ЛИНГВИСТИКА

Н. А. Миронова,
магистр филологических наук,
преподаватель кафедры иностранных языков
Учреждение образования Федерации профсоюзов Беларуси
«Международный университет «МИТСО», г. Минск

АВТОМАТИЗАЦИЯ СОЦИОЛИНГВИСТИЧЕСКОЙ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА В ПИСЬМЕННОЙ ВИРТУАЛЬНОЙ КОММУНИКАЦИИ

Аннотация: Данная статья посвящена проблеме автоматической идентификации автора в письменной виртуальной коммуникации в контексте компьютерной социолингвистики, представленной как одно из наиболее перспективных и инновационных междисциплинарных научных направлений.
Ключевые слова: социолингвистика, компьютерная лингвистика, компьютерная социолингвистика, идентификация автора текста, виртуальная коммуникация.

Annotation: The abstract is devoted to the problem of automatic author profiling in written virtual communication in the context of computational sociolinguistics, which is represented as one of the most perspective and innovative interdisciplinary scientific studies.
Key words: sociolinguistics, computational linguistics, computational sociolinguistics, author profiling, virtual communication.

Коммуникация в наши дни характеризуется тенденциозной динамичностью в специфике ее осуществления: проще говоря, приоритетные среди большинства людей способы обмена информацией меняются, а средства его осуществления с течением времени приобретают статус символа технократической эпохи. В этой связи, разумеется, нельзя не рассматривать виртуальную коммуникацию и значение киберлингвистики в ней, в особенности в контексте вопроса о деанонимизации.

Огромное количество современных пользователей самостоятельно публикуют значительное количество личной информации в интернете. В частности, в социальных сетях часто указаны настоящие имя и фамилия, также могут быть публично доступны фотографии (из которых можно сделать вывод о связях с другими людьми), телефоны, информация об учебных заведениях, группах, иллюстрирующих интересы пользователя. Многое может принести поиск с помощью соответствующих систем по электронной почте или редкому имени пользователя. Часто одно и то же имя используется в социальных сетях, на различных форумах и в иных сервисах,

что позволяет найти некоторое ограниченное множество «подозреваемых» пользователей. Также информация о владельце сайта содержится в сервисе WHOIS (правда, многие регистраторы предоставляют опцию сокрытия данных владельца). Идентификация пользователя (не реального лица) может быть произведена по статическому IP-адресу, использованным cookie-файлам (которые сохраняют информацию о сессии на сайте), ответам протокола TCP. Все интернет-сайты предоставляют возможность проверки IP, с которых посещался сайт. При помощи таких способов проводятся деанонимизации – раскрытие информации, которая доступна всем желающим в сети Интернет, при том что сами пользователи не предусматривали возможность ее сбора и систематизации. Также деанонимизации способствует поиск по ключевым фразам, сопоставление различной информации, имеющейся в разрозненном виде, анализ стилистики высказываний пользователей, личная переписка. В Российской Федерации система СОРМ записывает весь интернет-трафик, что также позволяет легко распознать автора любых сообщений. Аналогичные средства имеются в распоряжении полиции и специальных служб иных государств. Более того, интернет-провайдеры обязаны предоставлять следствию все возможные данные о пользователе.

Благодаря прогрессирующему распространению технологий, обеспечивающих виртуальную коммуникацию, стали возникать и сопутствующие этому процессу социальные и психологические феномены (например: catfish), что, в свою очередь, прекрасно иллюстрирует факт: с точки зрения социолингвистики и киберлингвистики социальное пространство в интернете стало синергетической системой, отдельным сегментом ноосферы, характеризующимся конкретными, постоянно возникающими явлениями и закономерностями, возросшей значимостью поиска инновационных методов атрибуции нелитературного текста и расширения возможностей систем искусственного интеллекта в целом.

