УДК: 311.2+330.88

1.10. Обзор развития методов анализа текста: от ручной обработки в психолингвистике к современным автоматизированным программам в маркетинге

Кашкин В.В.[1], Андросик Юрий,[1]
[1] Агентство международного маркетинга и исследований Kashkin.com.cn

В статье сделан краткий исторический обзор развития методов анализа текста. Значительную роль в этом сыграли психолингвистические исследования и информационные технологии. Они явились решающими факторами, которые и определили современное состояние этой области научных изысканий. Психолингвистические исследования позволили связать сначала письменный текст, а позднее и устную речь, с психологическими особенностями индивидуума, его чертами личности и мышлением, что в последующем в сочетании с возможностями современных вычислительных машин дало новый импульс в изучении поведения человека, расширив границы научных исследований в экономике и маркетинге. Авторы протестировали и проанализировали свыше 50 программных продуктов по анализу текста, выделили основные функции ПО и предложили основные направления их использования в маркетинговых исследованиях.

Введение

Каждую минуту в мире генерируется огромное количество информации, а совокупный объем данных, накопленных человеком, превышает несколько десятков миллиардов терабайт. При этом значительная часть информации генерируется и хранится в виде текстовых данных. Эта информация является неструктурированной и, на первый взгляд, разрозненной и бесполезной. Например, бесчисленное множество комментариев, мнений, высказываний и отзывов в социальных сетях кажется малопригодным с точки зрения рядового человека и излишним. Тем не менее, это не так. Подобная информация отражает внутренний мир человека и является цифровым отпечатком личности. Анализ такой информации открывает широкие возможности как для бизнеса, так и для науки. Особенный интерес вызывает перспектива использования анализа текста в маркетинге, так как каждый человек представляет собой покупателя, потребителя, пользователя или клиента. Знание особенностей его поведения, того, как он мыслит и как реагирует на стимулы, позволяет создавать поведенческие модели и предсказывать действия человека. Эта информация ценна как в научном, так и в коммерческом отношении. Фундаментом этого знания является связь текста и психологии личности, так как текст является продолжением мысли и речи человека.

Обзор развития методов анализа текста

Текстовую аналитику (или анализ текста) обобщенно можно рассматривать как подход к получению информации из текстовых материалов. Сейчас такой анализ называют интеллектуальным анализом текста либо майнингом текста.

Текстовая аналитика является частью интеллектуального анализа данных и предполагает применение алгоритмов искусственного интеллекта и машинного обучения к текстовым данным. Разница между интеллектуальным анализом данных и интеллектуальным анализом текста заключается в том, что в первом случае шаблоны извлекаются из текста на естественном языке, а во втором – из структурированных баз данных [Hearst, 2005]. Поэтому чаще текстовую аналитику определяют как область, которая занимается открытием новой, ранее неизвестной информации путем автоматического извлечения информации из различных письменных источников с помощью компьютера [Tan, 1999]. Основная цель, по сути, состоит в том, чтобы превратить текст в данные для анализа с помощью применения технологий обработки естественного языка, различных типов алгоритмов и аналитических методов. Это повышает полезную отдачу текстовой аналитики, но и при этом повышает ее сложность.

Анализ текста является междисциплинарным направлением. Его ядро основано на таких областях научного знания, как лингвистика, статистика, математика, информатика, наука о данных, а области применения повсеместны: экономика, социология, психология, медицина, история, антропология и другие [McLaughlin, 2022].

В то же время аналитику текста можно представить как совокупность нескольких прикладных направлений [Miner, 2012] – поиска информации, кластеризации документов, веб-майнинга, классификации документов, извлечения идей, извлечения информации, обработки естественного языка, – в рамках которых решаются определенные задачи обработки текста (см. рисунок 1). Однако большую часть задач стало возможным решить только в 21-м веке с применением производительных компьютеров.

Историю текстовой аналитики начинают отсчитывать с момента появления вычислительных машин. Однако работы, в которых осуществлялся анализ текста, появляются уже в середине 19-го века. В 1851 году английский логик Огастес де Морган предположил, что вопросы авторства могут быть решены путем определения того, «*не имеет ли один текст более длинных слов, чем другой*» [Holmes, 1998]. В 1887 году Т.С. Менденхолл начал изучать длину слов у Шекспира, Марло и Бэкона и показал, что длина слова не является эффективным средством установления авторства, хотя Менденхолл

нашел сходство между Шекспиром и Марло, которое по сей день все еще исследуется, но с помощью более продвинутых методов (Holmes, D. I. (1998)). В 1888 году в исследовании слов Б. Бурдон проанализировал Исход Библии и рассчитал частоты, переставив и классифицировав их, исключив стоп-слова [Bourdon, 1892].

В начале 20-го века Ж.Б. Эсту (Jean-Baptiste Estoup), один из основных основоположников статистического анализа текста (представитель французской стенографии), определил понятие ранга как позицию, занимаемую словом в списке слов, отсортированных по убыванию частоты [Estoup, 1916], и сформулировал закон, позже получивший название Закона Ципфа (который устанавливает связь между рангом слов в тексте, упорядоченном в порядке убывания частоты их встречаемости, и этой частотой [Petruszewycz, 1973, Zipf, 1945]). Он сделал это до Ципфа, однако последний стал популяризатором закона и распространил его на другие сферы жизни.

