ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доктора филологических наук, профессора
Колмогоровой Анастасии Владимировны
о диссертации Мамаева Ивана Дмитриевича на тему: «Лингвистическое исследование скрытых сообществ в корпусе социальных медиа с применением мультимодальных тематических моделей», представленной на соискание ученой степени кандидата филологических наук по специальности 5.9.8. Теоретическая, прикладная и сравнительно-сопоставительная лингвистика

Рецензируемая диссертация посвящена описанию принципов выявления скрытых сообществ и описания их лингвистических профилей.

Актуальность рассматриваемой проблематики связана с прикладными аспектами компьютерной лингвистики, а именно: во-первых, выявление скрытых сообществ в социальных сетях представляет особую важность для обеспечения национальной безопасности; во-вторых, в такой технологии заинтересованы коммерческие и государственные структуры, занимающиеся мониторингом и прогнозированием социальных процессов.

Новизна полученных результатов связана с задачей лингвистического профилирования сообществ, имеющих разные интересы, с помощью метрик лингвистической статистики. Для выявления самих сообществ используется метод тематического моделирования. Метод уже достаточно хорошо зарекомендовал себя в качестве эффективного инструмента выявления в социальных сетях социальный акторов, представляющих потенциальную группу риска для социума. Так, в работах научной группы Е.Ю. Кольцовой при помощи алгоритмов тематического моделирования выявлялись зоны межэтнического напряжения и группы, вовлеченные в создание данного напряжения, был создан сентимент-словарь для подобных сообществ [Nikolenko, Koltsova, Koltsov, 2019; Koltsova 2023]. Однако в рецензируемой работе данные алгоритмы используются только как первый этап (этап

кластеризации текстов в группы), а дальнейшая задача ставится иначе, а именно – описать лингвистический профиль каждой группы. В отличие от работ Т.А. Литвиновой по выявлению в социальных сетях групп лиц экстремистской направленности на основе применения методов стилометрии [Литвинова 2019, 2020], в рецензируемой работе проблематика рассматривается шире, а фокус смещен с алгоритмов стилометрического анализа (разновидности дельты Берроуза и т.д.) на лингвостатистические характеристики текстов скрытых сообществ: учитываются внутритекстовые морфологические корреляции средних значений употреблений имен существительных и имен прилагательных, глаголов и наречий; синтаксическая корреляция средних значений длины предложений и длины структур зависимостей; лексическая корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия». Предложенная И.Д. Мамаевым комбинация тематического моделирования и внутритекстовых лингвостатистических метрик представляется новой и интересной.

Методологический аппарат работы выстроен ясно и непротиворечиво, согласуется со структурой работы и ее текстовой организацией.

Положения, выносимые на защиту, полностью находят свое отражение в тексте диссертации. Так, Положение 1 опирается на проведенный в Главе 1 детальный анализ научной литературы, посвященной лингвистической специфике интернет-коммуникации. Положение 2, постулирующее эффективность использования выработанных автором критериев для создания репрезентативной выборки, раскрывается в Главах 2 и 3. Положения 4-8 суммируют результаты, полученные соискателем уже при аналитической работе с алгоритмами тематического моделирования и лингвостатистическими метриками внутри каждой тематической коллекции и изложенные в Главе 3. Закономерно, что положения 7-8 возвращают читателя, но уже на новом уровне обобщения к Положению 1, уточняя и

конкретизируя его. При общей логической стройности данной части работы, на мой взгляд, Положения 7 и 8 можно было бы объединить.

К числу наиболее значимых результатов, полученных в работе, отнесем следующие:

– разработана новая научная идея о том, что сообщества людей, имеющих схожие интересы и публикующих с определенной регулярностью посты в социальных сетях, но незнакомых друг с другом непосредственно, потенциально детектируемы с помощью алгоритмов тематического моделирования, примененных к обширному корпусу публикаций пользователей;

– предложена оригинальная научная гипотеза о том, что у каждого такого сообщества есть доминирующая тема, а у его текстовой продукции, образующей массив текстовых данных, – специфический лингвистический профиль, включающий не только лексические маркеры, но и специфические лингвостатистические характеристики на уровне морфологии, синтаксиса и распределения лексических единиц;

– доказана перспективность использования разработанного перечня критериев отбора текстов социальных сетей для последующего детектирования так называемых скрытых сообществ;

– представлена типология из 23 скрытых сообществ, обнаруженных в исследованном материале, а также описаны их лингвистические профили.

Таким образом, теоретическая значимость исследования видится, прежде всего, в том, что разработана методология выявления скрытых сообществ с помощью применения к их текстовой продукции алгоритмов тематического моделирования; доказано наличие у текстов большинства таких сообществ собственного лингвистического профиля, а также предложена система линвостатистических метрик, позволяющих рассматривать сообщества как распределенную систему и анализировать степень близости лингвистических характеристик продуцируемых в сообществе текстов.

