Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Московский физико-технический институт
(государственный университет)»

Факультет управления и прикладной математики

Кафедра информатики и вычислительной математики

ФОРМИРОВАНИЕ СЕРВИСА ПОСТРОЕНИЯ ОБРАЗА АВТОРА КАК КОГНИТИВНАЯ ТЕХНОЛОГИЯ

Магистерская диссертация

Направление подготовки: 011600 Прикладные математика и физика

Выполнил:
студент 176а группы Родина София Викторовна

Научный руководитель:
к.ф.н. Рыков Владимир Васильевич

Москва 2017

Содержание

1 Введение

Огромное число людей в развивающихся странах в настоящее время имеют доступ к сети Интернет, благодаря сочетанию снижения затрат и экспоненциального совершенствования технологий, которые используются для создания ноутбуков, смартфонов и планшетных устройств. Их можно купить всего за несколько десятков долларов. Также произошел невероятный рост доступности мобильных сетей. Даже некоторые из наиболее отдаленных поселений на Земле могут воспользоваться Интернетом, благодаря развитию инфраструктуры на местах. В этом году количество пользователей интернета превысило 5 млрд. человек, вместе с тем растет и число пользователей различных социальных сетей для обмена сообщениями, ведения блогов, трансляции фото- и видеоданных. Число активных пользователей крупнейшей соцсети в мире - Facebook - достигло почти 2-х млрд. человек в месяц. На фоне такого активного роста аудитории различных интернет-сервисов появляется огромное количество пользовательских данных, которые находятся в публичном доступе и могут быть полезны для исследований в самых различных областях.

Актуальность. Задача построения образа пользователя становится все более актуальной с неукротимым ростом интернет-аудитории. Интернет-гиганты, такие как Google и Facebook, владеющие различными сервисами, имеют возможность аккумулировать различную информацию о человеке, исходя из личных данных и активности, которую пользователь проявляет в интернете и использовать эти данные для разного рода исследований. Публичность большого количества данных также позволяет проводить различные эксперименты независимым исследователям.

Новизна. Данная область исследования является новаторской и число подобных исследований невелико. Системы построения образов только начинают свой путь среди когнитивных сервисов. Тем не менее, уже сегодня компания IBM предлагает подобный продукт под названием Personality Insights [3], который может быть использован для создания образа автора.

Практическая значимость. Исследования, проведенные компанией Facebook показали, что принимая во внимание личные характеристики пользователя(вычисленные путем предварительного опроса), национальные особенности, политические убеждения и прочие данные, конверсия рекламы может быть увеличена многократно. Данное исследование было проведено в тестовом режиме и его результаты стали неожиданностью для многих. Другим

интересным примером использования персонализированных данных являются алгоритмы предложения новостей. По некоторым данным [4], подобная система была опробована на последних президентских выборах в США, что в свою очередь могло отчасти вызвать столь непредсказуемый для многих аналитиков результат выборов. Таким образом, более точная персонализация рекламных и новостных предложений являются подтвержденными приложениями для систем построения образов.

Построение образа автора в данном исследовании будет проведено средствами вычислительной лингвистики и анализа данных.

2 Вычислительная лингвистика и образ автора

2.1. Предмет вычислительной лингвистики

Вычислительная лингвистика(компьютерная или математическая лингвистика) - научное направление в области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллекта, которое ставит своей целью использование математических моделей для описания естественных языков.

Компьютерная лингвистика частично пересекается с обработкой естественных языков. Однако в последней акцент делается не на абстрактные модели, а на прикладные методы описания и обработки языка для компьютерных систем.

Полем деятельности компьютерных лингвистов является разработка алгоритмов и прикладных программ для обработки языковой информации.

Направления компьютерной лингвистики:

- Обработка естественного языка(англ. *natural language processing*; синтаксический, морфологический, семантический анализы текста). Сюда включают также:
    - Корпусная лингвистика, создание и использование электронных корпусов текстов.
    - Создание электронных словарей, тезаурусов, онтологий. Например, Lingvo. Словари используют, например, для автоматического перевода, проверки орфографии.
    - Автоматический перевод текстов.
    - Автоматическое извлечение фактов из текста(извлечение информации).
    - Автореферирование(англ. *automatic text summarization*).
    - Создание вопросно-ответных систем.
- Оптическое распознавание речи.
- Автоматический синтез речи.

В данной работе будет рассматриваться подход к обработке естественного языка с целью получения информации об авторе данного текста или текстов. Задача построения образа автора возникла как следствие накопления большого

количества пользовательских данных в сети Интернет. Базами данных такого рода сегодня обладают многие крупные компании, такие как Google, Facebook, Apple и многие другие. Помимо этого, большое количество пользовательских данных находится в публичном доступе и могут быть использованы независимыми исследователями для своих целей. В рамках данной работы будет произведена попытка построения личностного портрета пользователя в сети по блогам, которые он ведет. Принципиальным ограничением используемой модели является качество вводимых текстов, а именно тексты, используемые в исследовании обязательно должны отражать точку зрения их автора, т.е. следует избегать объемного цитирования и чересчур формального языка. Подобные исследования могут быть полезны для любых приложений, где учет личного портрета человека может улучшить качество получаемого результата.

2.2. Человек - часть информационной системы.

Человек давно уже является частью информационных систем. Конкретно в инновационных проектах необходимо знать объективные характеристики его участников или их образы. Этим определяется актуальность этой темы.
Знание принципиально отличается от информации. При выполнении информационных процессов информационная система (ИС) находит заданную ей по запросу информацию. При выполнении когнитивных процессов (КП) генерируется новое знание. С когнитивной информатикой (КИ) связаны большие надежды на то, что при помощи ее моделей можно будет находить решения различных проблем, недоступных человеку в силу самых разных причин. Начиная от тривиальных, но необходимых, – например, обсчет алгоритмов эвакуации из большого строящегося здания и до нахождения решений проблем глобальной экологии.

Нахождение решений различных когнитивных проблем осуществляется обычно в рамках определенных когнитивных моделей (КМ). В данной статье рассматривается модель автоматического нахождения или генерации образа автора (ОА) из некоторого корпуса авторских публикаций. Образ автора – риторический термин. Эта проблема является во многих отношениях новой и актуальной. Представить себе автора текста интересно многим читателям. Но сейчас человек давно стал частью многих глобальных ИС (например, соцсетей) и поэтому также важно знать хотя бы некоторые качества их участников [1, 2].

2.3. Образ автора(ОА).

Само определение образа автора имеет различные значения в различных

областях. Рассмотрим некоторые из них.

2.3.1. Образ автора в риторике.

Риторика развивалась Аристотелем, написавшим один из наиболее древних и авторитетных трактатов о ней. Но он же общепризнан как основатель логики, при описании которой он исходил из задач убеждения слушателя. И здесь возникает одно из древнейших заблуждений, что правильная, то есть убедительная (с точки зрения риторики) речь должна быть обязательно логичной. Она может быть логичной, почти логичной, паралогичной или алогичной – выбор осуществляется исходя из главного постулата - эффективности. Если нелогичная речь, обращенная к женщине с просьбой выйти замуж, достигнет нужного эффекта, а строго логичная - нет, то кто выберет логичную? Это уже скорее риторический вопрос. Но это в то же время реальная и массовая практика.