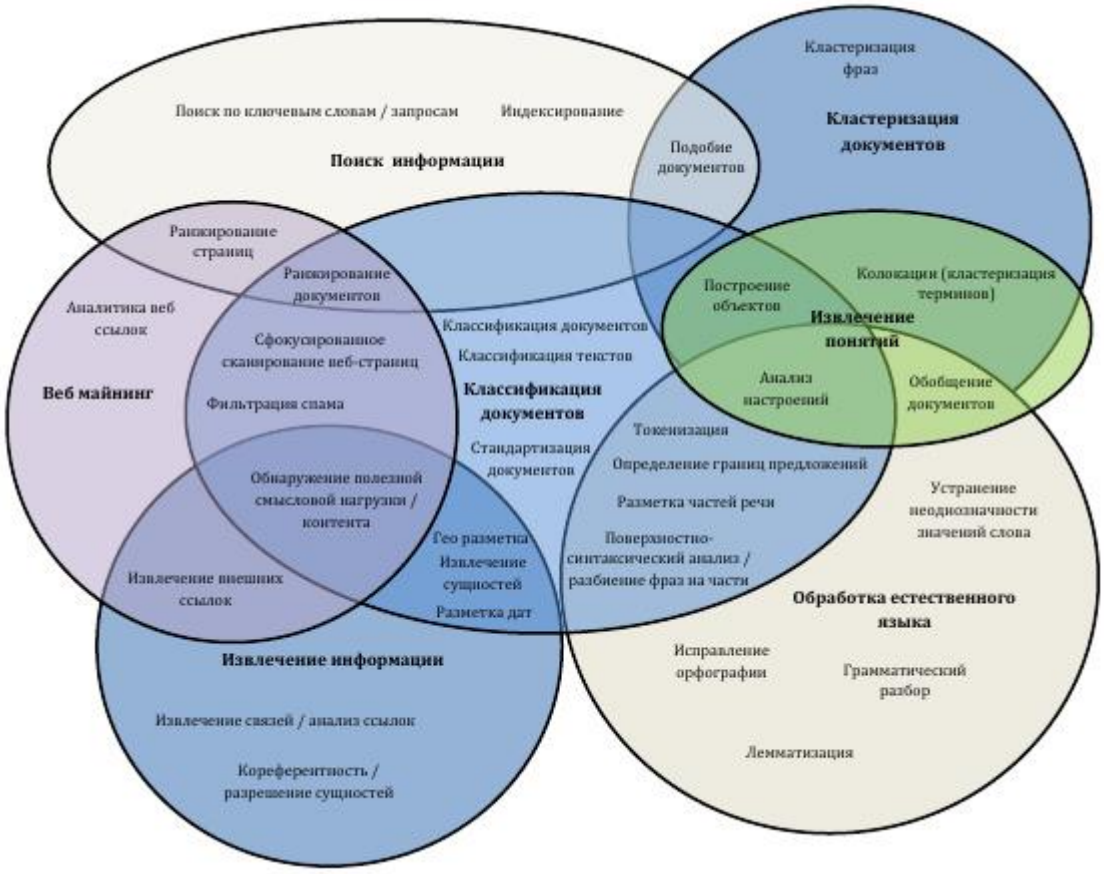

Рисунок 1 – Задачи обработки текстов
Источник: доработано авторами на основе [Miner, 2012]

В 20-м веке работу Менденхолла по определению авторских стилей и их черт продолжили такие лингвисты, как Дж. К. Ципф и Г.У. Юл. В 1932 году Ципф выявил связь между количеством слов, встречающихся в тексте ровно n раз, и самим n [Zipf, 1932]. Связь между логарифмами этих величин была линейной. Он обнаружил, что небольшое количество слов используется постоянно, а подавляющее большинство – очень редко. Слово первого ранга всегда используется вдвое чаще, чем слово второго ранга и втрое чаще, чем слово третьего ранга.

В 1938 году Юл предложил использовать длину предложений как идентификатор (дискриминатор) авторского стиля [Yule, 1938], а уже в 1944 году Юл разработал вычислительную меру постоянства текста, названную «характеристикой Юла K», которая предназначалась для идентификации автора, с учётом того, что она будет различаться для текстов, написанных разными авторами. Более поздние исследования показали, что мера K Юла недостаточно надежна, чтобы различать авторов [Kumiko, 2015].

В целом исследования велись в области стилометрии и лексической статистики – в том числе для создания систем стенографии, а также дипломатических и военных шифров, основу которых составляют стенография и криптография. В свою очередь, в основе последних лежат понятия частот и таблицы частот [Mandelbrot, 1966].

В 1946 году итальянский священник-иезуит Роберто Буса предложил IBM идею использовать компьютеры для изучения текстов и стал автором Index Thomisticus – полной лемматизации работ святого

Фомы Аквинского. В 1980 году, после 30-ти лет работы, печатное издание из 56 энциклопедических томов «Index Thomisticus» увидело свет. Этот труд собрал все произведения св. Фомы Аквинского в формате, читаемом и управляемом компьютером с использованием методики, разработанной отцом Бусой. Роберто Буса по праву считается отцом компьютерной лингвистики. На сегодняшний день компьютерная лингвистика занимается разработкой алгоритмов и прикладных программ для обработки языковой информации, а также разработкой математических моделей для описания естественного языка.