Данная работа посвящена процессу разработки автоматической системы социолингвистической идентификации автора в виртуальной коммуникации, позволяющей выявлять информацию об определенных социальных показателях автора из англоязычных текстов интернета. Ее результатом стало создание компьютерной программы LylBrother, написанной на языке программирования Python 3.6. Материалом разработки послужили 150 текстовых сообщений в англоязычной письменной разговорной речи. В рамках компьютерного моделирования была создана компьютерная программа, способная определять вероятную принадлежность автора текста к конкретным социальным группам:

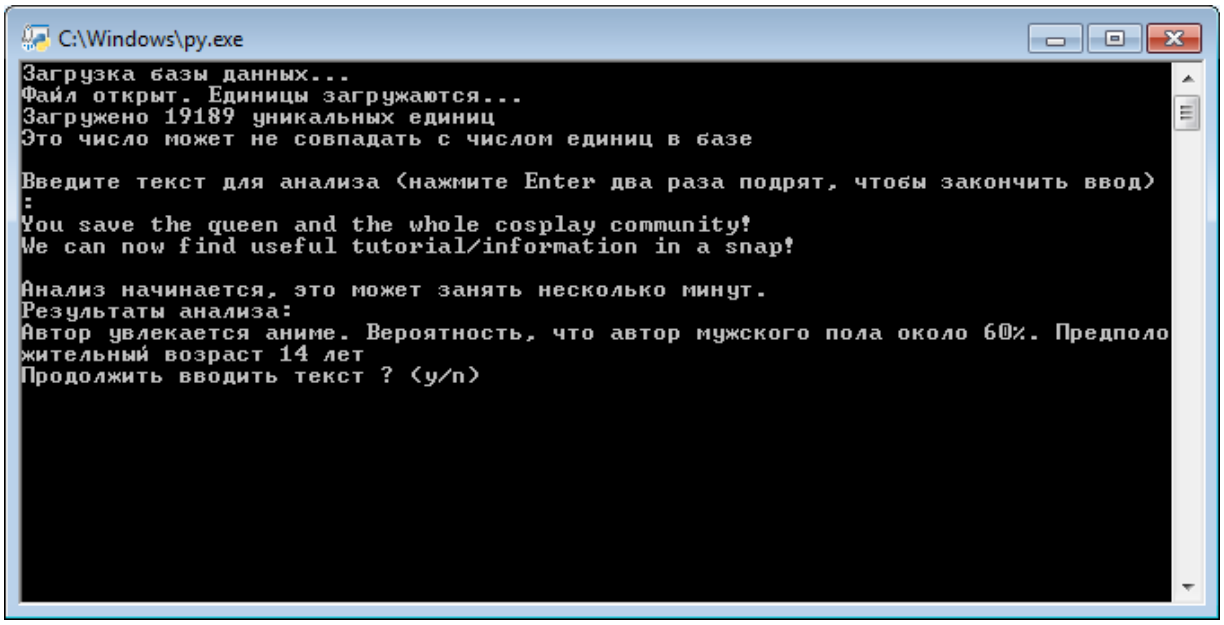

Рисунок 1 – Пример работы программы социолингвистической идентификации текста

Предложенная модель компьютерной системы работает с опорой на лингвистическую базу данных, которая имеет алфавитный словарь языковых единиц, включающий в себя, помимо слов, также лексикализованные аббревиатуры и символьные конгломераты. Отбор единиц базы данных проводился с опорой на знание английского языка, ряд словарей и инструментов, таких как, например, UrbanThesaurus, а также с помощью дальнейшей отладки с помощью тестирования. Объектом исследования в работе являлась система социолингвистических средств, составляющих основу автоматической системы социолингвистической идентификации автора текстового сообщения в виртуальной коммуникации. Предметом анализа является система языковых средств письменной коммуникации, составляющих лингвистическое обеспечение системы социолингвистической идентификации автора текстового сообщения в виртуальной коммуникации. Научная новизна состоит в формировании нового подхода к социолингвистической идентификации автора текстовых сообщений в сети Интернет, объединяющего в себе как традиционные, так и новые техники. Теоретическая значимость исследования заключается в том, что полученные результаты могут стать теоретической базой для последующих разработок в представленной области, а также использоваться при обучении прикладной лингвистике и компьютерной социолингвистике. Практическая значимость исследования заключается в возможности применения его результатов при создании деанонимизационного программного обеспечения, а также во внедрении представленной компьютерной программы в более крупные программы, такие как, например, виртуальные помощники, что позволит обычному пользователю избежать столкновения с феноменом «catfish» [3, с. 4–6].