Риторика - это словесное (знаковое) и страшное оружие в руках обученного человека. Это прекрасно знали в древности, преподавая риторику вместе с основами этики в древности или богословия в средневековье. И не только в отношении логики. По отношении к истине и поискам истины риторика всегда противопоставлялась диалектике. Риторика ставила во главу угла задачу убедить человека любой ценой. Скажем, убедить преподавателя на экзамене, что знаешь предмет. Или, например, можно ставить цель убедить ученый совет в том, что проделана ученая работа на достаточно высоком уровне. Тогда как другая наука, диалектика, по представлению древних, наоборот, преследовала цель совместного поиска истины. Примером этого могут служить диалоги Платона, научные семинары (быть может, в идеализированном варианте). Печально, хотя это часто бывает, что поиски истины (диалектика) иногда подменяются неистовым стремлением убедить собеседника в своей правоте.

Следовательно, нужно иметь в виду, что обе науки – риторику и диалектику (в античном понимании этого термина) - можно и нужно рассматривать в гораздо более обобщенном контексте, как знаковые технологии эффективной интеграции разных видов деятельности, в зависимости от обстоятельств их реализации. В случае риторической технологии доминирующей является цель воздействия на получателя (заказчика, инвестора, слушателя) с нужным эффектом. Для диалектики (в античном понимании этой науки или технологии), доминантой является истинность, понимаемая опять же в зависимости от условий и обстоятельств. Остается добавить, сделав еще один неизбежный экскурс в античность, что риторика как легальная и благородная наука убеждения противопоставлялась эристике, науке о том, как переспорить любой ценой. А диалектика имела свою неприятную противоположность – софистику,

в которой бескорыстные поиски истины заменяются абстрактными и бесполезными рассуждениями.

Как ни покажется странным, но к вопросам об истинности и логичности риторических технологий, по самой своей идее принципиально независимым от них, что и давало им необыкновенную широту, гибкость и практичность, а забвение этих постулатов привело риторику к гуманитарной катастрофе в новое время, тесно примыкает первая, также почти сразу забытая в средние века собственно риторическая категория – образ автора или образ ритора.

Речь идет именно об авторе (индивидуальном или коллективном) любого знакового произведения. Для того, чтобы сразу понять его важность и необходимость принимать во внимание, нужно представить, что одну и ту же команду в армии отдает сержант и полковник. Если кто-то возразит, что это слишком абстрактно или некорректно, то можно напомнить, что армия пока еще не абстракция. В ее рамках строят свою жизнь огромное количество людей, которых живо интересует эффект их деятельности (так же как и государство, которое их оплачивает). Игнорирование риторикой этих реалий деятельности самых разных людей в разнообразных ситуациях привело к тому, что саму риторику многие люди до сих пор рассматривают как ненужную абстракцию.

Однако, образ автора далеко не простая категория и не сводится к званию человека, который отдал команду. Эта категория была реконструирована и подробно описана в трудах академика В.В. Виноградова. Там было показано, что образ автора – неотъемлемая составляющая любого знакового произведения, в том числе литературного. Быть может не зная в явном виде об этой категории, ее мастерски использовал юный Лермонтов в романе «Герой нашего времени». Там впервые в мировой литературе (так, во всяком случае, утверждает В. Набоков в своих лекциях по русской литературе) образ автора предстает в динамике. Действительно, в серии отдельных повестей, составляющих этот роман, образ Печорина постепенно приближается к читателю. Этим достигается дополнительный художественный эффект. Не зря бабушка тратила на обучение и воспитание Лермонтова половину доходов от своего имения.

Для работников информационных технологий более понятными могут показаться примеры о том, что примерно одни и те же по своим возможностям программные продукты охотнее покупают у авторитетных фирм. Можно также мысленно сравнить эффект от одной и той презентации, сделанной представителем высшего руководства фирмы и простым инженером. Но, независимо от должности, современные журналы по ИТ технологиям единодушно отмечают ужасающе низкий их уровень почти по всем показателям, которые по своей сути являются риторическими. А ведь презентация – summa summarum деятельности фирмы или отдельного ее проекта [1].

2.3.2. Образ автора в Интернете(профиль пользователя).

Образом автора в сети может считаться совокупность объективных(измеряемых) характеристик о человеке и некоторый личностный(психологический) портрет. К объективным характеристикам относятся возраст, цвет глаз, цвет кожи, увлечения и прочие факторы, которые могут быть однозначно оценены. Личностный же портрет пользователя является характеристикой более эмпирической, но может оказаться намного более полезным, поскольку в большой степени позволяет предсказать поведение индивида в тестируемых ситуациях. Существует большое количество подходов к сбору таких данных. Объективные характеристики могут быть собраны из профилей социальных сетей, где в отдельных полях информация доступна и хорошо структурирована. Методы анализа изображений могут также быть полезны для определения внешних характеристик пользователя. Для определения личностного же портрета человека необходимы более сложные подходы, поскольку не всегда есть открытый доступ к результатам психологического тестирования, а при необходимости проанализировать большую базу пользователей возникает проблема того, что далеко не каждый даже теоретически может согласиться на прохождение таких тестов. Вследствие этого, возникает необходимость придумывать обходные варианты для получения личностных данных. Может быть использовано практически что угодно - от попытки считать скорость набора текста или скорость кликов на сайте до чтения истории в браузере или чтения личных блогов. Каждый из этих подходов имеет свои плюсы и минусы, но применяя их разумно можно добиться большой продуктивности в решении поставленной задаче.

Далее будет термин образа автора, используемый в рамках данной работы.

2.3.3. Образ автора в работе.

Определение. Образом автора мы будем называть совокупность личностных характеристик вычисленных для модели “Большой пятерки” [5].

Большая пятерка - диспозициональная(от англ. disposition - предрасположенность) модель личности человека. Продолжает линию исследований, начатую Г. Олпортом, Г. Айзенком и Р. Кэттелом, предполагавшими, что личность характеризует меру индивидуальных различий человека в степени и форме адаптации к социальной среде с учетом биологических свойств индивида.

В соответствии с названием, модель предполагает, что личность человека включает в себя пять общих и относительно независимых черт(диспозиций):

- экстраверсию
- доброжелательность(дружелюбие, способность прийти к согласию)
- добросовестность(сознательность)
- нейротизм(противоположный полюс - эмоциональная стабильность)
- открытость опыту(интеллект)

Модель выводится эмпирически, с использованием данных самоотчетов(вопросники, шкалы прилагательных), экспертных оценок(внешних наблюдателей за поведением) и данных поведенческих, получаемых в рамках исследований. Основным статистическим средством извлечения модели является эксплораторный факторный анализ. Таким образом, в эмпирических исследованиях 5 черт чаще всего предстают в виде сравнительно автономных факторов.

Таким образом, в данной работе будет предпринята попытка получения только психологических характеристик пользователя. В качестве подхода к получению этих характеристик будет использовано чтение блогов пользователей.

2.4. Обзор существующих работ по теме.

В данном разделе рассматриваются конкретно две когнитивные модели фирмы IBM – Персональный взгляд (Personality Insights) и Анализатор тона (Tone Analyzer). Обе модели или сервиса генерируют новую и неочевидную информацию об авторе. Однако между этими моделями есть различия т. к. они предназначены для получения несколько разных образов автора. Модель Персональный взгляд строит статичный образ автора (ОА), а Анализатор тона строит его конкретный тон или динамическую характеристику, реализованных при написании данного корпуса исходных текстов и даже выявить эмоциональное состояние автора [1, 2].
Итак, поставленную задачу можно декомпозировать на две более простые подзадачи:
● Обработка текстов на естественном языке и анализ полученного текста, вычленение потенциальных характеристик личности автора
● Построение модели личности автора по полученным характеристикам.
Если первая подзадача решается математическими или иными подходами строго формальными, то решение второй подзадачи лежит в большей степени в плоскости психологии. Тут также для анализа результата используется модель “Большая пятерка”.