В середине 20-го века благодаря популярности психолингвистических исследований [Iezzi, 2020] анализ текста получил новое развитие.

Считается, что термин “психолингвистика” был введен в научный оборот в 1954 году в книге “Психолингвистика: обзор теории и проблем исследования” [Osgood, 1965], где ее определили как занимающуюся *«в самом широком смысле отношениями между сообщениями и характеристиками человеческих личностей, которые их отбирают и интерпретируют*». По сути, она изучает взаимоотношения языка, мышления и сознания, включая связи между сообщениями (которые выступают в том числе как текст) и личностью (чертами личности). Сам термин “психолингвистика” был введен в книге американского психолога Дж. Кантора “Объективная психология грамматики” в 1936 году, но до 1946 года не использовался. Развитие получил после публикации Н. Г. Пронко в статье «Язык и психолингвистика» [Pronko, 1946]. В этой работе Пронко попытался систематизировать исследования о языке, которые были проведены в различных смежных с лингвистикой дисциплинах и были похожи тем, что основаны на “*существенных психологических особенностях языковых явлений*” [Willem, 2013]. В частности, он указал на наличие пробела в знаниях на стыке психологии и лингвистики, показав, что нужна единая теоретическая основа для подобных междисциплинарных исследований.

Однако психолингвистика рассматривает связь языка и мышления достаточно широко – от изучения проблем психического здоровья и оценки психического состояния человека до отдельных аспектов поведения человека, его мыслей и эмоций.

Уже в работах середины 20-го века предполагалось, что слова, которые используют люди, могут отражать психологические процессы, устойчивые и значимые паттерны поведения [Boyd, 2017]. То, какие слова использует человек, многое может сказать об его социальных и психических качествах, эмоциях и скрытых мотивах [Pennebaker, 2003]. Устойчивое словоупотребление может свидетельствовать как об индивидуальных различиях, так и об общих закономерностях. Однако подобные закономерности и различия варьируют в зависимости от региона / территории проживания человека [Giorgi, 2022]. При этом язык человека, также как и черты личности, являются почти константами в различные промежутки времени [Pennebaker, 1999b]. Это служит исходной предпосылкой того, что базовые черты личности, отраженные в способах мышления, ощущениях, эмоциях и поведении в целом, практически неизменны. Они кодируются языком человека неосознанно и позволяют выявлять личностно-языковые связи для психологический оценки личности и составления ее профиля. Например, чем выше частота слов, которые использует человек и которые могут быть отнесены к определенной категории (например, такой категории, как власть), тем выше вероятность того, что эти слова занимают центральное место в его психике. По этой причине анализ текста имеет потенциально высокую коммерческую ценность.

Во второй половине 20-го века в силу отсутствия широкого распространения компьютеров и их слабой вычислительной мощности основные расчеты при психологическом анализе текста выполнялись вручную. Поэтому исследования были сосредоточены на изучении языка в его связи с формами получения данных, например, анкетами или опросами пациентов. Высокая трудоемкость такой работы корректировала направления анализа в сторону решения первоочередных задач – изучения и лечения психических расстройств и проблем психического здоровья. Это обусловило и соответствующие методы исследований. Однако в современных условиях используются данные в том числе и из социальных сетей. Они позволяют прогнозировать развитие личности и диагностировать психологические проблемы, психические расстройства [Eichstaedt, 2018], проблемы умственного здоровья [Reece, 2017], позволяют дать оценку психическому здоровью [Saha, 2022]. Раньше часто использовались словари, которые содержали группы слов, объединенных по определенному признаку. На практике сложно было составить словарь, который бы однозначно определял слова или словосочетания, значимые для конкретного психического состояния. Поэтому данные методы были узкоспециализированы и зачастую неприменимы для более широкого круга исследований. С другой стороны, словари используются до сих пор, так как исследователи могут создавать свободно наборы слов под интересующую их тему или проблему.

Для устранения указанного выше недостатка подход дополнялся батареями тестов, опросниками, самоотчетами и любыми другими способами получения текстовых данных, но, как правило, в форме обратной связи. Обратная связь должна была быть строго регламентирована, так как ответы респондентов должны были согласовываться с неким эталонным теоретическим конструктом, который бы характеризовал человека как личность и идентифицировал его черты. Нужна была база для сравнения, чтобы выделять индивидуальные различия и отклонения от нормы. Наиболее популярным конструктом стала модель Большой пятерки [Soto, 2020]. Именно она является эталонной моделью оценки личности.

Альтернативой использования словарей были кодировщики [Iliev, 2014]. Это люди, которые хорошо интерпретировали письменный и устный текст. Такие люди, читая текст или слушая речь, могли определить эмоциональное состояние автора. Однако с ростом объемов текстовой информации подобный подход стал непрактичным. Существование подобных проблем привело к появлению идеи автоматизированного анализа текста. Тем не менее, обучить компьютер либо программу интерпретировать текст и находить в нем смысл является нетривиальной задачей, с которой человек справляется лучше. Одна из идей подобного обучения состоит в следующем. Многие проблемы человека можно заранее определить по набору слов, которые человек использует в своей речи. Эти слова обычно связаны с конкретными проблемами или темами. В анализируемом тексте программа ищет слова из словаря, считает их количество и определяет, в какой степени текст соотносится с темой. Такая идея изначально была заложена в автоматический анализ текста. Но и здесь появились внутренние противоречия, так как большинство компьютерных программ плохо оценивают контекст, иронию или сарказм, метафоры или многозначные слова [Chung, 2007]. Более того, чем чаще используется слово, тем более многозначным оно становится [Miller, 1999, Zipf, 1945]. Какое-то время существовало мнение, что автоматизированный анализ текста невозможен в силу того, что невозможно автоматически идентифицировать смысл слова [Bar-Hillel, 1960].

