ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему:

**Анализ тональности текстов политических новостей**

основная образовательная программа бакалавриата по направлению подготовки 45.03.02 «Лингвистика»

Исполнитель:

Обучающийся 4 курса

Образовательной программы

«Прикладная, экспериментальная и математическая лингвистика (английский язык)»

Профиль «Прикладная, экспериментальная и математическая лингвистика»

очной формы обучения

Николаева Дарья Игоревна

Научный руководитель:

к.ф.н., доц. Митренина О.В.

Рецензент:

к.ф.н., доц. Хохлова М.В.

Санкт-Петербург

2018

Оглавление

[Введение 2](#_Toc515885428)

[Глава 1. Анализ тональности новостей как инструмент финансового прогнозирования 5](#_Toc515885429)

[1.1. Гипотеза эффективного рынка 5](#_Toc515885430)

[1.2. Методологические основы анализа отображения СМИ политических событий 6](#_Toc515885431)

[1.3. Эмпирический анализ тональности политических новостей 9](#_Toc515885432)

[Глава 2. Автоматический анализ тональности новостных текстов 15](#_Toc515885433)

[2.1. Постановка задачи 15](#_Toc515885434)

[2.2. Основные методы анализа тональности документа 16](#_Toc515885435)

[2.3. Построение классификатора для определения тональности документа 17](#_Toc515885436)

[2.3.1 Постановка задачи 17](#_Toc515885437)

[2.3.2 Предобработка текста 18](#_Toc515885438)

[2.3.3 Извлечение признаков из текста 20](#_Toc515885439)

[2.3.3 Методы классификации текстов 23](#_Toc515885440)

[Глава 3. Эксперимент по предсказанию движения курса МосБиржи с помощью анализа тональности новостей 26](#_Toc515885444)

[3.1. Гипотезы о корреляции тональности новостей и индекса МосБиржи 26](#_Toc515885445)

[3.2. Исходные данные эксперимента 26](#_Toc515885446)

[3.3. Работа с корпусом отзывов 28](#_Toc515885447)

[3.4. Реализация программы классификации новостей по тональности 30](#_Toc515885448)

[3.5. Реализация программы предсказания движения курса МосБиржи 33](#_Toc515885449)

[Заключение 36](#_Toc515885450)

[Список литературы 38](#_Toc515885451)

[Приложение 1 42](#_Toc515885452)

# Введение

Предсказание изменений на фондовом рынке является важной задачей в области финансов, математики и инженерии. Принятие верного инвестиционного решения способствует стабилизации рынка и минимизации потерь конкретного лица, фирмы, государства. Финансовое прогнозирование – сложная задача, поскольку на рынок оказывают значительное влияние не только финансовые показатели и аналитические индикаторы, но также изменение политической ситуации, различные макроэкономические сигналы и малодоступная информация.

Особенно остро стоит проблема информационного влияния на нестабильных развивающихся рынках, где нет сформированного благоприятного климата. Экономическая и политическая нестабильность затрудняет возможность прогнозирования каких-либо событий. В условиях нестабильности информация о событии, которое никто не ожидал, возникает постфактум и оказывает доминирующее влияние на финансовые рынки. Средства массовой информации, конкурирующие друг с другом за право первым предоставить последние новости фондового рынка и экономической ситуации в целом, выпускающие в медиа-пространство постоянные сообщения о состоянии биржевых торгов, тем самым сами включаются в состав биржевой деятельности. Сенсационными новостями в области политики и экономики, заявлениями о новых тенденциях рынка они стараются привлечь внимание публики. В свою очередь инвесторы, получающие большое количество новой информации, оказываются не в силах оценить и обработать все информационные сообщения со стороны СМИ. Это приводит к упрощенному восприятию информации и **повышению значения ее эмоциональной оценки** [Чарахчян 2018].Именно этим и обусловлен интерес и **актуальность** исследования влияния настроений в политических и экономических новостях на ситуацию на российском рынке.

**Практическая значимость** исследования заключается в возможности использования тональности политических новостей для предсказания движения различных фондовых индексов и акций конкретных компаний. Полученные в ходе работы результаты могут быть востребованы при написании статей и учебных курсов в областях обработки естественного языка, контент-анализа, анализа данных и машинного обучения.

Актуальная и практическая значимость позволяют определить **цель** работы — создание программы анализа тональности политических новостей с опцией предсказания движения индекса МосБиржи с интервалом один день.

Для достижения цели были поставлены следующие задачи:

1. проанализировать, как в СМИ отражается политическое событие и формируется его оценка;
2. проследить корреляцию между тональностью новостей и ситуацией на российском финансовом рынке, на основе этого выдвинуть гипотезу;
3. изучить основные методы обработки текстов и анализа тональности;
4. собрать корпус новостей и подготовить его для загрузки в программу машинного обучения;
5. реализовать программу машинного обучения для классификации новостей по тональности.

**Объектом** исследования являются политические новости российских СМИ, посвященные событиям в России и графики индекса МосБиржи за четырехмесячный период.

**Предметом** научного анализа является тональность новостей, конструкции и n-граммы, используемые для создания эмоциональной окраски сообщения.

Характеризуя использованную литературу, можно отметить, что в процессе исследования упор делался на классические труды по лингвистике и медиадискурсу Дейка и Гавра, а также на современные работы лингвистов, журналистов и экономистов Абдуллина, Боярского, Науменко и др. При написании главы об автоматическом анализе тональности использовались книги исследователя Б. Лиу, одного из самых известных ученых в области сентимент-анализа. При эмпирическом анализе новостей использовались статьи из известных интернет-изданий, таких как Ведомости, РБК, Эксперт, BBC. Широко использовались интернет-ресурсы по программированию и машинному обучению.

**Научная новизна** работы обоснована разработкой новых теоретических и практических подходов к анализу тональности новостных сообщений. Впервые в отечественной науке **автоматический** анализ тональности использован как инструмент финансового прогнозирования.

Результаты исследования были представлены в качестве доклада на конференции РГГУ СКИЛ в апреле 2018 года и получили высокую оценку жюри.

# Глава 1. Анализ тональности новостей как инструмент финансового прогнозирования

## 1.1. Гипотеза эффективного рынка

Проблема влияния информации на тенденции фондового рынка в целом и на динамику котировок ценных бумаг – одна из наименее изученных в теории инвестиций. Данный вопрос решает гипотеза эффективного рынка (англ. Efficient market hypothesis, EMH), сформулированная американским экономистом Юджином Фама. Согласно этой гипотезе рынок рационально реагирует на поступающую информацию, то есть она сразу и полностью отражается на цене активов. Это означает, что участники рынка однородно интерпретируют новую информацию и мгновенно корректируют решения при ее поступлении [Абдуллин 2015].

С формальной точки зрения эффективность оценивается характером получаемой рынком и его участниками информации. Различают три вида доступной информации:

1. информация, содержащаяся в прошлых значениях цен;
2. информация, содержащаяся в публично доступных источниках (газеты, радио, телевидение);
3. вся мыслимая информация [Ширяев 1998: 544].

Существуют три варианта воздействия информации на рынок бумаг:

1. слабая форма – цены отражают всю информацию, которая получена в результате изучения статистики и уже заложена в стоимость акции;
2. квази-сильная форма – доступная информация отражает перспективы развития компании и играет роль в изменении тенденции на фондовом рынке;
3. сильная форма – цены отражают информацию, доступную инсайдерам, информация является основной в плане изменении тенденции на фондовом рынке [Чарахчян 2018: 48].

Из сказанного выше можно сделать вывод о том, что информация уже заключена в ценах активов, а значит, использование информации в той или иной форме не может приносить сверхприбыль. Однако гипотеза говорит об идеальном рынке, в реальности же он стремится к эффективности, но не всегда достигает ее. Особенно это касается нестабильных развивающихся рынков, к которым относится и российский. Информация учитывается не мгновенно после ее появления в рыночной среде, то есть существует интервал времени прежде чем она отразится на цене актива. Более того, в условиях политической и экономической нестабильности нельзя предсказать возникновение того или иного события. Именно поэтому анализ информации, выпускаемой в медийное пространство имеет вес при принятии инвестиционного решения.

## 1.2. Методологические основы анализа отображения СМИ политических событий

Процесс отражения в средствах массовой информации политических и экономических событий определен множеством факторов, в числе которых находится степень включенности СМИ в структуру конфликта, следовательно, также перечень выполняемых ими функций. Являясь активным игроком на поле информационного противоборства, разные СМИ, поддерживающие отдельных участников конфликта, будут пытаться убедить в правильности продвигаемой позиции как можно большие аудитории, оказать воздействие на общественность и элиты.