Модель Персональный взгляд (Personality Insights) сервис IBM, который способен извлекать и анализировать персональные характеристики личности, помогая больше узнать о человеке и настроить с ним более персонализированное взаимодействие. Сервис делит выводимые характеристики на 3 группы: пятерка личностных характеристик, ценности и нужды. Для работы Personality Insights требуется минимум 1200 слов любого текста, написанного данным человеком.

Модель Анализатор тона использует лингвистический анализ для поиска трех основных характеристик: эмоции, социальные тенденции и стиль написания. Все эмоции делятся на: злость, страх, радость, грусть и раздражение. Социальные тенденции включают элементы из Большой пятерки персональных характеристик используемые некоторыми психологами. Стили написания делятся на: *уверенный, аналитический и осторожный.* Сервис может анализировать электронные письма и любые другие тексты. Результат работы сервиса можно использовать для определения того, несет ли письмо нужную заказчику эмоциональную окраску или другую характеристику. Сервис использует для анализа модель Большая пятерка.

Вот что было получено для портрета папы Римского:

Экспрессивный и уверенный в себе человек. Он легко заводит знакомства и уверенно чувствует себя в компании. Данный человек не очень консервативен и имеет собственную точку зрения по всем вопросам. Он стремится к самосовершенствованию и пытается проявлять себя профессионалом в нужный момент.

Можно ли проанализировать полученный результат? С этой точки зрения, результат работы сервиса можно охарактеризовать как предельно верный, но достаточно расплывчатый. Можно предположить, что любой известный в каких-либо кругах человек обладает уверенностью в себе,
общителен и безусловно разбирается в выбранной области [1,2].

Модель Анализатор тона также предоставляет несомненный практический интерес. Например, диалог разговор агента сервис центра с клиентом (Customer service chat). Запустить данный пример можно на официальной странице сервиса в разделе демоверсия. В данном диалоге клиент связывается с агентом сервис центра по вопросу удаления аккаунта, зарегистрированного на его почтовый адрес, но не принадлежащего ему. Агент сервис центра долго не может решить проблему, предлагая порой недопустимые решения. Данный диалог носит сердито-извинительный характер. В результате, видно, что главные эмоции в этом разговоре гнев и страх. Стиль общения преимущественно аналитический, строгий, формальный [1,2].

Представленные когнитивные модели генерации образа автора представляют собой несомненный научную и практическую значимость.

2.5. Применение на практике.

Исследования, проведенные компанией Facebook показали, что принимая во внимание личные характеристики пользователя(вычисленные путем предварительного опроса), национальные особенности, политические убеждения и прочие данные, конверсия рекламы может быть увеличена многократно. Данное исследование было проведено в тестовом режиме и его результаты стали неожиданностью для многих. Другим интересным примером использования персонализированных данных являются алгоритмы предложения новостей. По некоторым данным [3], подобная система была опробована на последних президентских выборах в США, что в свою очередь могло отчасти вызвать столь непредсказуемый для многих аналитиков результат выборов. Таким образом, более точная персонализация рекламных и новостных предложений являются подтвержденными приложениями для систем построения образов.

3 Построение образа автора

В данной главе будут изложены алгоритмы и принципы построения сервиса формирования образа автора(СФОА). В качестве входных данных сервис принимает набор текстов, главным требованием для которых является выражения личного отношения к вещам, личного мнения, т.е. ухудшить результаты работы данного сервиса может объемное цитирование или чересчур формальный безэмоциональный язык. В основе построения данного сервиса лежит алгоритм определения эмоциональной окраски предложения. Классификация предложений проводится в соответствии с таблицей Экмана [6], т.е. каждому из предложений присваивается один из следующих тэгов: *радость*, *грусть*, *удивление*, *злость*, *отвращение*, *страх*. После классификации всех предложений текста, применяется алгоритм обобщения полученных данных и формируется личностный портрет автора. Все исследования в данной работе проводятся с текстами на английском языке и, как следствие, используются правила английской грамматики.

3.1. Методология

Весь процесс обработки входного текста/текстов состоит из нескольких этапов:

1. Построение ERR(Emotion Recognition Rule) для набора данных для обучения.
2. Построение 3-х ступенчатого классификатора.
3. Классификация входных ERR [9].

3.2. Построение ERR

Целью данной фазы является построение множества ERR(Emotion Recognition Rule) для набора данных для обучения. Для построения этого множества

необходимы специально подготовленные для обучения данные, в которых каждое из предложений вручную размечено одной из эмоций по классификации Экмана. Каждое из размеченных предложений обрабатывается модулем обработки предложений для построения соответствующего ERR(Рис.1).

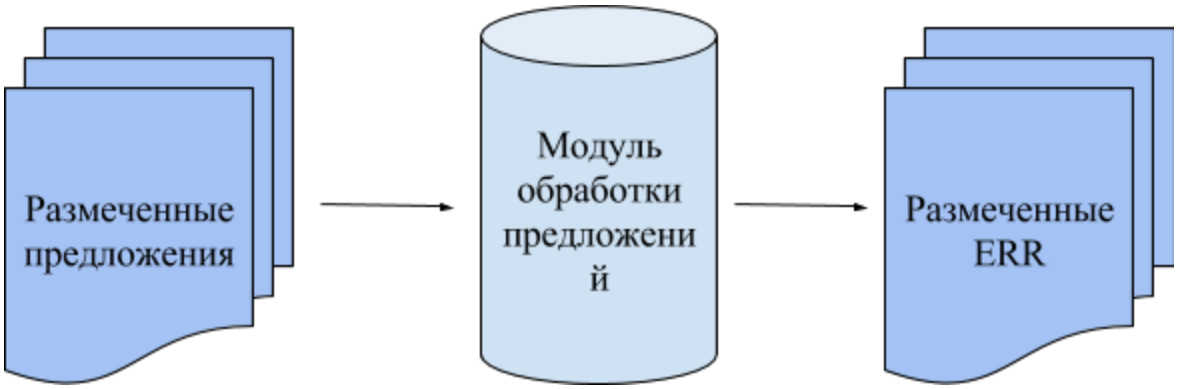

Рис.1

Для построения ERR каждое предложение проходит 3 этапа:

1. **Определение части речи каждого из слов.**

   Определение части речи каждого из слов происходит с использованием Stanford POS Tagger [7], поставляемого внутри пакета nltk.

   Используемые части речи и система обозначений:

   - PRP - личное местоимение
   - NN - существительное
   - JJ - прилагательное
   - IN - предлог
   - VB - глагол
   - RB - наречие
   - RP - частица или подчинительный союз
   - WP - вопросительное слово(wh-pronoun)

- CC - сочинительный союз

Данная система обозначений является стандартной для Stanford POS Tagger.

2. **Построение зависимостей в предложении.**

   Построение зависимостей в предложении реализуется путем использования Stanford Dependency Parser [8], который строит дерево зависимостей предложения.

   Вкупе эти два этапа приводят к построению графа зависимостей в предложении(Рис.2).