Вплоть до конца 20-го века в науке превалировал подход, согласно которому центральным измерением в психологии являлись слова, которые несут смысл, то есть имеют содержательный аспект (“open class” words – существительные, глаголы, прилагательные и наречия – играют лексическую роль). Однако благодаря работам таких исследователей, как George Miller, было установлено, что функциональные слова (“closed class” words – местоимения, предлоги, союзы, частицы – играют грамматическую роль) зачастую даже лучше либо на уровне “open class” words позволяют объяснять психологическое состояние

человека [Pennebaker, 2003]. Более того, выбор функциональных слов, например, таких как союзы, предлоги, частицы или междометия, происходит неосознанно, так как обычно человек думает о содержательной стороне своей речи. Поэтому исследователи начали изучать не только “что говорится”, но и “как говорится” – другими словами, кроме содержательных слов использовать еще и функциональные – местоимения, артикли, предлоги, союзы и вспомогательные глаголы.

Позднее анализ текста дополняется различными метриками, такими как длина слов и предложений, сложность слов, использование знаков препинания, тестовых мемов. Здесь на первый план выходят методы статистической обработки текста. Ключевым показателем становится частота употребления, или встречаемость слова. Однако в силу того, что некоторые слова чаще встречаются, чем остальные, их значимость для анализа текста снижается. Они не позволяют выделить отличия между текстами, найти в них существенные признаки и несколько искажают результаты анализа, если их не учитывать. По этой причине исследователи стараются вычленять значимые языковые признаки. Данные признаки должны позволять классифицировать тексты и находить максимальные различия между ними. Это достигается за счёт комбинирования методов обработки естественного языка.

Возможности использования автоматизированного анализа текста в маркетинге

Облегчило задачу анализа текста массовое внедрение информационных технологий в исследовательскую практику. Это позволило создать приложения для автоматического анализа текста. Одним из первых приложений стало Linguistic Inquiry and Word Count (LIWC, произносится как “Luke” [Pennebaker, 1999a]). Основная идея приложения заключается в том, что, если человек использует часто определенные слова и говорит на конкретную тему, то это отражает его психологические особенности и характеризует его как личность. Например, если человек недоволен, то большинство употребляемых им слов будет относится к теме недовольства. Программа подсчитывает частоту слов, которые относятся к теме или категории. Но она не учитывает контекст и требует для построения выводов наличия определенных наборов слов. Отчасти это было решено путем использования метода извлечения смысла (Meaning Extraction Method (MEM [Chung, 2008]). Суть метода в том, что он автоматически определяет слова, которые используются вместе и естественным образом составляют определенную тему. Алгоритм сопоставляет их с категориями и словарями программы либо формирует новую категорию. Реализация метода осуществлена в программе Meaning Extraction Helper [Boyd, 2016].

На данный момент времени существует множество программ для автоматического анализа текста (мы не будет рассматривать библиотеки и модули Python, R, Java, так как они требуют навыков программирования). Реализация данных программ включает множество алгоритмов обработки текстовых данных. При анонсе каждой программы разработчики указывают набор функций и спектр задач, которые она выполняет. Несмотря на различия в функционале прослеживается одна тенденция – синтез эффективных алгоритмов статистики, компьютерной лингвистики и машинного обучения (включая глубокое обучение и нейросети).

Ниже представлено наиболее популярное коммерческое и некоммерческое программное обеспечение (либо компании, разрабатывающие программный продукт): MonkeyLearn, LIWC, MEM, Meaning Cloud, SAS Text Miner, IBM Watson, SPSS Modeler Text analytics, Lexalytics, Knime, Rosette Text Analytics Platform, Sketch engine, Tisane, Twinword, Amazon Comprehend, NVivo, Luminoso, DiscoverText, MaxQDA, Aylien, Atlas.ti, Google cloud natural language, GATE, Alceste, Clustify, Eagle Online, Full Text Mapper, Cogito Discover, Intellexer, Keatext, Leximancer, Linguamatics, Loop Q, QDA Miner, Yoshikoder, Chinese

Text Analyser, Voyant Tools, Textometrie, Tagtog, Wordle, Ikanow, S-EM, LingPipe, VisualText2.0, Wmatrix, TagCrowd, Power Text Solution, HyperPo, Angoos Knowledge Reader, Cat Coding Analysis Toolkit, KH Coder, TAMS Analyzer, Textable (Orange Kanvas), TokenX.

Отдельно следует выделить программы текстовой аналитики, работающие с китайским языком: LIWC, SAS Text Miner, SPSS Modeler Text analytics, Lexalytics, Rosette Text Analytics Platform, MonkeyLearn, Amazon Comprehend, NVivo, Luminoso, DiscoverText, Google Cloud Natural Language, MaxQDA, Yoshikoder, Chinese Text Analyser, Atlas.ti, Voyant Tools, GATE, S-EM, LingPipe, KH Coder.