В отношении внутренней общественности превалирующая реализация той или иной функции СМИ зависит от типа события. В случае возникновения политического конфликта, например, государству как субъекту конфликтного взаимодействия необходима поддержка широких слоев общественности. В этом случае приоритет получают регулятивные функции СМИ (как социально- организующие, так и функции организации информационного фона), реализации которых, как правило, сопутствуют серьезные нарушения принципов объективности и взвешенности в подаче информации [Науменко 2015: 63]. В процессе отображения политического события СМИ способны оказывать влияние на различные аудитории, которые, в свою очередь, определяют экономическую ситуацию внутри государства (то есть, влияют на рынок). Коммуникативные действия, направленные на политические и бизнес-элиты, косвенным образом могут воздействовать на процесс принятия решений, касающихся экономической сферы в целом, отдельных отраслей и компаний. Более того, политический конфликт в качестве информационного продукта востребован аудиторией, чей интерес к событиям подобного рода связан с реализацией профессиональных практик. Сюда относятся и инвесторы, для которых изменение политической и экономической ситуации в стране становится ключом к принятию инвестиционного решения.

Введем несколько теоретических понятий функционирования средств массовой информации в рамках конфликта. Так как задача СМИ – информирование общественности о возникающих событиях, они становятся частью среды, в которой развивается данное событие. Следовательно, освещая событие, СМИ участвуют в формировании политического медиадискурса. Понятие медиадискурса является производным от общего понятия дискурса. Теун ван Дейк определяет дискурс как сложное коммуникативное явление, включающее в себя всю совокупность экстралингвистических факторов, сопровождающих процесс коммуникации: «Дискурс, в широком смысле слова, является сложным единством языковой формы, значения и действия, которое могло бы быть наилучшим образом характеризовано с помощью понятия коммуникативного события или коммуникативного акта» [Дейк 1989: 6].

К основным способам формирования медиадискурса относятся: эскалация конфликта, увеличения напряженности конфликта, урегулирование конфликта, наблюдение со стороны, умолчание. Также выделяют пять базовых типов отношения СМИ к субъекту конфликта: поддержка, одобрение, нейтральное отношение, неодобрение, критика [Гавра 2009: 103].

Основной метод исследования – это контент-анализ. Данный метод представляет собой достаточно строгий количественный и качественный анализ содержания документов. Ключевые инструменты контент-анализа позволяют определить тематическое содержание, уровень и характер представленности субъектов конфликта, оценочные характеристики, касающиеся участников события и оценка самого события, а также источники оценочных высказываний. При определении характера включенности тех или иных средств массовой информации в медиадискурс, необходимо провести количественный анализ и определить интенсивность освещения данного события. Следует рассмотреть предлагаемое редакцией позиционирование события в качестве важного, значимого; установить, каким образом оно встраивается в общий информационный поток, «повестку дня», и какое место в нем занимает. В качестве показателей для данного параметра могут выступать: количество публикаций, посвященных данному конфликту, за определенный период; расположение публикаций в номере; жанровая специфика публикаций, посвященных данному конфликтному противоборству (использование информационных/аналитических жанров, посвящение заявленной теме редакторской колонки либо коротких заметок в хронике) [Добросклонская 2008: 198].

Анализируя коммуникативную стратегию СМИ в освещении события, важно изучать процессы отбора фактов, их интерпретации, формирования отношения к конфликтному событию и его субъектам. На данном этапе рассматривается содержательный параметр новостных сообщений: количество, тематическая направленность, наличие/отсутствие и содержание фоновой информации. Данные параметры отражают репрезентативный уровень конструируемого средствами массовой информации медиадискурса. Интерпретация конфликтного события происходит на двух уровнях: десигнативном и коннотативном.

При оценке десигнативного компонента определяется распределение СМИ ролей между субъектами: присвоение статуса прямого/косвенного участника, виновника/жертвы и т.п. Это формирует у общественности оценочные характеристики того или иного актора. При анализе публикаций важно анализировать экономический и политический компонент сообщения. К экономической сфере относятся описание деятельности субъектов рынка, заключение договоров, параметры соглашений, фиксация результатов конфликта. Политический компонент связан с воспроизведением политических «месседжей», стереотипов, проявлением внеэкономических либо политизированных категорий. Интерпретация конфликта в конечном итоге должна давать завершенную картину события, предоставлять непротиворечащие друг другу факты и версии. В число компонентов интерпретации входят предмет интерпретации, коммуникативная направленность автора, оценка, идеологические факторы, влияющие на результат [Гавра 2009: 112].

При характеристике коммуникативной стратегии СМИ при освещении события важно учитывать коннотативный компонент сообщения. В его рамках журналист вводит экспрессивно-оценочные компоненты, формирующие дополнительные значения и ассоциативные представления. Оценка дается напрямую или опосредованно – через привлечение мнений экспертов, аналитиков, сторонников и противников субъекта. Коннотации формируются с помощью фигур речи и тропов. К языковым явлениям, с помощью которых формируется коннотативный компонент сообщения, относятся метафора, метонимия, ирония. Выявление и интерпретация коннотативных смыслов служит инструментом для определения связи тональности новостного сообщения с ситуацией на рынке акций.

## 1.3. Эмпирический анализ тональности политических новостей

В данной работе машинный анализ тональности текстов, предваряется эмпирическим исследованием с применением методов мониторинга и селекции, при отборе материала и ранжирования его по жанровому и тематическому признакам соответственно, а так же качественный и количественный контент-анализ, при дальнейшей работе с материалом. Работа с эмпирической базой в данном случае призвана наглядно показать и раскрыть механизмы, задействованные при аналогичном процессе при использовании автоматического анализа. Инфоповоды, выбранные для исследования, а именно судебное разбирательство между компаниями АФК Система и Роснефть по делу Башнефти и «Дело Скрипалей», были весьма резонансными, широко освещались в СМИ и соответственно, оставили обширный след на информационном поле. Ввиду важности данных событий, различные стороны конфликтов пытались утвердить свою точку зрения в информационном пространстве, с целью получения выгоды, что гарантирует широкий разброс мнений и явную эмоциональную окрашенность материалов, что является необходимым условием для исследования. Выбранные темы относятся к разным областям, что позволяет продемонстрировать, каким образом как политические так и экономические факторы влияют на фондовый рынок. Последним важным критерием является то, что оба события являются «форс-мажорами», что гарантирует, что информация не могла быть заранее заложена в рынок.

7 декабря 2017 года произошел очередной виток судебного процесса «О БАШНЕФТИ» между АФК СИСТЕМА и РОСНЕФТЬ, в результате которого акции компании АФК СИСТЕМА упали за пять минут с 11500 рублей до 9500 рублей (на 18 %). Дело «Башнефти» – корпоративный конфликт между госкорпорацией «Роснефть» и частной российской компанией АФК «Система». 3 марта 2017 года компании НК «Роснефть» и АНК «Башнефть» подали иск к АФК «Система» и ее дочерней компании «Система-Инвест» о взыскании 106,6 млрд рублей убытков, по словам истцов понесенных ими в связи с реорганизацией «Башнефти» в 2013-2014 годах, когда ее основным акционером была АФК «Система». Позднее сумма требований возросла до 170,6 млрд рублей. Новость об иске негативно повлияла на котировки акций частной компании. Судебный спор с крупнейшей нефтяной госкорпорацией нанес серьезный удар по рыночной капитализации одного из ведущих частных инвесторов в российскую экономику, став одним из самых резонансных после Дела ЮКОСа, а также оказал заметное влияние на инвестиционный климат в стране, полагают эксперты [BBC].

7 декабря в 15-00 ведущими бизнес-изданиями (RBK, БКС-экспресс, Ведомости) внезапно опубликована новость о новом иске «Роснефти» к АФК «Системе» на 131 млрд рублей. Новость явно оказалась неожиданной как для крупных игроков, так и для рядовых инвесторов, не была заложена в стоимость акций и произвела «паник-сейл», то есть продажу ценных бумаг под руководства чувства страха (см. рисунок 1).

Уже через 20 минут была опубликована следующая новость, где глава «Системы» В. Евтушенков заявил, что данный иск является «акцией устрашения», что позволило остановить падение акций и даже взывало небольшой отскок [РБК].



*Рисунок 1. Паник-сейл акций АФК «Система» 7.12.18*

Стоит заметить, что ответ Евтушенкова был срочно опубликован лишь независимо-оппозиционными РКБ и Медузой. После 16 часов был опубликован комментарий Пескова, в котором было выражено индифферентное мнение правительства к конфликту, что остановило дальнейший слив компании АФК «Система». Далее продолжилась эскалация конфликта, нашедшая отражение во всех СМИ. РБК занял в конфликте позицию «Системы», активно приводя точку зрения представителей компании и экспертов, считающих действия «Роснефти» «злоупотреблением правом» [РБК]. Мнение эксперта, где иск называют «кафкианским», подчеркивается, что слова представителей «Роснефти» о значении нового иска как «принуждения к миру» нельзя считать достоверными, ведь «Роснефтью» не было озвучено, что новый иск будет отозван, в случае согласия «Системы» погасить первые 131 млрд. В авторской статье адвоката Мельникова и вовсе обличается вся судебная система РФ, при которой возможен такой надутый процесс как в случае с двумя данными компаниями [РБК]. Схожую позицию заняли и Ведомости: например, статья под вычурным заголовком «Чем «Роснефть» похожа на Звезду Смерти из «Звездных войн». Хватит ли силы «Системе», чтобы дать отпор?», - хотя данный стиль все же не характерен для обычно крайне сдержанных Ведомостей [Ведомости]. И уже в статье ««Система» обвинила «Роснефть» в давлении. Компания Евтушенкова разъяснила, как насчитала убытки на 330,4 млрд рублей от исков «Роснефти» и «Башнефти»» от 12 декабря, издание подходит со свойственной ему бизнес-прагматичностью, обезличено приводя позиции обеих сторон. В том числе, сумма ответного иска АФК к «Роснефти» в 330 млрд называется трудно обосновываемой [Ведомости].