Итак, значительно возрос интерес пользователей сети Интернет к феномену «catfish» и способам его избегания, и, как следствие, возросла востребованность программного обеспечения, позволяющего идентифицировать социальный статус собеседника. Разработанное информационное и лингвистическое обеспечение можно использовать для создания программ, идентифицирующих принадлежность виртуального собеседника к конкретным социальным группам. Также алгоритм можно интегрировать с различным открытым программным обеспечением («виртуальными помощниками») либо создавать утилиты на его основе. Следует отметить возросшую значимость вопроса деанонимизации в криминалистике в последние годы.

Интенсивное развитие интернет-технологий за последние два десятилетия, а также все более быстрое проникновение интернета в повседневную жизнь каждого человека вызывают необходимость осмысления тех языковых особенностей, которые характеризуют общение и взаимодействие пользователей в онлайн-среде. Для наиболее полного понимания специфики письменной разговорной речи в виртуальной коммуникации следует определить суть каждого компонента этого комплексного понятия. Ядром такового является непосредственно текст.

Текст в виртуальной коммуникации можно называть письменным сообщением, объективированным в виде письменного документа, состоящим из ряда высказываний, объединенных разными типами лексической, грамматической и логической связи, имеющим определенный моральный характер, прагматическую установку и, соответственно, литературно обработанным [2, с. 67]. Распространенное среди исследователей мнение о том, что текст любого размера – это относительно автономное (законченное) высказывание, теряет свою актуальность, как и то, что к тексту можно подобрать заголовок. Правильно оформленный текст обычно имеет начало и конец, однако существует ряд условий и особенностей, когда речь идет о так называемой письменной разговорной речи.

Для письменной разговорной речи в виртуальной коммуникации характерны эллиптические предложения, эмфатический порядок слов использование разговорной лексики, краткие формы служебных частей речи, неправильный порядок слов. Общение в интернете совмещает в себе различные характеристики устной и письменной речи, не являясь в полной мере ни тем, ни другим. Объясняется это, разумеется, тем, что современные технологии впервые обеспечили возможность использования письма для мгновенной передачи информации, а также обеспечили приватное письменное общение двух собеседников. Это создало ситуацию, когда в условиях, характерных для устного общения, стало возможным использовать письмо [1, с. 63–67; 4, с. 70–89].

Идентификацией пользователей в интернете можно назвать использование набора методик и способов, позволяющего получить информацию о пользователе интернета из открытых источников. Средства

идентификации могут быть как техническими, так и лингвистическими. С помощью технических средств возможно определить некоторые передаваемые серверу характеристики браузера (тип, язык, встроенные расширения, поддержка приложений), просматриваемую страницу, ссылающуюся страницу, IP-адрес, данные прокси-сервера, поддержку cookie и Java, часовой пояс и др. С помощью лингвистических средств можно определить так называемую «языковую личность» (провести языковое портретирование), конкретного автора текста (при условии наличия нулевой гипотезы и узкого круга потенциальных авторов).

Личность детерминируется посредством специфических факторов: классовая принадлежность, принадлежность к социальным институтам, профессиональным общностям и т. д. Наиболее важным из этих факторов является классовая и слоевая принадлежность, которая в свою очередь детерминирует непосредственное окружение индивида – систему малых групп, в которых протекает его социальная деятельность (семья, трудовой коллектив, группы для удовлетворения совместных интересов и т. п.). К числу социологически существенных специфических факторов относятся также пол и возраст. Принадлежность ко всем вышеперечисленным группам с разной точностью можно идентифицировать лингвистически [5, с. 200, 201].

Существуют возможности применения в социолингвистике научных методов социального прогнозирования – экстраполяции, аналогии, моделирования. Экстраполяция способна дать нам определенные знания о будущем, поскольку оно всегда в известном смысле является продолжением настоящего, но поскольку будущее вместе с тем является и отрицанием настоящего, объективная ценность экстраполяции носит ограниченный характер. Аналогия улавливает определенную повторяемость в поступательном развитии общества, но не в состоянии предугадать возникновение чего-то существенно нового. Моделирование позволяет ограничить круг реальных вариантов, но само по себе не предопределяет их вероятности.