Указанные программы в целом выполняют схожие функции либо дополняют друг друга. Разница между ними, как правило, кроется в деталях и нюансах работы, а также заложенных алгоритмах. Разные программные решения могут выполнять одну и ту же функцию, но работать по алгоритмам разной эффективности. По этой причине, как правило, исследователи выбирают те программные решения, которые известны, доказали свою эффективность, имеют значимые научные результаты и принадлежат крупным компаниям или open source-проектам. Важным является наличие в программе размеченного и эмпирически обоснованного словаря.

Ниже представлен обзор функций, который выполняют данные программы (таблица 1). Эти функции являются программной реализацией методов и алгоритмов обработки текстовых данных.

Таблица 1. Обзор функций программ

Функции	**Описание**	**Предлагаемое применение в маркетинговых исследованиях/исследованиях клиентов**
Логические запросы	Тип поиска, который позволяет комбинировать слова с логическими операторами И или НЕТ для создания более релевантных запросов. Это ограничивает результаты поиска только теми документами, которые содержат логическое выражение, то есть два или более ключевых слов.	Он имеет самое широкое применение, так как предназначен для создания условий логического поиска и фильтрации информации в базах данных. В основном он используется для уменьшения обрабатываемого объема данных и времени обработки.
Фильтрация документов	Под фильтрацией документов понимается процесс, посредством которого система отслеживает поток входящих документов, классифицирует их по содержанию. Затем отбирает те, которые считаются актуальными для конкретного пользователя или темы. Эта функция позволяет отфильтровывать ненужную информацию и организовывать важную информацию по соответствующим категориям.	Функция аналогична предыдущей с той разницей, что она фильтрует и классифицирует документы по их содержанию или тематической направленности. Она имеет самое широкое применение. В основном используется для уменьшения обрабатываемого объема данных и времени обработки.
Распознавание языка	Определение языка, на котором написан текст. Обычно используют алгоритмы классификации текста	Это актуально в кросс-культурных маркетинговых исследованиях при работе с документами на разных языках. Определение языка позволяет выбрать необходимый словарь для дальнейшего исследования. Также используется в межкультурных исследованиях в области рекламы, продвижения и распространения продуктов
Анализ настроений	Анализ тональности текста, предназначенный для автоматизированного выявления эмоционально окрашенной лексики в текстах и оценки авторов по отношению к обсуждаемым в тексте объектам.	Позволяет определить отношение писателя или докладчика к продуктам, компаниям или событиям. Частично для уточнения психологического профиля человека. Позволяет задать реакцию потребителя на действие. Используется для разработки стратегий в рекламе, продвижении, таргетинге, сегментации, позиционировании.
Обобщение	Процесс уменьшения размера текста без потери смысла или поиск подмножества данных, содержащего информацию всего набора; своего рода создание репрезентативной выборки или резюме.	Он имеет самое широкое применение. Обычно используется для получения краткого описания (набора ключевых слов) клиентов, конкурентов, влиятельных лиц, референтных групп, сообществ, продуктов или брендов.

Маркировка тегами	Процесс маркировки слов или фраз. Тег, или метка — это своего рода метаданные, которые помогают описать элемент, а затем найти его при просмотре или поиске.	Процесс сопоставления любого слова с любой маркетинговой категорией. Например, создание словаря для отнесения целевого клиента по его описанию или параметрам к определенному сегменту; создание словаря для отнесения конкурентов к разным группам по описанию или ключевым словам.
Классификация	Иногда это называется маркировкой текста, или категоризацией текста. Это процесс разделения текста на организованные группы. Классификаторы текста могут автоматически анализировать текст. Затем определяется набор предопределенных тегов или категорий на основе его содержимого. Классификация текста включает определение темы, анализ настроений, определение языка. При классификации текста документ или фрагмент текста можно отнести к одному или нескольким классам или категориям.	Используется для поиска целевых сегментов или групп потребителей по признакам или классификации клиентов по интересам и характеру реакции.
Тематическая кластеризация (тематическое моделирование)	Процесс группировки содержимого документа или документов по темам или подтемам, в результате которого формируется тематический кластер, показывающий тесно связанное содержание.	Он используется для формирования групп пользователей или клиентов, которые обсуждают схожие темы, выражают схожие мысли или имеют одинаковые интересы, для управления жизненным циклом клиента, брендинга и контроля поведения потребителей.
Анализ сущности	Анализ сущностей, или распознавание названных сущностей (Распознавание именованных сущностей (NER)). Задача извлечения информации, направленная на поиск и классификацию ссылок на именованные объекты в неструктурированном тексте по заранее определенным категориям, таким как имена людей, организаций, местоположения, различные коды, даты, денежные значения, проценты и т. д.	Используется для поиска названий компаний, продуктов, мест, событий, дат, имен частных лиц и т. д., их упоминаний в социальных сетях или СМИ. Также для мониторинга конкурентов.
Графическое представление данных	Возможность визуализировать данные, полученные в ходе анализа текста (например, облако слов или частоту встречаемости слов).	Используется для сравнительного анализа частоты упоминаний брендов или отдельных продуктов. Также используется для сравнительного анализа отношения потребителей к продукции, конкурентам и маркетинговой деятельности компаний. Облака слов удобны для настройки ключевых слов, с помощью которых потребители описывают события, акции, продукты.
Категоризация	Общее направление категоризации текста еще называют задачей кластеризации. Процесс очень похож на классификацию с той разницей, что границы категорий размыты по сравнению с границами классов и устанавливаются не по формальным признакам, а путем сравнения категорий друг с другом.	Используется для формирования обобщающих признаков с целью создания категорий, например, для создания и уточнения категорий или типов заказчиков/покупателей; определения признаков и моделей поведения, общих для нескольких групп потребителей; поиска общих признаков в сообществах в социальных сетях.
Извлечение	Название общего направления извлечения текста или слов из документа.	Извлекает любую релевантную маркетинговую информацию из неструктурированного текста.