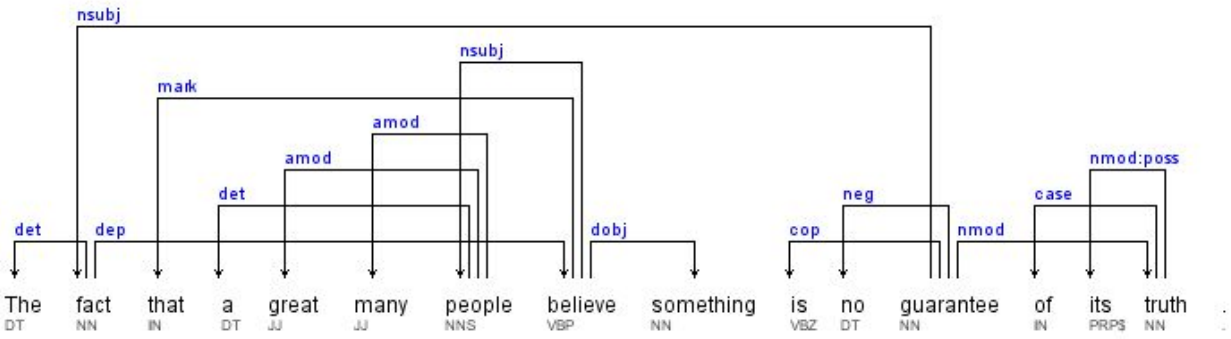

Рис.2. Пример графа зависимостей.

3. **Применение множества правил, удаляющих безэмоциональные части предложения.**

   Целью данного этапа является выделение эмоционально окрашенной части предложения.

Все применяемые правила делятся на 2 категории:

1. **Разделяющие правила.**
    a. **Разделяющие правило #1:**

Удалить часть предложения перед противительными союзами(*but* и другие). Поскольку противительные союзы подразумевают противопоставление, часть предложения после этого союза замещает эмоции, присутствующие в предыдущих частях предложения. Рис.3 показывает граф зависимостей предложения "*It was a bit complicated but we had fun*".

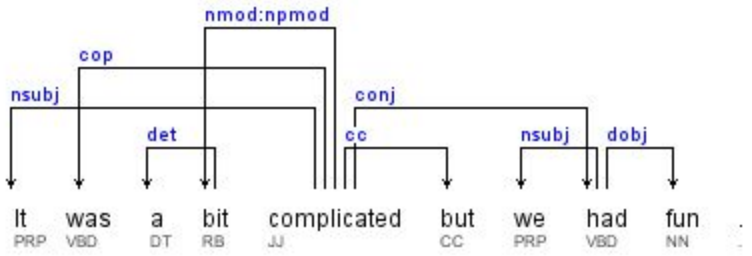

Рис.3

После удаления части до союза "but" остается следующий граф(Рис.4).

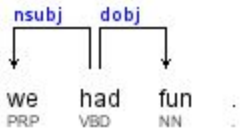

Рис.4

    b. **Разделяющие правило #2:**

Удалить часть предложения после слова *"as"* если после него стоит

местоимение. Такое же правило применяется к словам имеющим сходное с *“as”* значение. *“As”* является подчинительным союзом, что означает, что предложение после *“as”* является подчинительным. Подчинительное предложение может рассматриваться как дополнение к основному смыслу предложения и, следовательно, может быть удалено. Например, предложение *“People stare as I run”* будет рассматриваться как состоящее из двух частей: *“people stare”* и *“as I run”*. Однако, вторая часть будет удалена в соответствии с данным правилом.

2. **Удаляющие правила.**

    a. **Удаляющее правило #1:**

Удалить глагол, если у него нет объекта действия и он относится к wh-слову, поскольку он может рассматриваться как дополнение к основному эмоциональному значению предложения. Рассмотрим, например, предложение *“Where you are going is a disgusting place”* и его граф зависимостей. Часть *“where you are going”* будет удалена из графа зависимостей, который теперь будет содержать только часть *“disgusting place”*(Рис.5, Рис.6).

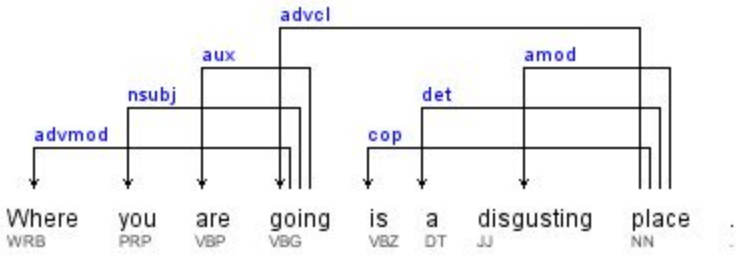

Рис.5

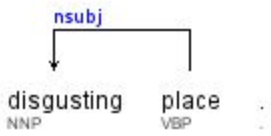

Рис.6

b. Удаляющее правило #2:

Удалить глагол, если он является или неэмоциональным или формой глагола *“быть”(“to be”),* поскольку он также может быть рассмотрен в качестве дополнения к эмоциональному значению предложения. Для того, чтобы определить имеет глагол эмоциональную окраску или нет используются WordNet-Affect [10], SentiWordNet [11] и эмоциональная вероятность глагола из множества для обучения. Мы считаем, что глагол имеет эмоциональную окраску, если он либо существует в множестве WordNet-Affect, либо имеет эмоциональную полярность в SentiWordNet, либо его эмоциональная вероятность(полученная из множества для обучения) выше некоторого порогового значения. Например, в предыдущем предложении *“we had fun”* будет минимизирована до *“we”* и *“fun”*.

c. Удаляющее правило #3:

Удалить местоимения если они не связаны с другими узлами графа. Например, если это правило было применено к предыдущему графу для предложения *“it was a bit complicated but we had fun”*, единственным оставшимся узлом графа будет *“fun”*, который и будет являться ERR для данного предложения.

Таким образом, то, что остается от предложения после применения всех

вышеуказанных правил и называется ERR - правилом распознавания эмоции.

3.3. Построение классификаторов.

Целью данного этапа является сравнение двух ERR, одно из которых представляет входящее предложение, второе - предложение из множества тестовых данных. На Рис.7 представлена общая структура сервиса распознавания эмоций в предложении.