Интересно и показательно для анализа отражение в СМИ недавнего дела Скрипаля. 4 марта 2018 года в Солсбери (Великобритания) ядом Новичок был отравлен бывший сотрудник ГРУ Сергей Скрипаль и его дочь Юлия, прилетевшая посетить отца в Великобританию. 12 марта премьер-министр Великобритании Тереза Мэй выступила с докладом в Парламенте со словами: «Отравление Скрипаля – акт государственной агрессии против граждан Британии [Youtube]». Великобритания прямо назвала Россию причастной к покушению на убийство и в нарушении Конвенции о запрещении химического оружия. 13 марта десятки стран европейского союза поддержали позицию Великобритании. Разразился громкий дипломатический скандал, и более 30 стран ЕС объявили о высылке российских дипломатов. В то же время в СМИ начинают появляться резкие высказывания о конфликте. Журнал «Эксперт» вообще позволяет себе экспрессивные высказывания вроде «Британка опять гадит», «Неприятный для России прецедент», «фейковые обвинения» [Expert]. Ресурс vesti.ru, поддерживающий курс правительства, заполняется оппозиционными к Британии выражениями, такими как «заранее разработанный сценарий», «провокация», «недопустимые действия», «новая модель демонизации России» [Вести]. В данной работе я не берусь оценивать, объективным ли являются данные источники, ведь как уже было сказано ранее, журналист всегда вносит коннотативный компонент в новостное сообщение, создавая медиадискурс. Здесь важно отметить, что при эскалации конфликта в СМИ создавался негативный эмоциональный фон, что привело к последующему обвалу российскому рынку. Более того, отрицательная тональность сохранялась и в оппозиционных изданиях, типа RBK. В статье от 13 марта под названием «Лондон дал сутки на детоксикацию [РБК]», приводится практически в переводном виде речь Т. Мэй, направленная на начало действий против России, таких как «блокировка активов россиян», «бойкот чемпионата мира по футболу» и «высылка российских дипломатов». О скандале писали все ведущие российские издательства, и инвесторы, предполагая введение новых санкций, стали массово выводить деньги с российского рынка, в первую очередь, со Сбербанка, который является самой ликвидной ценной бумагой в России. Поскольку Сбербанк привлекает множество иностранных инвесторов, то политический скандал привел к выводу иностранных капиталов из компании, и, как следствие, к падению акций Сбербанка на 10 %, и всего индекса МосБиржи на 4 %.



*Рисунок 2. Падение акций Сбербанка, вызванное выводом иностранного капитала из компании, в ре-те конфликта по «Делу Скрипаля»*

В результате анализа этого события можно сделать вывод, что важно рассматривать, как освещение в СМИ внутригосударственных событий, так и отношений России с другими странами, так как российский рынок привлекает множество иностранных инвесторов. Конечно, на фондовом рынке экономические события превалируют на политическими, поэтому к 21 марта (см. рисунок 1) российский рынок уже отыграл позиций, тем не менее, очевидно, что росрынок не до конца самостоятелен и зависит от иностранных вложений, а иностранные инвесторы читают как свою, так и зарубежную прессу, и принимают инвестиционные решения, основываясь на полученной информации. Больше информации о связи эмоционального фона новостей и ситуацией на фондовой бирже см. в Приложении 1.

# Глава 2. Автоматический анализ тональности новостных текстов

## 2.1. Постановка задачи

Анализ тональности текста (англ. Sentiment analysis and opinion mining) – это класс методов в компьютерной лингвистике, предназначенный для выявления и дальнейшего анализа эмоционально окрашенной лексики в тексте. Эмоциональная окраска, выраженная в тексте, также называется тональностью текста.

В общем виде задачу сентимент-анализа можно сформулировать следующим образом: в документе d, содержащим мнение, извлечь все пятерки (ei, aij, sijkl, hk, tl). Ввиду сложности проблемы, исследователи [Liu 2012: 23] выделяют шесть подзадач:

1. извлечение всех слов и словосочетаний, обозначающих объекты; группировка синонимичных выражений в кластеры, каждый из которых обозначает один объект
2. извлечение всех слов и словосочетаний, обозначающих атрибуты; группировка синонимичных выражений в кластеры, каждый из который обозначает один атрибут определенного объекта
3. извлечение авторов текста/сообщений
4. извлечение времени и приведение всех дат к единому формату
5. классификация по тональности – определить является ли мнение положительным, отрицательным или нейтральным или приписать рейтинг объекту/атрибуту
6. выдать все пятерки (ei, aij, sijkl, hk, tl), содержащиеся в документе d, основываясь на результатах пунктов 1-5

У сентимент-анализа есть несколько областей исследования. Поскольку данная работа посвящена лишь одному из них, автор здесь ограничится простым перечислением этих направлений:

1. анализ тональности на уровне документа (document-level sentiment classification);
2. анализ тональности на уровне предложения (sentence-level sentiment classification);
3. классификация по шкале "субъективность/объективность»;
4. аспектный анализ тональности (aspect-based sentiment analysis);
5. анализ мнений-сравнений;
6. обобщение мнение;
7. определение спама в отзывах.

## 2.2. Основные методы анализа тональности документа

В данном исследовании задача анализа тональности будет решаться на уровне документа – новостного сообщения. При таком анализе ставится цель определить, какую тональность в целом – положительную, отрицательную (негативную) или нейтральную (отсутствие сентимента), несет документ. В предыдущем разделе уже была сформулирована задача анализа тональности текста, теперь введем ее уточнение для документа. В документе d, содержащим оценочное высказывание об объекте, определить тональность s касательно данного объекта:

(\_ , GENERAL, s, \_, \_ ), где атрибуты рассматриваются как одно целое, объект заранее известен, а автор и время не учитываются.

Для того, чтобы задача имела смысл, вводится следующие допущение: документ d содержит мнение, высказанное одним автором, об одном объекте [Liu 2010: 10]. Конечно, в новостях часто приводят прямую речь, мнения и оценки разных людей. В рамках данной работы определяется общая тональность новости, на допущении, что она написана одним автором (журналистом) и освещает одно событие (конфликт).

Для автоматического определения тональности текстов существуют следующие подходы:

1. На основе правил с использованием шаблонов (rule-based with patterns), где для анализа создается жесткая система правил «Если условие, то заключение». Данный метод дает самую большую точность, но действует в рамках определенной тематики (например, отели). Для набора отзывов другой тематики нужно разрабатывать другую систему правил [Хохлова 2017].
2. Машинное обучение без учителя [Turney 2002: 417] (unsupervised learning). Подход заключается в том, что в отзывах учитываются позитивные и негативные прилагательные и наречия, а после вычисляется семантическая ориентация по специальной формуле. Отзыв считается положительным, если результат семантической ориентации всех фраз отзыва больше нуля.
3. Машинное обучение с учителем (supervised learning). В данном подходе требуется обучающая коллекция отзывов, на базе которой строится статистический классификатор.
4. На основе тонального словаря (affective lexicon), где используется список слов со значением тональности для каждого слова. В этом подходе также используется классификатор.

## 2.3. Построение классификатора для определения тональности документа

### 2.3.1 Постановка задачи

Машинное обучение – это раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Задачу определения тональности текста можно представить как задачу классификации. Неформально задача классификации может быть сформулирована следующим образом: имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества. Классифицировать объект — значит, указать номер (или наименование класса), к которому относится данный объект [Machinelearning].

Классификация может быть двухклассовой и многоклассовой. В данной работе будет решаться задача бинарной классификации. Теперь мы можем переформулировать задачу с использованием математических терминов:

Пусть *X* — множество описаний объектов, *Y* — конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение , значения которой известны только на объектах конечной обучающей выборки Требуется построить алгоритм способный классифицировать произвольный объект .

### 2.3.2 Предобработка текста

Для обучения классификатора нужно разделить коллекцию документов на обучающую и тестовую выборки. Выборка (sample, set) — конечный набор прецедентов (объектов, случаев, событий, испытуемых, образцов, и т.п.), некоторым способом выбранных из множества всех возможных прецедентов, называемого генеральной совокупностью [Machinelearning]. Обучающая выборка – это тот набор данных, который используется собственно для обучения программы. По этим данным производится настройка математических закономерностей. Тестовая (контрольная) выборка служит для оценки построенной модели. Как правило, обучающая выборка составляется 75-80 % от всех данных.