Традиционно в компьютерной и социолингвистике для профилирования автора текста (определения его социальных и психологических показателей) используется лингвостатистический анализ текста. Одними из наиболее актуальных показателей для анализа являются пол, возраст и психологическое (или психическое) состояние. Одним из наиболее популярных идентификаторов пола и возраста традиционно считают употребление автором служебных частей речи в письменной разговорной коммуникации (например, считается, что более частое употребление личных местоимений характерно для женщин). Однако данный идентификатор недостоверен, так как при изменении психологического состояния меняется и «языковая личность» автора, что приводит к ошибкам в идентификации. Наиболее точным идентификатором возраста автора текста в письменной виртуальной коммуникации можно считать частоту употребления так называемых «трендовых» единиц. Существует зависимость

между употреблением акронимов, сленгизмов и неологизмов и возрастом автора: так как чем старше человек, тем менее он подвержен трендам, соответственно, чем младше автор – тем больше трендовых лексических единиц он употребляет в письменной разговорной речи. Чем больше автор использует акронимы, сленгизмы, неологизмы и графические способы выражения эмоций, тем моложе он является [3, с. 44].

Идентификатором сферы интересов либо занятости в письменной виртуальной коммуникации можно считать специфические лексические единицы, известные и используемые обычно только участниками конкретной социальной группы. Идентификатором пола можно также считать данные единицы при условии наличия привязки статистических данных к корреляции гендерной группы и группы по интересам либо сфере занятости.

Нельзя не отметить тот факт, что представляется возможным совмещение традиционных техник профилирования автора с техниками, предложенными в данной работе, для повышения точности результатов [3, с. 46].

Статистический анализ специфических языковых единиц, употребление которых характерно для определенной социальной группы, позволяет определять не только вероятную сферу интересов/занятости автора, но и иные социальные показатели, такие как гендер или возраст. Наибольшую точность при социолингвистической идентификации автора даст комбинация из традиционных методов профилирования автора текста (например, по служебным частям речи) и методов, предложенных в данной работе при интеграции элементов интеллектуального анализа текста и машинного обучения [3, с. 53–56].

Перспективы применения и развития идентификационного программного обеспечения в сфере компьютерной социолингвистики зависят от специфики построения и создания такого обеспечения. Открытость системы, возможность свободной отладки и интеграции системы как модуля повышают потенциал системы. При пополнении лингвистической базы данных и других списков, используемых системой в процессе работы, компьютер сможет более точно выделять необходимую для поиска информацию. Точность работы системы можно повысить также путем добавления дополнительных статистических данных (вероятности принадлежности представителя конкретной социальной группы по сфере интересов/занятости к конкретным возрастным и гендерным группам) и создания новых категорий в базе данных [3, с. 28–60].

Возможность применения полученных в исследовании результатов видится во внедрении представленной компьютерной программы в более крупные идентификационные системы с целью оперативно и точно извлекать информацию из текстов на естественном языке, а также в дальнейшем производить портретирование языковой личности автора.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ахманова, О. С. Словарь лингвистических терминов / О. С. Ахманова. – М. : Советская энциклопедия, 1969. – 605 с.

2. Гальперин, И. Р. О понятии «текст» / И. Р. Гальперин // Лингвистика текста : материалы науч. конф. – М., 1974. – Т. 1. – С. 67–72.

3. Ивашко, Н. А. Лингвистическое и информационное обеспечение автоматической системы социолингвистической идентификации автора текста : автореф. дис. … магистра филол. наук / Н. А. Ивашко ; Минский гос. лингвистический ун-т, Минск, 2016. – 78 с.

4. Chrystal, D. Language and the Internet / D. Chrystal. – Cambridge : Cambridge Univ. Press, 2006. – 257 p.

5. Davis, J. Electronic discourse: linguistic individuals in virtual space / J. Davis [etc.]. – Albany, NY : State University of New York Press. – 217 p.