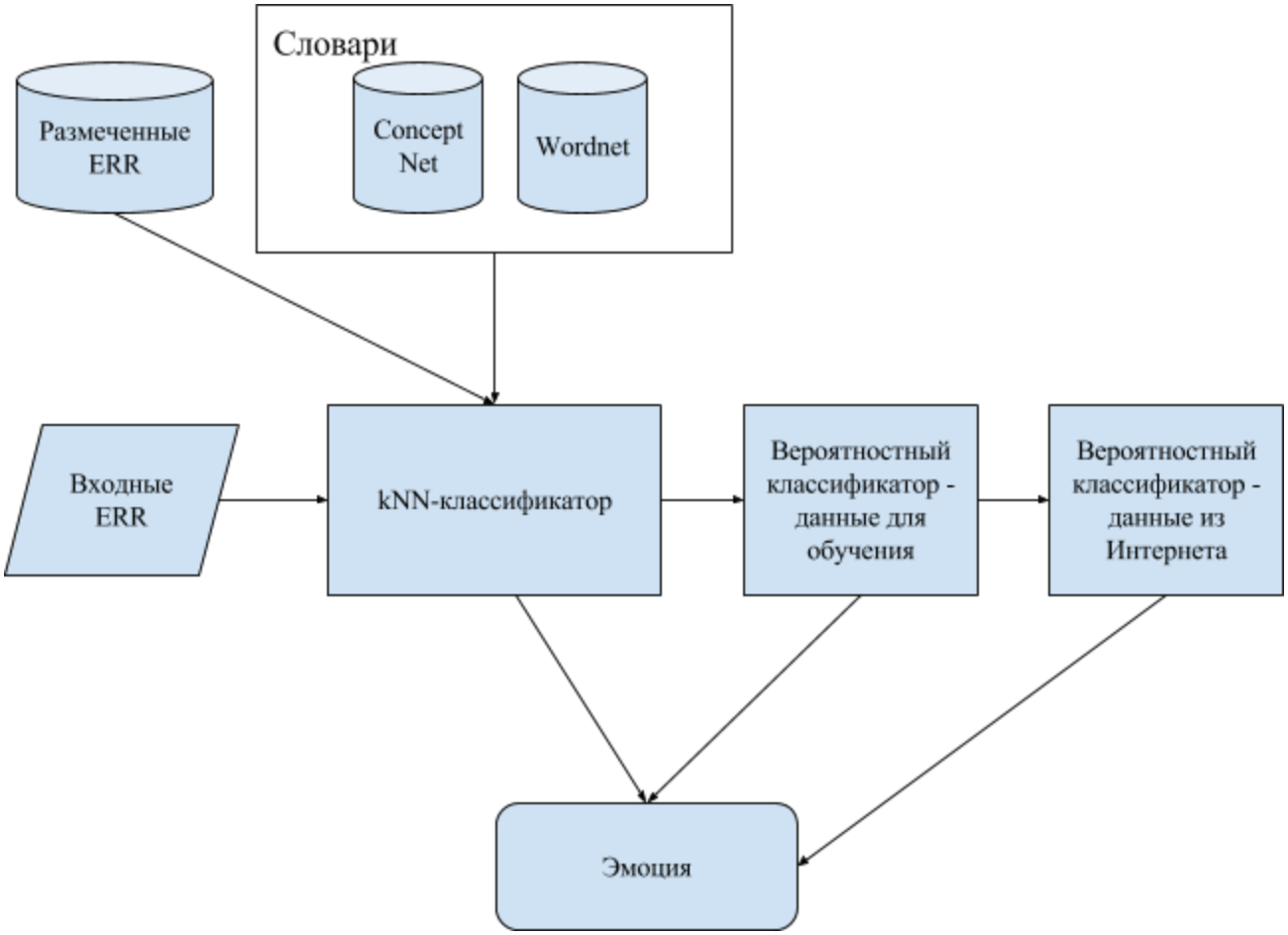

Рис.7

Для каждого входного предложения система строит соответствующее ERR используя те же шаги, что и для данных для обучения. Построенное ERR сравнивается с каждым размеченным ERR из данных для обучения с использованием вариации kNN-алгоритма(алгоритм k ближайших соседей).

Эмоцией входного предложения будет являться эмоция размеченного ERR, с которым было обнаружено максимальное сходство(для 1NN алгоритма). Интуитивный подход здесь заключается в том, чтобы найти ERR, которое которое похоже по структуре и значению с входным ERR. KNN-классификатор использует словари WordNet и ConceptNet для обобщения и сравнения ERR. Чтобы измерить подобие между входной ERR и ERR из набора для обучения, мы строим kNN-классификатор, основанный на двух измерениях: семантическое подобие и подобие по ключевым словам. Семантическое подобие определяет насколько два ERR схожи в своих значениях, в то время как подобие по ключевым словам определяет количество схожих по значению слов в этих ERR. Выбранное в итоге ERR - это правило, которое имеет максимальное семантическое сходство с входным ERR.

В результате эмоция либо была классифицирована, либо была отвергнута классификатором(подобие с любым размеченным ERR равно нулю), тогда классификация проводится методами вероятностной классификации.

3.3.1. KNN-классификатор.

Для вычисления подобия между двумя ERR используется функция подобия следующего вида:

$$F_{similarity} = Sim(VerbNounClauses) + Sim(NounClauses) + Sim(AdjectiveClauses) + \\ + Sim(AdverbClauses)$$

Мы вычисляем схожесть каждой пары однотипных(по части речи) слов путем сравнения двух множеств синонимов для каждого из этих слов. Данное сравнение проводится с использованием технологии WordNet. Для каждой пары схожих по значению слов функция подобия этих ERR увеличивается на единицу. Мы также используем ConceptNet для сравнения законченных фраз

или устойчивых выражений. Полученное значение для сравнения двух фраз лежит в интервале от нуля до единицы и также добавляется к итоговой функции подобия для двух ERR. В KNN-классификации мы также делаем различие между семантических подобием и подобием по ключевым словам. Функция *Sim()* возвращает числовое значение вместе с типом подобия; либо подобие по ключевым словам, либо семантическое. Однако, если одна из *Sim()*-функций возвращает семантический тип подобия для своих аргументов, общее подобие также будет считаться семантическим, иначе - подобие по ключевым словам. Далее будут разъяснены принципы работы KNN-классификатора, но прежде небольшой обзор технологий WordNet и ConceptNet, которые будут активно использоваться в работе.

3.3.1.1. WordNet

WordNet - это система, позволяющая получить набор синонимов для данного слова. Мы сравниваем пару слов путем сравнения их множеств синонимов, чтобы добавить элемент обобщения. Если пара слов является синонимичной, то к итоговой функции подобия добавляется единица.

3.3.1.2. ConceptNet

ConceptNet - это система, позволяющая проводить сравнение фраз. Важно, что схожие по значению фразы могут иметь разную грамматическую структуру, что ставит дополнительную подзадачу определения грамматических альтернатив для каждой введенной фразы. После сравнения двух фраз мы добавляем значение, полученное из ConceptNet к итоговой функции подобия.

Для сравнения фраз мы используем предварительно сформированные паттерны для типов фраз. Так, например, фраза имеющая структуру JJ NN(прилагательное с существительным) имеет в качестве альтернативы всякую фразу, имеющую

структуру типа JJ? NN - знак ? означает, что прилагательное может встречаться а может и не быть в структуре фразы.

Все правила альтернатив, используемые для сравнения фраз в данной работе:

- JJ NN = JJ? NN
- VB NN = JJ NN? | VB NN
- NN = JJ | NN | JJ NN

Для повышения точности сравнения могут быть использованы механизмы обобщения: для выбранного слова могут быть получены 10 наиболее используемых синонимов и для каждого из этих синонимов(при подставлении в фразу) также вычисляется подобие со второй фразой. В качестве значения функции подобия выбирается максимальное значение среди всех подобий.

$$Score = Max\{ConceptNetSim\{phrasePart\ term^i,\ phrase\}\}$$

3.3.1.3. Правила сравнения грамматических структур.

1. **Сравнение форм вида глагол-существительное(VerbNounClause - VNC).**

   Для начала выбираются все части предложения, которые содержат связи глаголов к существительным. Для сравнения двух VNC используется следующая процедура:

   - Сначала определяем являются ли глаголы синонимами или принадлежат схожим концептам(используя ConceptNet).
   - Если глаголы схожи в значении, то мы сравниваем существительные, с которыми эти глаголы связаны по аналогичному принципу.
   - Если какие-то из существительных схожи по значению, мы

сравниваем их прилагательные и наречия по тому же принципу.

2. **Сравнение форм, состоящих из существительных(NounClauses - NC).**

   Мы считаем ERR формой, состоящей из существительных, если она содержит только существительные. В связи с этим рассмотрим 2 случая:

   - Если оба ERR состоят из существительных, мы сравниваем все существительные между собой. Подобие глаголов увеличивает значение функции подобия для этих ERR.
   - Если только входное ERR состоит из существительных, тогда все существительные входного ERR сравниваются с существительными из ERR из данных для обучения.

3. **Сравнение прилагательных и наречий.**

   Мы сравниваем свободные прилагательные и наречия из ERR, т.к. к этому шагу только несвязанные прилагательные и наречия остались неучтенными в функции подобия. После того, как была вычислена функция подобия между двумя ERR, необходимо решить является это подобие семантическим или подобием по ключевым словам. Подобие считается семантическим, если выполнено одно из следующих утверждений:

1. Эмоциональные VNC(глагол имеет эмоциональную окраску) оказались полностью подобны - т.е. и глаголы, и его существительные, и их прилагательные и наречия оказались подобны.
2. Все существительные оказались подобны и оба ERR могут рассматриваться как формы состоящие только из существительных.
3. Прилагательные и наречия подобны и оба ERR не имеют других компонентов(нет VNC & NN).