Однако перед тем как разделить данные на обучающую и тестовую выборки, нужно привести каждый текст к единому формату. Следует привести все слова к нижнему регистру, так как строки «Автомобиль» и «автомобиль» будут распознаваться компьютером как разные слова. Далее документ должен пройти стадию «очистки»: нужно удалить html теги, знаки пунктуации, ссылки. На данном этапе возникают две проблемы. Стоит учитывать, что некоторые слова пишутся через дефис (напр. какой-либо, Санкт-Петербург), и удаление такого знака пунктуации ведет к потери целостности токена.

Следующий этап обработки документов – токенизация, то есть выделение минимальных неделимых при дальнейшем анализе единиц текста [Боярский 2013: 217]. Недостаточно просто разделить текст на токены по пробелам, так как в языке существует ряд случаев, когда такой подход не будет работать. Например, название города «Нижний Новгород» состоит из двух слов, но при обучении нужно, чтобы оно было одним токеном [Coursera]. Подобный случай также представляют собой составные числительные вроде «триста двадцать первый». Для решения подобных проблем существуют специальные программы – токенизаторы. Их применение будет расстроено во второй главе.

Заключительный этап препроцессинга – нормализация, то есть приведение все слов к начальной форме. От количества уникальных токенов зависит размерность признакового пространства и для улучшения производительности нужно, чтобы оно было меньше. Нормализация в данном случае помогает успешно уменьшить количество признаков. Существуют два подхода к решению этой задачи: стемминг и лемматизация.

Стемминг – это процесс нахождения основы слова. Самый известный алгоритм стемминга называется стеммер Портера, заключающийся в последовательном отсечении окончаний и суффиксов. Например, если в тексте встречаются слова «длинный» и «длинная», стеммер усечет их до «длин». Однако, у процедуры стемминга есть ряд недостатков. Во-первых, метод стемминга приводит к слишком большому обобщению (также называется overstemming). Это значит, что под усеченные формы будут попадать разные слова. Например, алгоритм Портера сопоставит слова «universal», «university» и «universe» основе «univers», а считать их одним и тем же токеном будет ошибкой. Во-вторых, в русском языке есть ряд слов, имеющих супплетивные формы. Примерами таких слов будут глаголы «идти» (шел, иду) и «быть» (был, буду). Стеммер с такими случаями не справляется, а они довольно распространены.

Хотя и реализация стеммера более проста и не требует подключения внешней базы данных, процедура лемматизации дает лучшие результаты. Лемматизация – это процесс приведения слов к начальной форме. Лемматизатор устроен таким образом, что в нем содержится словарь, в котором записано большое количество словоформ. Поскольку в языке часто встречаются омоформы типа «прекрасен стих» - «ветер стих», лемматизатор выдает все возможные варианты, в соответствии со словарными статьями. Некоторые лемматизаторы разрешают омонимию, опираясь на размеченный корпус словоупотреблений. В таком случае лемматизатор подгружает вероятностную модель в оперативную память и использует ее для выбора наиболее вероятной леммы. Так, для разрешения неоднозначности *роем – рой / рыть* лемматизатор будет опираться на самое частотное употребление (пчелиным роем) и учитывать синтаксические особенности (мы роем – после местоимения в именительном падеже идет личная форма глагола рыть) [Solarix].

Также уменьшить количество признаков поможет удаление так называемых «стоп-слов» - вспомогательных слов, имеющих высокую встречаемость в текстах, но несущих мало информации о содержании текста. Чаще всего это предлоги, союзы, частицы вроде «а», «кроме того», «потому что» и т.д. Для решения задачи сентимент-анализа важно оставить отрицательные частицы и местоимения «не», «ничего», «никакой» так как наличие или отсутствие их в тексте влияет на тональность (напр. «мне понравился фильм» - «мне не понравился фильм).

### 2.3.3 Извлечение признаков из текста

Признаком (feature) называется отображение , где *X* – множество описаний объектов, – множество допустимых значений признака. Если заданы признаки , то вектор называется признаковым описанием объекта . Признаковые описания допустимо отождествлять с самими объектами. Множество называют признаковым пространством [Machinelearning].

То, какие признаки будут выбраны для обучения модели, сильно влияет на качество результатов. Присутствие неинформативных и слабоинформативных признаков приводит к снижению точности некоторых моделей. Отбор признаков перед классификацией обеспечивает:

1. уменьшение переобучения – существуют так называемые шумовые признаки, которые не относятся к решаемой задаче [Datareview]. Чем меньше избыточных данных, тем меньше будет возможность, что решение будет принято на основе «шума»
2. повышение точности
3. сокращение времени обучения – чем больше признаков, тем более сложная модель получается, и тем больше времени необходимо, чтобы построить прогноз

Для классификации текстов по тональности часто используется модель «мешок слов» (bag-of-words). В данной модели текст представляется как неупорядоченное множество (англ. bag, multiset) слов, в котором порядок и грамматика не учитывается. Более формально описать этот метод можно следующим образом. Пусть всего в выборке N различных слов: w1, . . . , wN . В этом случае каждый текст кодируется с помощью N признаков, причём признак j — это доля вхождений слова wj среди всех вхождений слов в документе, называемая также TF – term frequency, частота термина [Coursera]. Сумма всех значений признаков документа составляет единицу. Как уже отмечалось ранее, в данной схеме взвешивания не учитываются стоп-слова и редкие слова.

Более сложный подход называется TF-IDF (*TF — term frequency, IDF — inverse document frequency*). В схеме взвешивания TF все слова имеют одинаковую важность. Рассмотрим пример: существует коллекция экономических текстов. Очевидно, что слово «экономика» будет встречаться во многих текстах, следовательно, это слово не несет особого значения при определении тональности. Чтобы устранить этот недостаток, для каждого слова считается его обратная документная частота по формуле:

idft =

Перемножая значения tf и idf получают схему взвешивания TF-IDF. Таким образом, значение idf для редко встречающегося слова будет высокой, а для слова с высокой частотой – низкой. Если слово встречается во всех документах, значение признака будет нулевым.

Однако даже учитывая обратную документную частоту, в подходе «мешок слов» не берется во внимание порядок слов, а для задачи определения тональности он имеет большое значение. Например, словосочетание «нравится» будет говорить о положительной окраске предложения, в то время как «не нравится» об отрицательной. Для того, чтобы компьютер учитывал эти различия, используют n-граммы. N-грамма – это последовательность следующих друг за другом n слов. В предложении «Трамп ввел санкции против России » можно выделить следующие n-граммы:

* униграммы (n=1): Трамп, ввел, санкции, против, России
* биграммы (n=2): Трамп ввел, ввел санкции, санкции против, против России
* триграммы (n=3): Трамп ввел санкции, ввел санкции против, санкции против России

Не стоит брать слишком большие n, так как с увеличением n растет количество уникальных признаков, а это может привести к переобучению. Имеет смысл также использовать в качестве признаков буквенные n-граммы – это поможет учитывать орфографические ошибки, незнакомые формы известных слов.

Существует более расширенный поход к n-граммам – k-skip-n-граммы. Skip-граммы – это наборы из n токенов, причём расстояние между соседними должно составлять не более k токенов [Coursera]. Использование skip-граммы в качестве признаков помогает преодолеть проблемы разреженности данных (data sparsity problem).

Для вычисления признаков в текстовых документах часто используется хэширование (feature hashing, hashing trick). При машинном обучении документы представляются в виде матриц [документ[слово[признак…]…]…]. Хранение таких структур данных занимает много места, которое растет при увеличении размера выборки. Вместо этого можно применить функцию хэширования h(x), которая принимает на вход и выдает некоторый хэш. В результате построим вектор длины 2n, где n – некоторое небольшое число. Все слова в тексте можно заменить на их хэши и уже для них вычислять признаки. Использование хэширования помогает понизить признаковое пространство и быстрее его построить.

### 2.3.3 Методы классификации текстов

В данном параграфе будут описаны статистические методы классификации, используемые для решения задачи определения тональности текстов в этой работе.

**Наивный Байесовский классификатор**

Наивный Байесовский классификатор – это алгоритм классификации, опирающийся на теорему Байеса. Наивный байесовский классификатор принимает решение о принадлежности документа d к классу c на основе вероятности P(d|c) принадлежности документа к этому классу.

Знаменатель P(d) в данной формуле может быть опущен, так как вероятность для одного и того же документа d будет одинаковой. Так как документ представлен в виде вектора *d =*{ *w1* *, w2* *,…, wn* }, то условную вероятность P(d|c) можно выразить как P(w1…wn | c) или же *P* (*w1* | *c*) \* (*w2* | *c*) \*...\* ( *wn* | *c*) = Пi *P*(*w*i | *cj*). Таким образом, для нахождения наиболее вероятного класса c нужно посчитать условные вероятности для каждого из возможных классов и выбрать тот класс, что имеет самое высокое значение вероятности:

Метод наивной байесовской классификации строится на допущениях о том, что

1. признаки условно независимы – то есть наличие одного признака в классе не связано с наличием другого
2. позиция слова в предложении неважна – не учитывается, что вероятность встретить конкретное слово в той или иной позиции различна

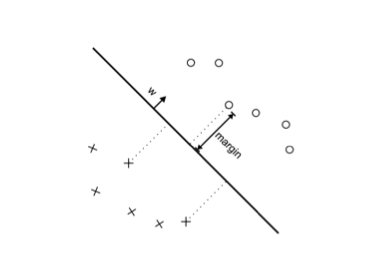
Из-за таких серьезных допущений модель получила название «наивной». Однако наивный байесовский классификатор является простой и эффективной моделью для решения задачи сентимент-анализа.