Прогнозное моделирование	Функция позволяет создавать модели прогнозной аналитики. Это означает, например, прогнозирование того, принадлежит ли тег/метка или текст определенной теме, принадлежит ли извлеченное слово или часть текста определенной метке, соответствует ли извлеченный текст запросу или соответствует ли содержимое запросу.	Обычно используется для соотнесения клиента (обычно нового) или потребителя, а также интересов, публикаций и мнений с ранее известной группой или категорией.
Аспектный анализ настроений	Он определяет не общую тональность текста, а различные аспекты тональности каждой части текста. Обеспечивает более детальный анализ текста. Другими словами, выявляет высказывает положительные или отрицательные мнения по различным темам или аспектам чего-либо.	Используется для установления тональности (эмоциональности) текста/поста по отношению к какому-либо аспекту, например, положительным или отрицательным эмоциям в рамках предоставления дополнительной услуги, добавления новой функции продукта; в связи с рекламной акцией, мероприятием с конкурентами или деловыми партнерами.
Анализ настроений на уровне организации	Этот анализ определяет не общую тональность текста, а тональность текста относительно конкретного именованного объекта.	Используется для определения тональности или эмоций в отношении названий, например брендов, торговых марок, названий компаний или событий.
Анализ настроений на уровне документа	Анализ тональности на уровне документа, то есть возможность определить тональность всего документа.	Используется для определения тональности любого документа, например, файла с деловой перепиской, файла с описанием конкурента или описания клиента, файла с отзывами или комментариями к товару.
Предложение хэштегов	Функция автоматического предложения актуальных хэштегов для публикации контента в социальных сетях.	Используется для анализа сообщений и автоматического создания хэш-тегов.
Маркировка изображений	Отмечает не только текст, но и изображения, найденные на веб-страницах.	Используется для анализа сообщений и автоматического создания хэш-тегов для изображений.
Семантическое сходство	Поиск не по ключевым словам, а по смыслу. Если слова неоднозначны, это вызывает определенные трудности. Семантический поиск позволяет искать не по словам, а по смыслу. Используется в основном для межъязыкового поиска, когда не может быть полной лексической переводимости слов. А также для поиска соответствующих терминов и понятий или их генерации на других языках, для поиска повторов в документах, для поиска плагиата.	Используется в международных маркетинговых исследованиях для поиска информации по смыслу, как правило, для межъязыкового поиска, когда нет полной лексической переводимости слов.
Анализ субъективности	Задача, связанная с анализом настроений, основная цель которого — обозначить мнение как субъективное или объективное.	Используется для определения субъективности или объективности мнения, выраженного в комментариях или сообщениях. Удобно при работе с возражениями и жалобами, для повышения уровня обслуживания клиентов.
Психометрический анализ текста	Определение психометрических свойств текста, то есть основных психологических/когнитивных характеристик автора, в частности уровня аналитического мышления, уровня лидерства, степени честности и открытости и эмоционального фона.	Используется для составления психологического профиля человека, определения его личностных качеств и установления параметров целевых групп и сегментов пользователей или клиентов.

Основные функции анализа текста и их прикладные аспекты можно обобщить и представить следующим образом (рисунок 2).

В целом наиболее ценными функциями являются психометрический анализ, анализ настроений и тематическое моделирование. Они позволяют наиболее точно описать профиль целевого клиента, понять и установить свою целевую аудиторию – основу для разработки комплекса маркетинга. Для получения более точной информации о целевом клиенте целесообразнее использовать узкоспециализированные программные продукты. Они относятся к программные решениям по аналитике социальных медиа, которые включают и аналитику социальных сетей и имеют более совершенные алгоритмы.