Практическая ценность также весома: предложены перечни лингвистических маркеров разных сообществ, описаны лингвистические профили сообществ, доказана целесообразность применения подсчета значений внутритекстовой корреляции имен существительных и прилагательных, с одной стороны, а также глаголов и наречий – с другой для дифференциации групп пользователей. Особенно интересными мне представляются рис. 50-52 (КД), где показаны расстояния между тематическими сообществами и кластеры, которые они образуют. Например, тематическое сообщество «политика» входит в один кластер с «эзотерикой» и «историей», а «культура и искусство» – с «экономикой и финансами» (рис. 52). Думаю, что это может стать интересным инструментом для наблюдения за динамикой сближения способов говорить о темах (способах дискурсивизации тем) на определенных отрезках жизни общества.

Личный вклад соискателя состоит в непосредственном участии в сборе материала исследования, его описании, систематизации, валидации и апробации в рамках научных мероприятий различного уровня. Основные положения и результаты работы отражены в семи публикациях, три из которых – в научных журналах, входящих в Перечень рецензируемых научных изданий Высшей аттестационной комиссии Министерства науки и высшего образования (по научной специальности 5.9.8), а также в изданиях, индексируемых в международной базе данных Scopus.

Особо хотелось бы подчеркнуть методологическую комплексность дизайна исследования, аккуратность соискателя в работе с данными, математическую фундированность использованного подхода.

Автореферат и публикации Ивана Дмитриевича Мамаева отражают содержание диссертационного исследования.

По прочтении текста диссертации возникли следующие замечания:

1. Не вполне удачным считаю название п. 3.4. «Процедура ручного аннотирования тематических моделей». Оно наводит на мысль о какой-то дополнительной аннотации данных для тематических моделей или самих

моделей по каким-то критериям. Однако в параграфе речь идет о процедуре назначения меток и/или оценки релевантности уже сформулированных меток тем – стандартной задаче внутри методологии тематического моделирования.

2. Стилистически «шероховатой» считаю формулировку новизны исследования на стр. 9 КД «Разработана процедура создания скрытых сообществ на основе современных семантических анализаторов». Как представляется, речь идет не о процедуре *создания* сообществ (они формируются, создаются стихийно, по крайней мере, пока контролировать этот процесс мы не можем), а о процедуре детектирования или, возможно, моделирования скрытых сообществ.

Рецензируемая работа, будучи новаторской в своей области, мотивирует задать несколько вопросов дискуссионного характера:

1. На с.47 КД указывается, что первым этапом составления исследовательского корпуса постов была сплошная выборка из ВКонтакте, в результате которой было отобрано 7000 пользователей, но после фильтрации осталось 700, а после тематического моделирования кластеризованными оказались только 376 пользователей. Иными словами, половина пользователей оказалась «отсечена» тематическим моделированием при том, что отбирались посты более 200 слов, чтобы обеспечить максимальный захват тематической моделью. В чем видится основная причина малой гранулярности захвата пользователей тематическими моделями? Качество самой использованной модели? Ее гиперпараметры? Предобработка?

2. Второй вопрос связан с первым: из 376 пользователей, попавших в «сети» тематического моделирования, 227 оказались жителями Петербурга, 62 – Москвы, жители других регионов – единичны. Это что-то сообщает нам о специфике речевого стиля жителей мегаполисов или, скорее, говорит о том, что сплошная выборка была все-таки несколько предвзятой? С чего стартовала сплошная выборка?

Заданные вопросы и сформулированные замечания ни в коей мере не умаляют несомненных достоинств работы.

В заключение резюмируем, что диссертационное исследование Ивана Дмитриевича Мамаева «Лингвистическое исследование скрытых сообществ в корпусе социальных медиа с применением мультимодальных тематических моделей» соответствует требованиям п. 9 – 14 «Положения о присуждении ученых степеней», утверждённого Правительством Российской Федерации от 24.09.2013 г. №842 (с изм. от 20.03.2021 г. N 426), предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор заслуживает присуждения ученой степени кандидата филологических наук по специальности 5.9.8 Теоретическая, прикладная и сравнительно-сопоставительная лингвистика.

Колмогорова Анастасия Владимировна
доктор филологических наук, 10.02.19 – Теория языка,
профессор, профессор Департамента филологии Санкт-Петербургской Школы гуманитарных наук и искусств,
заведующий лабораторией языковой конвергенции ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики» в Санкт-Петербурге

дата 02.12.2024 подпись

Контактные данные
Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики» в Санкт-Петербурге
190121, Санкт-Петербург, Союза Печатников ул., д.16
Телефон: +7 (812) 644-59-11 доб. 61222
E-mail: office-spb@hse.ru.
Веб-сайт: https://spb.hse.ru/

Против включения персональных данных, заключенных в отзыве, в документы, связанные с защитой указанной диссертации, и их дальнейшей обработки не возражаю.

Подпись руки Колмогоровой А.В. удостоверяю.
НАЧАЛЬНИК ОТДЕЛА КАДРОВ
Л.Н. Возиян