**Метод опорных векторов**

Метод опорных векторов (SVM, support vector machine) – это набор алгоритмов, относящихся к семейству линейных классификаторов. В данном походе каждый объект представляется как вектор (точка) в p-мерной пространстве. Каждая из точек принадлежит одному из двух классов. Задача метода — преобразовать пространство при помощи оператора ядра так, чтобы нашлись такие гиперплоскости, которые разделяют примеры из разных классов обучающей выборки. Если такая гиперплоскость существует, она называется оптимальной разделяющей гиперплоскостью, а соответствующий ей линейный классификатор называется оптимально разделяющим классификатором [Wikipedia].

Классифицирующая функция *F* принимает вид *F(x) =* sign*(<w, x> + b)* . Выражение   называется *ядром* классификатора. С математической точки зрения ядром может служить любая положительно определенная симметричная функция двух переменных [Habrahabr]. В задачах классификации текстов линейное ядро – самый лучший выбор. В таком случае ядро совпадает со скалярным произведением векторов:

K(w, x) = < w, x>



*Рисунок 3. Двоичная классификация SVM с линейным оператором ядра. 𝑚𝑎𝑟𝑔𝑖𝑛 — расстояние от гиперплоскости до каждого из классов. 𝑤 — вектор нового примера, для которого делается предсказание.*

# Глава 3. Эксперимент по предсказанию движения курса МосБиржи с помощью анализа тональности новостей

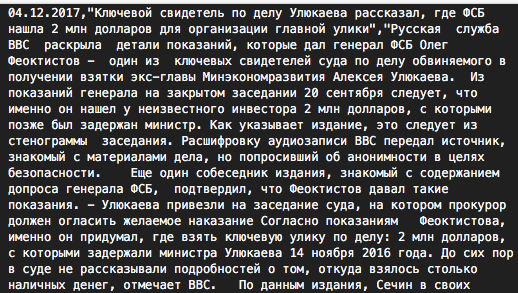
## 3.1. Гипотезы о корреляции тональности новостей и индекса МосБиржи

Проведенный в главе 1 эмпирический анализ новостей, а также обнаруженное влияние их на курсы акций российского рынка позволяет выдвинуть две гипотезы. Гипотезы задаются в общем виде, относительно индекса МосБиржи. Данное решение обсуловлено тем, что индекс МосБиржи – это композитный ценовой индекс, включающий в себя акции 50 наиболее ликвидных компаний русского рынка (такие, как GAZPROM PJSC, LUKOIL Oil Company, Aeroflot, Megafon, Norilsk Nickel JSC и т. д.). Следовательно, данный индекс отражает движение российского рынка в целом, и если предсказание будет верно для него, то в дальнейшем становится возможной работа с отдельно взятыми акциями, однако для них может потребоваться дополнительный отбор текстов для обучения. Таким образом, были сформулированы гипотезы о корреляции тональности новостей и индекса МосБиржи:

1. Тональность политических новостей влияет на значение индекса Мосбиржи
2. Пусть s – индекс тональности документа, а m – индекс МосБиржи. Если – индекс тональности документа i-го дня и > 0, то возрастает, а если , то убывает.

## 3.2. Исходные данные эксперимента

Для написания программы предсказания движения индекса МосБиржи требовались данные для обучения. Это данные двух типов: извлеченные из новостных текстов (тональность новостей) и технические индикаторы (цена открытия и закрытия индекса).

Для решения задачи сентимент-анализа был собран корпус новостей с сайта newsru.com за 4-месячный период с 1 декабря 2017 года по 31 марта 2018 года. Корпус был собран автоматически с помощью фреймворка Scrapy для Python. Корпус содержит 3011 новостей с раздела «В России», в котором отражены актуальные политические и экономические новости. Новости записаны в формат.csv: одна ячейка представляет собой текст формата дата – название – собственно новость (корпус доступен по ссылке <https://github.com/daranikolaeva427/sentimentstockpredict2018>).

*Рисунок 4. Фрагмент корпуса политических новостей.*

Технические индикаторы – значения открытия и закрытия торгов были скачаны с сайта МосБиржи moex.com за тот же период, что и новости. Значения индекса устанавливаются в рабочие дни, с понедельника по пятницу.



*Рисунок 5. Фрагмент таблицы со значениями индекса Мосбиржи.*

Используемые индикаторы – значение открытия и закрытия торгов.

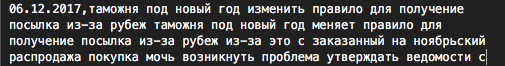
## 3.3. Работа с корпусом отзывов

Для решения задачи классификации новостей по тональности имеющийся корпус текстов был обработан и полуавтоматически размечен. Во-первых, были удалены знаки препинания, а все слова приведены к нижнему регистру. Во-вторых, корпус был лемматизирован для уменьшения признакового пространства и для упрощения задачи определения тональности.

Для решения задачи лемматизации была выбрана программа Mystem. Mystem производит морфологический анализ текста на русском языке, а также строит гипотезы для слов, не входящих в словарь [Mystem]. Программа работает через терминал: на вход подается текстовый файл с отзывами, далее указываются параметры анализа. В нашем случае это:

* c – возврат к полному представлению текста
* l – не печатать исходные словоформы, только леммы и граммемы
* d - применить контекстное снятие омонимии

На выходе получаем файл формата:



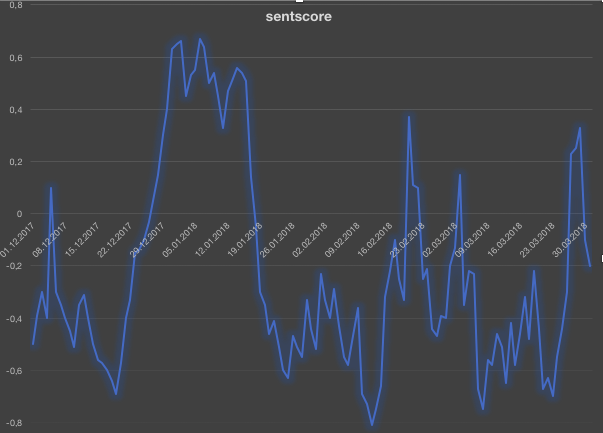
*Рисунок 6. Лемматизированный корпус новостей*

Следующим шагом было составление частотного списка слов, с приписанной им тональностью. Это требовалось для создания распределения тональности новостей, чтобы проследить тенденции за выбранный период. Частотный список корпуса был составлен автоматически (код программы доступен по ссылке <https://github.com/daranikolaeva427/sentimentstockpredict2018>). Программа просматривала элементы лемматизированного корпуса и записывала в массив слова с наращиванием их частоты. Оценки не проставлялись наиболее частотным словам, вроде «и», «но», «а», «он», «сказать», так как они в большинстве случаев не несли эмоциональной окраски. Итого было размечено 5867 слов. Разметка уровня тональности была выполнена вручную. Было выбрано два уровня тональности: 1 для положительного уровня и -1 для отрицательного. Оценку -1 получили такие частотные экспрессивные слова, как санкция, война, удар, критика, чиновник, падение, чекист, пожар, катастрофа, беспорядок, фальсификация, позорный и др. Оценка +1 была проставлена следующим словам: развитие, демократия, договор, сотрудничество, вырасти, бесстрашный, благо, культовый, знаменитый, успех и др.

Таким образом, каждый документ корпуса был размечен по формуле:

где Pos означает количество положительно окрашенных слов, а Neg – отрицательно окрашенных слов (основываясь на списке, составленным до этого). Положительные и негативные слова учитывались столько раз, сколько они появлялись в тексте. Значение +1 определяет полностью позитивно окрашенную статью, 0 – нейтральную, а -1 – полностью отрицательную статью.

Далее строим график распределения тональности новостных сообщений за выбранный период:



*Рисунок 7. Распределение sentiment score по выбранному диапазон*

Этот график можно сопоставить с графиком изменения курса МосБиржи:

*Рисунок 8. График изменения курса акций МосБиржи*

Таким образом, можно сделать вывод, что выдвинутые гипотезы верны, так как график распределения тональности соответствует графику индекса МосБиржи. Следовательно, становится возможным написать программу автоматического предсказания индекса МосБиржи.

