for stock market prediction from financial news articles. 10.1109/CIVEMSA.2017.7995302.
19. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).
20. CatBoostClassifier method from Catboost library documentation. Date Views 01.04.2022 catboost.ai/en/docs/concepts/python-reference_catboostclassifier.
21. Comparing Iconic Indices: The S&P 500® and DJIA®. Date Views 01.04.2022 www.spglobal.com/spdji/en/documents/education/education-comparing-iconic-indices-the-sp-500-and-djia.pdf.

22. Current List of All Non-U.S. Issuers. Date Views 01.03.2022. www.nyse.com/publicdocs/nyse/data/CurListofallStocks.pdf
23. feature_extraction.text.CountVectorizer method from sklearn library documentation. Date Views 01.04.2022 scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.
24. Gensim library documentation. Date Views 01.04.2022 radimrehurek.com/gensim/.
25. Historical data: Dow Jones Industrial - U.S. (^DJI). Date Views 01.04.2022. https://stooq.com/q/d/?s=%5Edji.
26. SHAP library documentation. Date Views 01.04.2022 github.com/slundberg/shap
27. Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [Date You Retrieved This Data] from https://www.kaggle.com/aaron7sun/stocknews
28. tokenize method from nltk library documentation. Date Views 01.04.2022 www.nltk.org/api/nltk.tokenize.html.
29. World News subreddit. Date Views 01.04.2022 www.reddit.com/r/worldnews/.

Приложения

Приложение 1 — jupyter-ноутбук с проводимыми расчетами

Приложение 2 — jupyter-ноутбук с проводимыми расчетами

Приложение 3

Модель	accuracy	ROC-AUC	F1-score (macro)
TF-IDF + Logistic Regression (Baseline)	0.57	0.58	0,56
Word2Vec + Gradient Boosting	0,56	0,58	0,55
Word2Vec + LSTM	0,55	0,54	0,53
FinBERT + LSTM	0,54	0,54	0,52
TF-IDF + Gradient Boosting	0,57	0,59	0,57
Target encoding + Gradient Boosting	0,54	0,53	0,51

[Источник: проведенные расчеты]