4. Все слова в обоих ERR подобны(то же самое предложение) или входное ERR является частью ERR из набора для обучения.

Если одно из вышеуказанных условий было выполнено, то подобие считается семантическим, иначе - подобием по ключевым словам.

3.3.2. Альтернативная вероятностная классификация.

В качестве запасного метода классификации мы используем вероятностные алгоритмы - APA(Alternative probabilistic algorithm). Если предложение было отвергнуто kNN-классификатором, то предлагается следующий набор действий. Для каждого слова входного предложения, которое требуется классифицировать вычисляется следующая величина - эмоциональная вероятность среди данных для обучения:

$P(E,\ w)\ =\ log($*#Sentences with emotion E that contain the word w / # total sentences that contain w*$)$, где E принимает значения каждой из эмоций по Экману, а w - каждое из слов входного предложения. Данная величина вычисляется для предложений из данных для обучения. После вычисления всех значений мы получаем 6 * (количество слов во входном предложении(ERR)) эмоциональных вероятностей, т.е. фактически для каждого слова входного предложения мы имеем распределение значений его эмоционального контекста. Теперь необходимо причислить такое предложение к одному из тэгов. Для этого просуммируем вероятностные вклады для каждой эмоции для всех слов предложений и возьмем максимум по этим вкладам. Строго данное рассуждение может быть описано так:

$E_{result}(ERR)\ = max_E(\ \sum_{n=1}^{|ERR|} P(E,\ w_n) / |ERR|)$, где E - множество эмоций по Экману, |ERR| - количество слов во входном ERR.

Такой алгоритм также может отвергнуть ERR, если ни одно из слов не находится в датасете для обучения. Такую проблему можно решить, обращаясь к большему скоплению данных, нежели данные для обучения, например, к хранилищу Интернета.

Для начала в любом поисковом движке(Google, например) вводятся слова, составляющие ERR предложения без указания каких-либо эмоций. Записывается число вхождений, возвращаемое движком. Далее аналогичная процедура проводится для всех слов из ERR и каждой из эмоций по Экману и также записывается число вхождений. Очевидно, что для комбинации ERR + эмоция число вхождений будет всегда меньше, чем просто для ERR. Результирующий тэг вычисляется по следующим формулам:

$$P'(E,\ ERR) = log(\#Query\ hits\ for\ ERR + E / \#Query\ hits\ for\ ERR)$$

$$E_{result}(ERR) = max_E(P'(E,\ ERR))$$

3.4. Классификация ERR.

Для обучения классификатора использовались данные Aman 2007 [12, 13], в которых каждое предложение размечено одной из 6 эмоций классификации Экмана. Датасет состоит из эмоционально богатых предложений, собранных из различных блогов и размеченных эмоциями по Экману. Этот набор данных было решено выбрать, поскольку блоги предоставляют большое различие стилей написания и тем. Таким образом, предложения хорошо дифференцированы по тематикам и используют множество разноплановых фраз.

Необходимо отметить, что для данного датасета потребовалась предварительная обработка с целью исправления опечаток и прочих недочетов данных.

Для данного эксперимента метод kNN-классификации использовался с параметром k = 1.

Данные из датасета делятся на две равные части: первая часть - данные для непосредственно обучения и вторая часть - данные для оценки качества алгоритма.

После обучения классификатора(т.е. агрегации данных ERR для данных для обучения) проводится проверка алгоритма. Т.е. для тех предложений, для которых известна их эмоциональная окраска, проводится классификация алгоритмом и сравнивается реальная окраска предложения с результатом классификации. Для этого была использована вторая часть данных Aman-датасета.

Для большей точности для обучения использовались предложения до 10 слов длиной, поскольку более длинные предложения могут содержать несколько эмоциональных частей, которые более сложны для анализа и могут быть противоречивыми. Оценка качества алгоритма проводилась на предложениях от 10 до 15 слов длиной. С учетом этих соображений данные для обучения составили 500 предложений длиной до 10 слов. Данные для оценк качества составили также 500 предложений длиной от 10 до 15 слов.

Эмоция	Precision *true positive / (true positive + false positive)*	Recall *true positive / (true positive + false negative)*	F-score
Радость	0.89	0.92	0.91
Грусть	0.84	0.81	0.83
Отвращение	0.79	0.78	0.79
Злость	0.82	0.76	0.79

Удивление	0.80	0.75	0.77
Страх	0.91	0.84	0.88

Рис 9. Таблица результатов классфикации.

По таблице видно, что результаты классификации получаются достаточно хорошими и данный классификатор может быть использован для реальных данных.

3.5. Построение образа автора.

После оценки качества алгоритма можно приступать к анализу реальных данных. Для анализа личности необходим текст или совокупность текстов объемом не менее 1400 слов, написанный человеком, личность которого хочется проанализировать. Важным условием хорошего результата являются эмоционально насыщенные тексты, в которых человек активно высказывает свое мнение по разным темам.

Личностный портрет человека в системе “Большой пятерки” формируется путем отображения множества размеченных предложений из входных текстов во множество черт “Большой пятерки”.

Необходимо также обратить внимание, что не каждое предложение среди входных данных обязательно будет размечено одной из эмоций Экмана. Часть предложений могут быть отвергнуты классификаторов вообще или оказаться безэмоциональными. Такие предложения составляют отдельную группу и просто удаляются перед началом отображения. Большое количество удаленных предложений может быть как следствием недостаточно высокого качества входных данных, так и следствием безэмоциональности индивида. Для

уточнения по какой из причин было удалено много предложений необходимо дополнительное исследование, поэтому в рамках данной работы система просто их отбрасывает без выяснения корня проблемы. Если более половины предложений оказались удалены, то система показывает предупреждение о том, что вероятность неточности результата высока ввиду низкого качества входных данных.

Далее будет описан алгоритм отображения.

Входными данными для алгоритма отображения являются размеченные предложения из текстов, выбранных для анализа. Таким образом, для каждой эмоции по Экману можно подсчитать количество предложений, соответствующих ей. Вспомним какие эмоции включает данная классификация: *радость, грусть, удивление, отвращение, злость* и *страх*. То есть эти эмоции(для каждой из которых может быть подсчитано процентное содержание среди всех) необходимо отобразить на другое множество - множество характеристик “Большой пятерки”. Напомним также какие это характеристики: *экстраверсия, нейротизм, доброжелательность(способность прийти к согласию)*, *добросовестность(сознательность)* и *открытость опыту(интеллект)*.

В качестве результата отображения для каждой из характеристик “Большой пятерки” будет указано процентное содержание данной характеристики в личностном портрете исследуемого человека - т.е. число от 0 до 100.