## 3.4. Реализация программы классификации новостей по тональности

Для написания программы был выбран язык Python 2.7 и библиотека Scikit-learn. Данная библиотека предоставляет реализацию алгоритмов как для обучения с учителем (supervised learning), так и обучения без учителя (unsupervised learning) как модуль языка Python. Библиотека удобна тем, что помимо реализации большинства алгоритмов, в ней есть также функции автоматического разделения выборки на обучающую и тестовую, перекрестной проверки (cross-validation), извлечения и отбора признаков (feature selection) и сокращения размерности (dimensionality reduction) [Субботин].

Корпус был разделен на обучающую и тестовую выборку с помощью встроенной функции train\_test\_split. Тестовая выборка составила 30 % от общего числа новостей, остальные сообщения вошли в обучающую выборку. Для извлечения признаков был использован TfidfVectorizer – метод, преобразующий текст в матрицу TF-Idf весов. При обучении не учитывались слова, которые встречаются больше чем в 95 % документах и менее чем в трех документах. Выборка была распределена на 3 класса: отрицательный, положительный, нейтральный. Отрицательных новостей было 1389, нейтральных – 869 и положительно окрашенных – 753.

Решение об использовании n-грамм было принято с помощью GridSearch – модели поиска, помогающей оценить униграммы, биграммы и триграммы по среднему значению (mean score) и z-значению (z-score, standard score). Триграммы получили высшие значения по обеим мерам:

 *Рисунок 9. Значения mean score и standard score для уни-, би- и триграмм*

Для оценки работы классификаторов использовались метрики точности (precision) и полноты (recall). Точность системы в пределах класса – это доля документов действительно принадлежащих данному классу относительно всех документов которые система отнесла к этому классу. Полнота системы – это доля найденных классификатором документов принадлежащих классу относительно всех документов этого класса в тестовой выборке [Баженов].

Значения точности и полноты рассчитываются по матрице ошибок:

*Рисунок 10. Матрица ошибок для категории (класса) i*

В матрице содержится информация о том, сколько раз система приняла верное решение или «ошиблась», где:

1. TP – True positive, истинно-положительное решение, то есть пример принадлежит классу и был правильно распознан системой
2. TN – True negative, истинно-отрицательное решение, то есть пример не принадлежит классу и не был отнесен к нему системой
3. FP – False positive, ложно-положительное решение, или ложное срабатывание. Данный документ не принадлежит классу, но был отнесен к нему системой.
4. FN – False negative, ложно-отрицательное решение, или пропуск события. Данный документ принадлежит классу, но система не отнесла его к нему.

Тогда, точность и полнота вычисляются следующим образом:

Для решения задачи были выбраны следующие алгоритмы: наивный байесовский классификатор и метод опорных векторов. Для метода опорных векторов было использовано линейное ядро и функция потерь hinge loss. Функция потерь характеризует величину отклонения ответа *y=a(x)* от правильного ответа на произвольном объекте . Hinge loss определена как ), где , а y – множество проставленных ответов (а не предсказанных ответов). При решении задачи методом наивного байеса была выбрана мультиномиальная (multinomial) модель. Мультиномиальный классификатор оценивает вероятность класса документа, опираясь на частоты слов, то есть в соответствии с униграммной моделью языка. В сочетании с tf-idf схемой весов были получены следующие значения метрик:

Таблица 1. Значения точности и полноты для программы классификации по тональности

|  |  |  |
| --- | --- | --- |
|  | precision | recall |
| Naïve Bayes | 0.79 | 0.62 |
| SVM | 0.80 | 0.69 |

Обе модели получили достаточно высокие результаты по мере точность, что позволяет использовать их в работе программы-предсказателя. Однако метод опорных векторов получил лучшие оценки, поэтому было решено использовать данную модель в финальной программе.

## 3.5. Реализация программы предсказания движения курса МосБиржи

В рамках данной работы программа машинного обучения должна предсказывать тенденцию движения индекса МосБиржи. Предсказание производится по двум классам – движение вверх (1) и движение вниз (0). Таким образом, решение данной задачи сводится решению задачи бинарной классификации.

При решении задачи использовалась модель классификации новостей по тональности SVM, полученная ранее. На вход программе подаются новости, вышедшие на n-й день для предсказания движения курса Мосбиржи на n+1-й день. Основываясь на модели классификации SVM определяется тональность каждой новости. Далее значения меток класса складываются, и на основе полученного числа принимается решение. При положительной сумме программа делает предсказание вверх (1), при отрицательной предсказание вниз (0).

Для эксперимента было подано 150 новостей в период с 1 по 20 апреля. Каждый день программе давалось разное количество новостей для определения эмоционального тона в СМИ, в зависимости от освещаемости тех или иных событий. Из 20 предсказаний верными оказались 16. Поскольку программа делала предсказания на коротком промежутке времени, сложно оценить ее работу. Однако для меры оценки было выбрано значение достоверности. Достоверность программы прогнозирования – это процент верных предсказаний. Для данной программы на текущей выборке достоверность составила 80 %.

Если говорить о более практическом применении программы, интересно рассмотреть следующее политическое событие. 6 апреля 2018 года, в пятницу, Ведомости выпускают статью о введении США новых санкций против российских миллиардеров и госменеджеров. В этот день программа сделала предсказание о движении курса МосБиржи вниз. Интересно, что участники биржи не придали сильного внимания этой новости, на финансовых форумах не было панических настроений. Однако уже в понедельник утром, 9 апреля, случился грандиозный обвал на бирже. Стоимость акций компании Русал, владельцем которой являлся О. Дерипаска, попавший под санкции США, упала более чем на 40 % на Гонконгской бирже и на 46,9 % на Московской бирже [BBC]. Падение акций Сбербанка превысило 20 %. На бирже в Лондоне снизились в цене бумаги почти всех российских компаний, в том числе "Лукойла", "Новатэка", НЛМК, "Фосагро", "Полюса Золота", "Роснефти", "Сургутнефтегаза". Из-за санкций инвесторы старались минимизировать риски и стали продавать акции компаний. В итоге, в конце дня индекс МосБиржи опустился на 9 % и достиг отметки в 2075,04 пункта.



*Рисунок 10. Падение индекса МосБиржи в ре-те введения новых санкций США*

Таким образом, программа позволила бы принять решение за 2 дня до случившихся событий и избежать огромных финансовых потерь инвестора.

# Заключение

В ходе данной работы были решены следующие задачи:

1. были выявлены и изучены практики российских СМИ в процессах отражения политических конфликтов и определена корреляция между настроениями в СМИ и ситуацией на фондовом рынке;
2. рассмотрены различные способы структурирования, приведения к единому формату и предварительной обработки текста: приведение к единому регистру, токенизация, стемминг, лемматизация, удаление стоп-слов;
3. изучены и описаны основные алгоритмы машинного обучения, применяемые при решении задачи классификации: наивный байесовский классификатор и метод опорных векторов;
4. реализована программа машинного обучения для классификации отзывов на два класса, использующая библиотеку scikit-learn; проведена классификация корпуса новостей, выбрана эффективная модель SVM;
5. реализована программа, предсказывающая на основании анализа тональности новостей движения индекса МосБиржи с интервалом в 1 день, которая показала 80 % уровень достоверности (программные коды исследования доступны по ссылке

<https://github.com/daranikolaeva427/sentimentstockpredict2018>).

Таким образом, впервые в отечественной лингвистике была предпринята попытка применения классификатора тональности политических новостей для финансового прогнозирования.

Дальнейшим развитием проекта является автоматиза ция подачи новостей на вход, так как сейчас это происходит вручную. Также предполагается агрегация новостей из разных новостных источников, для анализа медиадискурса не только качественными, но и количественными методами, что позволит проведение более глубоко анализа десигнативного и коннотативного компонентов. Более высокие результаты и точные предсказания может давать глубокий лингвистический анализ, вычленение ключевых предложений, выделение семантических ролей, определение тональности в пределах предложения. Кроме того, помимо анализа тональности в программу-предсказатель можно добавить технические индикаторы – значения открытия и закрытия торгов, это поможет увеличить точность системы.