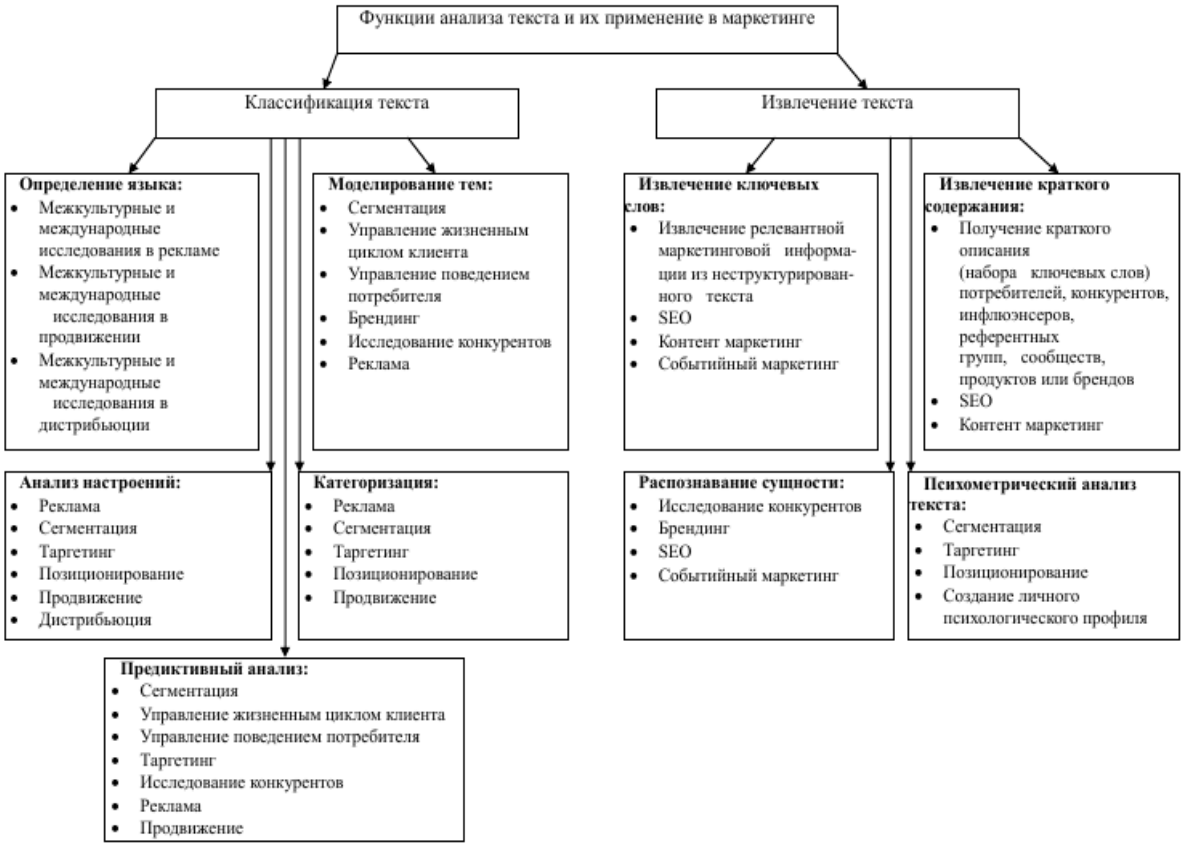

Рисунок 2 – Функции анализа текста и их применение в маркетинге
Источник: собственная разработка авторов

Заключение

Современные методы и алгоритмы обработки текста в историческом срезе ушли достаточно далеко в своем развитии. Изначально, нацелившись на установление авторства, стилометрию, стенографию и шифрование, анализ текста постепенно распространялся и на другие сферы научных исследований.

Ключевой вехой в становлении текстовой аналитики стало появление вычислительных машин и бурное развитие психолингвистики. Именно это позволило автоматизировать анализ текста и связать его с поведением и внутренним миром человека. Сознательное и бессознательное проявляется в мыслях, речи, решениях и поступках личности, что находит отражение в тексте.

Современные программные решения проникли и широко используются во всех сферах, где встречается текст, а теперь уже аудио- либо видеозапись. С точки зрения вычислительных устройств любая информация представляет собой нули и единицы, поэтому изображения, аудио- и видеозаписи дополняют текстовую информацию и в определённом смысле между ними можно поставить знак равенства.

Современные программные средства способны использовать различные форматы данных и комплексные алгоритмы анализа, позволяя точнее описывать человека, особенности его поведения, психологические черты и черты личности. Эта информация извлекается из социальных сетей, берется на вооружение бизнесом и используется для сегментации потребителей, оптимизации рекламного воздействия, исследования конкурентов, таргетирования и позиционирования продуктов, их продвижения и других коммерчески значимых действий.