Для получения корректного отображения из одного пространства признаков в другое, необходимо использовать таблицу эмоций, определяющих каждую из черт “Большой пятерки”. Для нахождения синонимичных эмоций используется дерево эмоций WordNet-Affect. Также для уточнения отображения будет

использован график корреляции между чертами “Большой пятерки” и эмоциями, полученный в рамках работы [18]

Big Five Dimensions	Facet (and correlated trait adjective)
Extraversion vs. introversion	Gregariousness (sociable) Assertiveness (forceful) Activity (energetic) Excitement-seeking (adventurous) Positive emotions (enthusiastic) Warmth (outgoing)
Agreeableness vs. antagonism	Trust (forgiving) Straightforwardness (not demanding) Altruism (warm) Compliance (not stubborn) Modesty (not show-off) Tender-mindedness (sympathetic)
Conscientiousness vs. lack of direction	Competence (efficient) Order (organized) Dutifulness (not careless) Achievement striving (thorough) Self-discipline (not lazy) Deliberation (not impulsive)
Neuroticism vs. emotional stability	Anxiety (tense) Angry hostility (irritable) Depression (not contented) Self-consciousness (shy) Impulsiveness (moody) Vulnerability (not self-confident)
Openness vs. closedness to experience	Ideas (curious) Fantasy (imaginative) Aesthetics (artistic) Actions (wide interests) Feelings (excitable) Values (unconventional)

Рис 10. Таблица связей “Большой пятерки”

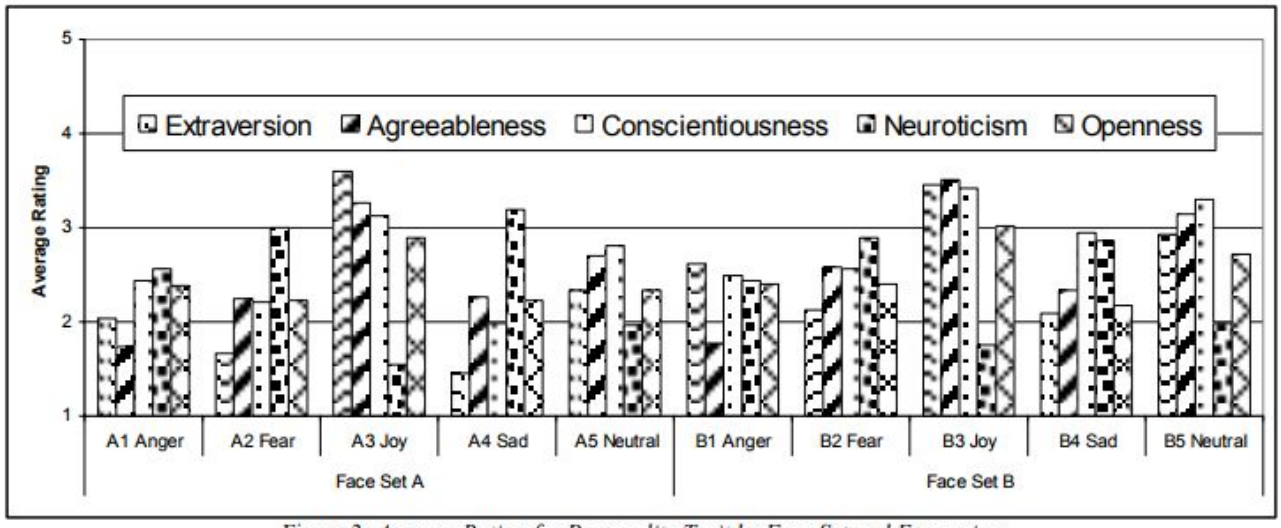

Figure 3: *Average Rating for Personality Trait by Face Set and Expression*

Рис 11. Корреляция черт “Большой пятерки” и эмоций.

Первой характеристикой является *экстраверсия*. *Интроверсия - экстраверсия* - распространенный в психологии критерий категоризации или показатель измерения черт личности. Термины *интроверсия* и *экстраверсия* впервые были введены Юнгом, однако их понимание и употребление в психологии отличаются от первоначального значения. Скорее фокусируясь на межличностном поведении, Юнг, однако, определял интроверсию как «поведенческий тип, характеризуемый направленностью жизни на субъективное психическое содержание» (фокус на внутреннюю психическую активность); и экстраверсию как «поведенческий тип, характеризуемый концентрацией интересов на внешних объектах», (внешний мир). Экстраверсия проявляется в дружелюбном, разговорчивом, энергичном поведении, в то время как интроверсия проявляется в более замкнутом и уединенном поведении. Экстраверсия и интроверсия обычно рассматриваются как единое пространство измерений. Поэтому высокие показатели одной характеристики подразумевают низкие показатели другой.

Ганс Айзенк заимствует у Юнга термин «экстраверсия» при создании своей

диспозициональной модели. Айзенк обнаружил, что в разных исследованиях, проведенных разными исследовательскими группами, параметры личности согласовано варьируются по степени своей ориентации на социальные отношения в противовес ориентации на рефлексию, переживания, чувства. Эти понятия являются полюсами супер-фактора — комплекса коррелирующих между собой черт личности, который детерминирован генетически. Типичный экстраверт по Айзенку общителен, оптимистичен, импульсивен, имеет широкий круг знакомств и слабый контроль над эмоциями и чувствами. Типичный интроверт спокоен, застенчив, отдалён от всех, кроме близких людей, планирует свои действия заблаговременно, любит порядок во всём и держит свои чувства под строгим контролем [14].

Как видно из таблицы свойств “Большой пятерки”, радость является эмоцией определяющей экстраверсию. Оставшиеся эмоции классификации Экмана будут считаться нейтральными для формирования уровня экстраверсии личности.

$$Extraversion = Norm[Count(Hp) / Count(All)]$$

$Norm$ − функция нормировки, будет определена ниже.

Следующей характеристикой является *нейротизм(противоположный полюс - эмоциональная стабильность)*. Невротизм(нейротизм) - черта личности, характеризующаяся эмоциональной неустойчивостью, тревогой, низким самоуважением. В широком смысле нейротизм можно определять как неспособность эффективно регулировать негативные эмоции (van Egeren, 2009). Негативные эмоции склонны возникать, когда люди полагают, что они плохо справляются с достижением своих целей (Carver & Scheier, 1990).

Сильная корреляция между невротизмом и негативным аффектом получена во множестве работ (Clark & Watson, 1999; Costa & McCare, 1980; Tellegen, 1985;

Warr et al., 1983; Watson & Clark, 1984, 1992). Elliot and Thrash (2002) обнаружили, что невротизм, негативная эмоциональность и поведенческое торможение высоко нагружают один и тот же фактор, названный авторами «избегающим темпераментом» (avoidance temperament). Watson and Clark (1984), проанализировав данные 20 исследований, продемонстрировали сильную положительную связь между нейротизмом и реакцией стресса на экзамены, интервью и другие ситуации, в которых может случиться неудача и возникнуть разочарование.

Тремя основными эмпирически получаемыми компонентами нейротизма являются возбудимость (irritability), незащищенность (insecurity) и эмоциональность [17].