# Список литературы

1. Абдуллин А. Р., Фаррахетдинова А.Р. Гипотеза эффективности рынка в свете теории финансов // УЭкС. 2015. №4 (76)
2. Азарова И.В., Гордеев C.C., Дёгтева А.В. Использование нейронной сети для автоматического определения тональности отзывов // Материалы XLIII Международной филологической конференции, 2014 — СПб., 2014 С. 3–9.
3. Гавра Д.П. Понятие и структура современной деловой журналистики // Сборник материалов Всероссийской научно-практической конференции. М. Факультет журналистики МГУ имени М.В. Ломоносова. МедиаМир. 2009. с.103, с. 112-144
4. Дейк Т. Ван. Язык. Познание. Коммуникация. // М. 1989. с.6
5. Добросклонская Т.Г. Медиалингвистика: системный подход к изучению языка СМИ. // М. 2008. с. 198
6. К.К. Боярский, Е.А. Каневский. Предсинтаксический модуль в анализаторе SemSin. // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), Санкт-Петербург, 9 — 11 октября 2013 г. СПб.: НИУ ИТМО, 2013, с. 281
7. Науменко Т.В. Коммуникативные процессы в глобальном экономическом пространстве// Credo New. 2015. №1. с.63
8. Паничева П. В. Система сентиментного анализа ATEX, основанная на правилах, при обработке текстов различных тематик // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). — М.: Изд-во РГГУ, 2013
9. Хохлова М.В. Анализ тональности // Прикладная и компьютерная лингвистика. М.: URSS, 2016
10. Чарахчян К.К., Чарахчян В.К. Некоторые аспекты воздействия новостной информации на биржевую оценку эмитентов // Теория и практика общественного развития. Краснодар: Хорс, 2018. № 1. С. 48-52
11. Ширяев А. Н. Основы схоластической финансовой математики // Соч.: том 2. М.: Фазис, 1998. с. 544
12. Aisopos, F., Papadakis, G., Varvarigou, T. Sentiment Analysis of Social Media Content using N-Gram Graphs // Scottsdale: WSM 2011*.*
13. Bing Liu. Sentiment Analysis and Opinion Mining // Morgan & Claypool Publishers, 2012. p. 23
14. Bing Liu. Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing // Morgan & Claypool Publishers. Second Edition, 2010. p.10
15. Bollen J., Mao H. and Zeng X. Twitter mood predicts the stock market // CoRR, abs/1010.3003, 2003
16. Domingos, P. & Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss // Machine Learning, № 29, 1997. pp. 103-­137
17. Fung, Gabriel, et. al. The Predicting Power of Textual Information on Financial Markets // IEEE Intelligent Informatics Bulletin. Vol. 5. No. 1. June, 2005
18. Gidofalvi G. Using News Articles to Predict Stock Price Movements. // Department of Computer Science and Engineering, University of California, San Diego: 2001
19. Hoeffding. W. The large-sample power of tests based on permutations of observations // The Annals of Mathematical Statistics, 1952. pp. 169–192
20. Huang W., Nakamori Y. et al. Forecasting stock market movement direction with support vector // Computers & Operations Research, Elsevier : 2005
21. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval 2(1-2), 2008
22. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002. pp. 417—424
23. Wang S., Manning C. Baselines and bigrams: Simple, good sentiment and topic classification // In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), 2012

**Электронные ресурсы**

1. Блог Игоря Субботина

// URL: <http://igorsubbotin.blogspot.ru/2015/01/intro-to-scikit-learn.html>

1. Ведомости // URL: <https://www.vedomosti.ru/business/news/2017/12/07/744505-rosneft-podala-sistema-novii-isk>

<https://www.vedomosti.ru/politics/news/2018/04/10/764399-tramp-rasskazal-o-vozmozhnom-voennom-otvete-vlastyam-sirii>

1. ВЕСТИ // URL: <https://www.vesti.ru/doc.html?id=2995634&tid=111662>
2. РБК // URL: <https://www.rbc.ru/business/07/12/2017/5a292f1c9a7947e2cf0a70f7>, <https://www.rbc.ru/business/07/12/2017/5a29315d9a7947fea4ea6f35>

<https://www.rbc.ru/business/07/12/2017/5a293b459a7947132066b2b8>

<https://www.rbc.ru/newspaper/2018/03/13/5aa64aba9a7947fe67942cf9>

<https://www.rbc.ru/opinions/politics/07/12/2017/5a293da39a794716addbe505>

1. Сайт Московской биржи // URL: <https://www.moex.com/ru/index/archive-data.aspx>
2. Сайт Юрия Баженова // URL: <http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html>
3. BBC // URL: <https://www.bbc.com/russian/news-43696073>
4. Coursera // URL: <https://www.coursera.org/learn/data-analysis-applications/lecture/6pei3/priedobrabotka-tieksta>
5. Datareview // URL: <http://datareview.info/article/otbor-priznakov-dlya-mashinnogo-obucheniya-na-python/>
6. EXPERT // URL: <http://expert.ru/2018/03/13/ubijstvo-v-britanskom-marazme/>
7. Finanz // URL: <https://www.finanz.ru/indeksi/sostav/micex>
8. Habrahabr // URL: <https://habrahabr.ru/post/105220/>
9. MachineLearning.ru // URL: <http://www.machinelearning.ru/wiki/index.php?title=Классификация>

<http://www.machinelearning.ru/wiki/index.php?title=Выборка>

<http://www.machinelearning.ru/wiki/index.php?title=Эмпирический_риск>

1. Mystem // URL: <https://tech.yandex.ru/mystem/>
2. MK // URL: <http://www.mk.ru/economics/2016/11/09/ekonomisty-tramp-snimet-sankcii-s-rossii-v-2017-godu.html>
3. Newsru // URL: <https://www.newsru.com/>
4. Scrapy // URL: <https://scrapy.org/>
5. Solarix // URL: <http://www.solarix.ru/for_developers/api/lemmatizator-api.shtml>
6. Wikipedia // URL: <https://ru.wikipedia.org/wiki/метод_опорных_векторов>
7. YOUTUBE // URL: <https://www.youtube.com/watch?v=dwkKrQHEvVg>

# Приложение 1

В приложении приведены примеры новостей, значительно оказавших влияние на фондовый рынок. Жирным выделены выражения, создающие эмоциональный фон новостных сообщений. После каждой новости указано ее отражение на ценах активов. Приведены яркие примеры как российских, так и иностранных событий.

1) Брексит

26 июня 2016

Ведущие мировые банки предлагают свои сценарии развития событий после британского референдума, на котором британцы проголосовали за выход страны из ЕС.

Большинство банков заявляют, что исход референдума оказался для них неожиданным. Все сходятся во мнении, что "брексит" и отставка премьер-министра страны Дэвида Кэмерона приведут к **росту неопределенности** на рынках и **массовому падению акций**.

Например, в банке Citi прогнозировали, что с небольшим перевесом в референдуме победят сторонники членства Великобритании в ЕС. Вероятность нынешнего исхода они оценивали в 40%. По мнению экспертов американского банка, рынки были **совершенно не готовы** к подобному исходу.

В Bank of America заявили, что на рынках, вопреки всем прогнозам и опросам, воплотился в жизнь один из основных **рисков**, который спровоцирует длительный **период волатильности**.

**Замедление экономики**

"Брексит" обернется **серьезными потерям** для британской и европейской экономики, уверены банкиры.

В Citi обещали радикально пересмотреть прогнозы по росту ВВП Великобритании и ЕС на ближайшие три года.

С этим согласны в Bank of America Merrill Lynch: высокий уровень неопределенности в отношении экономики Великобритании заставит инвесторов, в том числе внутренних, **отказаться от вложений** в страну. Экономисты банка прогнозируют, что в Великобритании после референдума **начнется рецессия**, а прогноз по росту ВВП на 2017 год они **снизили** с 2,3% до 0,2%. Прогноз по экономике еврозоны на следующий год в банке пересмотрели с 1,6% до 1,1% .

В Morgan Stanley прогнозируют **снижение потенциала роста** британской экономики в будущем, а также **повышение ее уязвимости**. По прогнозу банка, рецессии удастся избежать, если выход из ЕС будет мирным и беспроблемным. В противном случае британскую экономику ждет рецессия [BBC].

Очевидно, новость была громкой и неожиданной, что привело к панике инвесторов. 24 июня 2016 года фунт продемонстрировал самое быстрое падение в истории по отношению к доллару США и оказался на минимальном значении с 1985 года. Падение с 1,5 до 1,25 было очень неожиданным.

Более того, индекс FTSE (ведущий индекс Британской биржи) уже 27 июня упал на 5,6 %.

2) Выборы президента США

9 ноября 2016 года

**Дональд Трамп выиграл** президентские выборы  
Кандидат в президенты США **от Республиканской партии** Дональд Трамп одержал победу в президентской гонке, свидетельствуют подсчеты Associated Press и Bloomberg

Кандидат в президенты США от Республиканской партии Дональд Трамп **одержал победу** в президентской гонке, **победив демократа** Хиллари Клинтон. Об этом свидетельствуют данные ​Associated Press. Позже поражение признала и сама Клинтон.  
О **победе Трампа** также сообщает Bloomberg. У него 276 голосов выборщиков, следует из подсчетов агентства.  
СМИ дают свои прогнозы по исходу выборов, основанные на экзитполах и предварительном подсчете голосов. Для избрания президентом необходимо 270 голосов выборщиков. О том, что Трампу удалось набрать их, победив в достаточном количестве штатов, сообщает также The Washington Post, однако официальный подсчет голосов продолжается. Глава избирательного штаба Клинтон Джон Подеста ранее предложил дождаться утра и окончательных цифр.