Литература

1. Bar-Hillel Y. (1960) A demonstration of the nonfeasibility of fully automatic high quality translation. In Advances in Computers, ed. FL Alt, 1:158-163. New York: Academic.
2. Bourdon B. L'expression des émotions et des tendances dans le langage, Félix Alcan; 1892., цит. по McLaughlin JE (2022).
3. Boyd, R. (2017) Psychological Text Analysis in the Digital Humanities. 10.1007/978-3-319-54499-1_7.
4. Boyd, R.L. MEH: Meaning Extraction Helper (Version 1.4.13) [Software]. Available from http://meh.ryanb.cc (2016).
5. Chung, C., Pennebaker, J. (2007) The psychological functions of function words. In K. Fielder (Ed.), Frontiers in social psychology, pp. 343-359. New York: Psychology Press.
6. Chung, C.K., Pennebaker, J.W. Revealing dimensions of thinking in open-ended self-descriptions: an automated meaning extraction method for natural language. J. Res. Pers. 42(1), 96-132, (2008).
7. Eichstaedt J.C., Smith R.J., Merchant R.M., Ungar L.H., Crutchley P., Preoţiuc-Pietro D., Asch D.A., Schwartz H.A. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A. 2018 Oct 30; 115(44): 11203-11208.
8. Estoup J.B. (1916) Gammes sténografiques. Institut sténografiques de France, Paris.
9. Giorgi S., Nguyen K.L., Eichstaedt J.C., Kern M.L., Yaden D.B., Kosinski M., Seligman MEP, Ungar LH, Schwartz HA, Park G. Regional personality assessment through social media language. J Pers. 2022 Jun; 90(3): 405-425.
10. Hearst, M.A. Text Data Mining. In The Oxford Handbook of Computational Linguistics; Mitkov, R., Ed.; Oxford University Press: Oxford, UK, 2005; pp. 616-662.
11. Holmes, D.I. (1998) The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3), p. 111-117.
12. Iezzi, D.F., Celardo, L. (2020) Text Analytics: Present, Past and Future. In: Iezzi, D.F., Mayaffre, D., Misuraca, M. (eds) Text Analytics. JADT 2018. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham.
13. Iliev, R., Dehghani, M., & Sagi, E. (2014) Automated text analysis in psychology: methods, applications, and future developments. Language and Cognition, 7(02), 265-290.
14. Kumiko Tanaka-Ishii, Shunsuke Aihara; Computational Constancy Measures of Texts – Yule's K and Rényi's Entropy. Computational Linguistics 2015; 41 (3): 481-502.
15. Mandelbrot, B., Les constantes, chiffrées du discours, Encyclopédie de la Pléiade Le Langage, vol, publié sous la direction d'A. Martinet, Paris, Gallimard, 1966, pp. 46-56.
16. McLaughlin J.E., Lyons K, Lupton-Smith C., Fuller K. An introduction to text analytics for educators. Curr Pharm Teach Learn. 2022 Oct; 14(10): 1319-1325.
17. Miller, G.A. (1999) ON KNOWING A WORD. Annual Review of Psychology, 50(1), 1-19.
18. Miner G. Practical text mining and statistical analysis for non-structured text data applications. – Academic Press, 2012.
19. Osgood, C.E., Sebeok, T.A. Psycholinguistics: A survey of theory and research problems. – Indiana University Press, 1965.
20. Pennebaker, J., Francis, M., Booth, R. (1999) Linguistic inquiry and word count (LIWC).
21. Pennebaker, J.W., King, L.A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77(6), 1296–1312.
22. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G. (2003) Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 54(1), 547-577.
23. Petruszewycz, M. L'histoire de la loi d'Estoup-Zipf: documents, Mathématiques et sciences humaines, vol. 44, 1973, p. 41-56.
24. Pronko, N.H. (1946) Language and psycholinguistics: a review. Psychological Bulletin, 43(3), 189-239.
25. Reece, A.G., Reagan, A.J., Lix et al. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 7, 13006 (2017).
26. Saha, K., Yousuf, A., Boyd, R.L. et al. Social Media Discussions Predict Mental Health Consultations on College Campuses. Sci Rep 12, 123 (2022).
27. Soto, C.J., Jackson, J.J. (2020). Five-factor model of personality. In Dana S. Dunn (Ed.), Oxford Bibliographies in Psychology. New York, NY: Oxford.
28. Tan, A.-H. Text Mining: The state of the art and the challenges, Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, Beijing, 1999, pp. 65-70.
29. Willem J.M., Levelt A. History of Psycholinguistics: The Pre-Chomskyan Era. Oxford University Press, 2013, p. 653.
30. Yule, G.U. (1938) On sentence-length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. Biometrika, 30: 363-90.
31. Zipf, G.K. (1929) Relative frequency as a determinant of phonetic change. Harvard Studies in Classical Philology, 40, 1-95.

32. Zipf, G.K. (1932) Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA.
33. Zipf, G.K. (1945) The Meaning-Frequency Relationship of Words. The Journal of General Psychology, 33(2), 251-256.

Ключевые слова
Психометрические характеристики, исследование аудитории, маркетинговые исследования, искусственный интеллект

Василий Кашкин – к.э.н., доцент, руководитель агентства международных исследований Kashkin.com.cn,
ORCID: 0000-0002-7568-6188,
vasily@kashkin.com.cn

Юрий Андросик – старший аналитик агентства международных исследований Kashkin.com.cn,
ORCID: 0009-0009-1474-9315,
cosadesl@gmail.com

Vasily Kashkin, Yuri Androsik, Review of the development of text analysis methods: from manual processing in psycholinguistics to modern automated programs in marketing

Keywords
Automated text analysis, text analytics, software, text analysis functions, psycholinguistics, marketing research, marketing analysis.

DOI: 10.34706/DE-2023-05-10
JEL classification – C81 – Методология сбора, оценки и организации микроэкономических данных. Анализ данных

Abstract
The article provides a brief historical overview of the development of text analysis methods. Psycholinguistic research and information technology played a significant role in this. They were the decisive factors that determined the current state of this area of scientific research. Psycholinguistic research has made it possible to connect first written text, and later oral speech, with the psychological characteristics of the individual, his personality traits and thinking. Which subsequently, combined with the capabilities of modern computers, gave new impetus to the study of human behavior, expanding the boundaries of scientific research in economics and marketing. The authors tested and analyzed over 50 text analysis software products, identified the main functions of the software and suggested the main directions for their use in marketing research.