Нейротизм характеризуется явным, плохо контролируемым проявлением негативных эмоций. По таблице это эмоции *злость*, *отвращение* и *страх*.

$$Neuroticism = Norm[(Count(Ag) + Count(Dg) + Count(Fr)) / Count(All)]$$

Доброжелательность(способность прийти к согласию) - это следующая характеристика “Большой пятерки”. Данная черта отвечает за формирование социальных альянсов и связей. Как видно из графика корреляции, на данный параметр влияют такие эмоции, как *радость* и *страх*.

$$Agreeableness = Norm[(Count(Fr) + Count(Hp)) / Count(All)]$$

Добросовестность(сознательность) рассматривается как черта, характеризующая ответственность индивида и может отвечать за эффективное управление реакциями приближения и избегания. Добросовестность может быть достаточно хорошим предиктором академической успеваемости и производственной эффективности, в том числе, при контроле показателя интеллекта. Как видно из графика корреляции, на данную черту влияют такие

эмоции как *грусть* и *страх*.

$$Conscientiousness = Norm[(Count(Fr) + Count(Sd)) / Count(All)]$$

Последней характеристикой требующей уточнения является *открытость опыту(интеллект)*. Эту категорию черт отличают такие характеристики, как воображение и проницательность — люди с сильно развитыми качествами этой группы имеют достаточно широкий круг интересов. Данная черта будет характеризоваться эмоцией *удивления* в нашей модели.

$$Openness = Norm[Count(Sp) / Count(All)]$$

Определим теперь функцию нормировки $F = Norm[arg]$, аргументом этой функции всегда является доля нужных для вычисления характеристики эмоций в датасете. Если эта доля оказалась меньше нуля, то ее необходимо приравнять к нулю, т.к. это говорит о наличии признаков противоположных искомым. После этого можно утверждать, что аргументом функции нормировки будет всегда являться положительное число от 0 до 1. Это число показывает процентное содержание искомых эмоций для данной характеристики в датасете. Т.е. можно утверждать, что если это число оказалось больше 0.5, то более половины предложений указывают на наличие искомой характеристики. Будем считать число 0.5 - порогом, после которого искомая характеристика считается ярко выраженной и имеет 100% содержание в портрете личности. Строго функцию нормировки можно определить так:

$$Norm[arg] = \{arg \geq 0.5 \; then \; 100\%, \; else \; arg * 200\%\}$$

Таким образом, каждая характеристика “Большой пятерки” будет иметь значение от 0% до 100%, что отражает содержание этой характеристики в личности человека.

4 Анализ результатов

Для тестирования написанной системы были выбраны несколько популярных личностей, о которых известно достаточно много, чтобы можно было оценить правильность работы системы. Для каждой личности и соответствующего ей набора текстов будет проведено исследование как написанной системой, так и сервисом от IBM “Personality Insights”, а также будет произведено сравнение результатов.

В качестве первого исследуемого возьмем Илона Маска и его речь об освоении Марса “Making humans a multiplanetary species.” Проанализируем данный текст на нашей системе и на системе IBM “Personality Insights”.

“Personality Insights”:

Экстраверсия ---------- 9%

Открытость опыту(интеллект) ---------- 99%

Доброжелательность(способность приходить к согласию) -------- 0%

Добросовестность(сознательность) ---------- 31%

Невротизм ---------- 71%

СФОА(система формирования образа автора):

Экстраверсия ---------- 15%

Открытость опыту(интеллект) ---------- 58%

Доброжелательность(способность приходить к согласию) -------- 20%

Добросовестность(сознательность) ---------- 87%

Невротизм ---------- 52%

Следующим текстом для анализа выберем знаменитую “Фулотонскую речь” Черчилля, в дальнейшем считавшуюся сигналом для начала Холодной войны.

“Personality Insights”:

Экстраверсия ---------- 16%

Открытость опыту(интеллект) ---------- 100%

Доброжелательность(способность приходить к согласию) -------- 7%

Добросовестность(сознательность) ---------- 88%

Невротизм ---------- 78%

СФОА(система формирования образа автора):

Экстраверсия ---------- 9%

Открытость опыту(интеллект) ---------- 53%

Доброжелательность(способность приходить к согласию) -------- 11%

Добросовестность(сознательность) ---------- 47%

Невротизм ---------- 54%

Следующим текстом для анализа была выбрана речь “Quit India” Махатмы Ганди.

“Personality Insights”:

Экстраверсия ---------- 3%

Открытость опыту(интеллект) ---------- 99%

Доброжелательность(способность приходить к согласию) -------- 5%

Добросовестность(сознательность) ---------- 77%

Невротизм ---------- 69%

СФОА(система формирования образа автора):

Экстраверсия ---------- 17%

Открытость опыту(интеллект) ---------- 49%

Доброжелательность(способность приходить к согласию) -------- 40%

Добросовестность(сознательность) ---------- 52%

Невротизм ---------- 24%

В целом, СФОА дает адекватные результаты, хотя, и порой сильно расходящиеся с результатами “Personality Insights”. Полученные данные при правильной обработке и, возможно, некотором их расширении(добавление объективных данных о пользователях) могут принести практическую пользу во многих приложениях.

Ограничения модели.

Безусловно, данный сервис не лишен недостатков. К ограничениям данного подхода к анализу текстов можно отнести невозможность отличить сарказмы в предложениях, также при наличии противоположных эмоций в предложении, невозможно отследить какая из эмоций является основной и правильно классифицировать предложение.

Еще одним ограничением на данном этапе разработки является отсутствие возможности оценить результат данного сервиса иначе, как сравнить со схожим сервисом. Для решения этой проблемы необходимо получить статистику следующего характера: для набора испытуемых, чьи тексты были исследованы на данном сервисе необходимо получить результаты сертифицированного психологического тестирования для характеристик “Большой пятерки”. И,

таким образом, выявить подобие данных результатов.

5 Заключение

Представленная работа проводит свое исследование на стыке математики, лингвистического анализа и психологии.

В рамках данной работы была построена система формирования образа автора(СФОА), который принимает в качестве входных данных набор текстов, написанных человеком, которого требуется проанализировать и на выходе выдает модель “Большой пятерки” для этого человека. Проведенные эксперименты показали, что сервис выдает адекватные результаты, но их точная оценка требует дополнительных усилий по причине необходимости проведения психологического тестирования для испытуемых.

Также необходимо отметить, что в рамках данной работы был построен классификатор, распознающий эмоции в каждом отдельном предложении, который может сам по себе быть использован в практических целях.

Литература

1. *Рыков В.В*. Обработка нечисловой информации. Управление знаниями. – М.: МФТИ, 2008.
2. *Рыков. В.В*. Формирование образа автора как когнитивный процесс. - М.: МФТИ, 2016.
3. Personality Insights, IBM, 2017. https://www.ibm.com/watson/developercloud/personality-insights.html
4. Cambridge Analytica, Wikipedia, 2017. https://ru.wikipedia.org/wiki/Cambridge_Analytica
5. *Goldberg, L.R*. The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 1992.
6. P. Ekman, Basic emotions. In T. Dalgleish and T. Power (Eds.) The handbook of cognition and emotion. Pp. 45-60. New York.: John Wiley & Sons, 1999.
7. K. Toutanova, Klein D., Manning C., Singer Y., StanfordPOStagger, [Online].Available:http://nlp.stanford.edu/software/tagger.shtml,Stanford, 2003.
8. Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. "Generating typed dependency parses from phrase structure parses." Proceedings of LREC. Vol. 6. 2006.
9. Main article.
10. George A. Miller, "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.
11. Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." Proceedings of LREC. Vol. 6. 2006.
12. Saima Aman, and Stan Szpakowicz. "Identifying expressions of emotion in text." Text, Speech and Dialogue. Springer Berlin Heidelberg, 2007.

13. Saima Aman, and Stan Szpakowicz. "Using Roget's Thesaurus for Fine-grained Emotion Recognition." IJCNLP. 2008.
14. Wikipedia, 2017. https://en.wikipedia.org/wiki/Big_Five_personality_traits
15. Bernardo Magnini and Gabriela Cavaglia'. *Integrating Subject Field Codes into WordNet*. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pp. 1413-1418 Athens, Greece, 31 May - 2 June, 2000.
16. Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*. In Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, 2004, pp. 101-108.
17. Айзенк Г. Ю. Структура личности. — СПб.: Ювента; М.: КСП+, 1999. — 464 с.
18. Brian E. Tidball, Sasanka Prabhala, and Jennie J. *Making faces: exploring perceptions of personality based on emotional expressions.* Gallimore Department of Biomedical, Industrial and Human Factors Engineering Wright State University, Dayton OH 45435, 2006.