Накануне выборов большинство социологических опросов показывали, что больше шансов на победу у Клинтон. Сама она не стала выступать перед сторонниками после того, как стало **очевидно лидерство Трампа**, еще накануне считавшегося менее вероятным победителем президентской гонки. [РБК]

Новость была неожиданной, как для США, так и для России, так как ожидалась победа Клинтон. СМИ сразу же стали приписывать Трампу образ пророссийского президента, было опубликовано множество новостей, связанных с ожиданием снятия санкций:

9 ноября 2016

Экономисты: Трамп **снимет санкции с России** в 2017 году

Финансовые рынки трепетно следили за ходом выборов президента США. Ведь речь идет о первой экономике мира и о судьбе доллара. Сегодня утром **с победой Дональда Трампа**стало понятно, что фондовые рынки большинства развитых стран отреагировали на итоги голосования значительным падением. То же касается и нефти, стоимость которой нырнула ниже $45. Курс рубля в начале торгов на внутреннем рынке также немного снизился. Между тем эксперты считают, что Трамп будет делать **шаги по снятию санкций** с России.

Экономисты уверены, что до января российская валюта будет находиться в подвешенном состоянии, но серьезного падения ее курса, скорее всего, не произойдет. А вот на нового президента США Трампа они **возлагают большие надежды.**

«Конечно, один человек не может развернуть внешнюю политику такой империи, как США. Но с республиканцами **Кремлю всегда легче**. Значит, можно рассчитывать **на компромисс как по Сирии**, так и по Украине», — считает эксперт.

Более того, по мнению главного экономиста датского Saxo Bank Стина Якобсена, Дональд Трамп **выступит за снятие с России санкций**, причем сделает это уже в 2017 году. «**Санкции будут сняты**, конечно не все, но часть точно уже в 2017 году», — полагает он [MK].

На фоне ожидания снятия санкций с России индекс МосБиржи вырос до 2293 и достиг исторического максимума. Однако, как уже было отмечено в параграфе 1.3, политические события имеют сильное, но кратковременное влияние на фондовые рынки. Из-за широкого освещения в СМИ будущей позитивной политики Трампа, реакция рынка была импульсивна. Тем не менее, индекс МосБиржи ждало резкое падение из-за неоправдавшихся ожиданий и к концу февраля 2017 года он вернулся к прежним значениям.

3) Торговые войны США с Китаем

22 марта 2018

Трамп подписал меморандум **о торговых мерах против Китая**

Американский президент Дональд Трамп подписал меморандум о торговых мерах в отношении Китая. Видео подписания документа размещено в YouTube-канале Белого дома.

Во время церемонии подписания документа Трамп заявил, что речь идет о поставках китайских товаров «примерно на $60 млрд» в год.

Трамп поручил торговому представителю США Роберту Лайтхайзеру в ближайшие 15 дней изучить **возможность введения таможенных пошлин** в отношении отдельных видов китайской продукции. Ему также поручено составить соответствующий список, инициировать **разбирательство** в рамках Всемирной торговой организации (ВТО) **против Китая за нарушения международных правил торговли**, говорится на сайте Белого дома.

Министру финансов Стивену Мнучину Трамп поручил в течение двух месяцев подготовить предложения **по ограничению инвестиций Китая** в американские компании.

«Это первый пакет мер, первый из множества», – сказал президент, подписывая документ.

Ранее Лайтхайзер рассказал Трампу об итогах расследования о предположительном **нарушении Китаем интеллектуальных прав** американских компаний. Расследование было начато еще в августе 2017 г. Американская сторона пришла к выводу, что **Китай использует нечестные торговые практики** для приобретения новейших технологий, которые потом используются китайскими компаниями для получения конкурентных преимуществ [Ведомости].

Эта новость нашла отражение на фондовых рынках США и Китая уже на следующий день:

23 марта 2018

США и Китай обменялись предупредительными залпами в торговой войне

**Фондовым рынкам это очень не понравилось**

Президент США Дональд Трамп в четверг вечером подписал меморандум, в котором поручил правительству принять меры по ограничению деятельности Китая, который Белый дом **обвиняет в нечестной торговле** и нечестных же попытках заполучить американские технологии путем устрашения, финансируемых государством поглощений и прочими хитрыми способами. Среди этих мер – импортные пошлины на поставки из Китая стоимостью $60 млрд, ограничения на приобретение американских компаний и передачу Китаю технологий, пишет The Wall Street Journal. Конкретные действия не будут предприняты еще минимум месяц, отмечает газета: в президентской администрации надеются, что Китай за это время согласится пойти на определенные уступки.

Трамп говорил, что ждет от Пекина предложений по сокращению американского дефицита в торговле товарами с Китаем, составляющего $375 млрд в год, на $100 млрд.

…

Китай же в пятницу сделал свой предупредительный залп в пока еще не начавшейся в полную силу торговой войне – объявил о планах ввести пошлины на американские товары, от фруктов до свинины, переработанной алюминиевой продукции и стальных труб, на сумму $3 млрд. Министерство торговли Китая назвало эти меры ответом на стальные и алюминиевые пошлины США, которые, по его мнению, нарушают правила международной торговли. США «создают опасный прецедент», заявил китайский минторг, пообещав, что Китай будет защищать свои интересы и готовить ответ на новые ограничительные меры США.

**Происходящее не понравилось участникам фондовых рынков**. Индекс Dow Jones упал в четверг на 2,9%, S&P 500 – на 2,5%. Сильнее всего упали акции компаний, чьи прибыли могут пострадать из-за введения 25%-ных пошлин на импорт стали и 10%-ных – на импорт алюминия. Акции таких производителей промышленных машин и оборудования, как Caterpillar, Deere и Boeing, упали на 5,7%, 4,1% и 5,2% соответственно. Бумаги сельскохозяйственных производителей Archer Daniels Midland (-1,6%) и Bunge (-1,3%) подешевели в ожидании ответных торговых мер Китая.

Индекс фондовых рынков Азиатско-Тихоокеанского региона MSCI Asia Pacific упал в пятницу на 2,6%, гонконгский Hang Seng – на 2,5%, китайский Shanghai Composite – на 4,7%. Китайские власти провели в пятницу интервенции для поддержки фондового рынка, сообщает Bloomberg со ссылкой на людей, знакомых с ситуацией. В Европе индекс Stoxx Europe 600 потерял 0,9%. Американские индексы в пятницу закрылись падением на 1,8% (Dow Jones) и 2,1% (S&P 500); падение по итогам недели оказалось самым значительным более чем за два года. Индекс акций развивающихся рынков MSCI Emerging Markets снизился на 2% [Ведомости].

2 апреля 2018 года Китай предпринимает ответные меры:

**Повышенные пошлины** будут распространяться на 128 наименований, импортируемых в Китай из США. На 120 наименований, в том числе фрукты, пошлина составит 15%, на 8 наименований, в том числе свинину, — 25%. Решение вступает в силу 2 апреля.

«**Принятая мера** призвана восполнить ущерб, нанесенный нашей стране введением в США пошлин на стальную и алюминиевую продукцию из КНР», говорится в сообщении Министерства финансов Китая [БКС].

После принятых мер индекс S&P (ведущий американский индекс) упал на 110 пунктов. Данный пример снова показывает, как политические события в кратковременной перспективе превалируют над экономическими.

4) Рост цен на нефть в РФ после событий 9 апреля 2018 года

10 апреля 2018

США быстро и решительно ответят на предполагаемую **химическую атаку** в сирийском городе Дума. Об этом американский президент Дональд Трамп заявил на встрече с военачальниками и советниками по национальной безопасности своей администрации, передает Reuters.

У Штатов есть «множество вариантов военного ответа», сказал Трамп. Он пообещал принять решение к вечеру 9 апреля или вскоре после этой даты. «Мы не можем позволить **зверствам**, которым мы были свидетелями <...> мы **не можем позволить происходить** этому в нашем мире <...> особенно, когда в силах Соединенных Штатов, мощь нашей страны позволяет это остановить», – подчеркнул Трамп.

США считают ответственными за возможную химическую атаку в Думе в Восточной Гуте правительственные войска президента Сирии Башара Асада. Сирийские власти **отрицают свою вину** и называют сообщения о химатаке недостоверными. МИД России заявил, что российские военные специалисты не обнаружили следов применения химического оружия в Думе.

В понедельник Трамп также заявил, что если президент России Владимир Путин причастен к химатаке, то **реакция США будет жесткой**. «Он [Путин] может, да, может. И если он это сделал, [ответ] будет очень жестким», – подчеркнул Трамп, отвечая на соответствующий вопрос журналистов.

Ранее постпред США при ООН Никки Хейли пообещала, что Вашингтон **ответит на химическую атаку** независимо от того, какое решение примет Совбез.

О химической атаке в пригороде Дамаска 7 апреля заявила сирийская оппозиция и сирийско-американское медицинское общество. По неподтвержденным данным, в результате инцидента могли **погибнуть** около 60 человек, сотни людей получили ранения [Ведомости].

Это событие очень интересно разбирать. После невероятного падения росрынка 9 апреля 2018 (см. параграф 3.5) цены на нефть стали резко расти, так как во-первых, рос курс доллара, а во-вторых, увеличились ожидания нестабильности на Ближнем Востоке. Соответственно, Роснефть выросла с 284 руб. до 410 руб., другие нефтяные компания (Татнефть, Сургутнефтегаз, Газпром) после 9 апреля также восстановили прежние значения и прибавили 10 %.