Выявление проблемных вопросов по социально-направленным тематикам на основе данных открытых источников

О.К. Головнин
Самарский университет
Самара, Россия
golovnin@ssau.ru

И.Н. Дубинина
Самарский государственный технический университет
Самара, Россия
vartaric@yandex.ru

А.В. Иващенко
Самарский государственный медицинский университет
Самара, Россия
anton.ivashenko@gmail.com

А.В. Кривошеев
Самарский государственный технический университет
Самара, Россия
arkas19@gmail.com

П.В. Ситников
ООО «Открытый код»
Самара, Россия
sitnikov@o-code.ru

***Аннотация*—В работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик. Подход предполагает использование текстов и метаданных публикаций в открытых тематических группах и на публичных страницах пользователей в социальных сетях. Выполняется многоэтапная обработка данных по публикациям: сбор данных, очищение от спама, фильтрация по проблемной области, определение тональности, классификация по темам, кластеризация. Программная реализация предложенного подхода выполнена на основе Цифровой платформы интегрального мониторинга. Экспериментальная апробация проведена на данных социальной сети VK. Применение подхода позволяет выявить остросоциальные проблемные вопросы административного региона в оперативном режиме.**

Ключевые слова—большие данные, социальная сеть, классификация текстов, Text Mining.

1. Введение

Социально-экономическое развития регионов России затрудняется без оперативного реагирования на возникающие остросоциальные вопросы [1]. С развитием различных платформ, обеспечивающих общение граждан посредством сети Интернет, в частности, социальных сетей и форумов, появились методы, модели и технологии, позволяющие извлекать информацию о качестве жизни и благополучии населения, а также по социальному самочувствию [2, 3]. В работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик, использующий методы на основе искусственного интеллекта для обработки данных. Подход обеспечивает анализ обсуждаемых вопросов не только по тематикам, но и по роли в социально-экономическом развитии региона, выявляя не только проблемы, но и достижения.

2. Методология интегрального мониторинга

Выявление проблемных вопросов по социально-направленным тематикам выполняется на основе анализа публикаций в открытых тематических группах и на публичных страницах пользователей в социальных сетях.

На *первом этапе* осуществляется сбор исходных данных, связанных с заданным анализируемым административным регионом. Исходные данные для исследования состоят из следующих атрибутов: текст публикации, время и дата поста, местоположение пользователя, пол и возраст пользователя. Информация собирается за указанный период из тематических групп социальной направленности административного региона с вопросами по оказанию помощи, начислению единовременных выплат и т.п.

На *втором этапе* собранные данные очищаются от спама, для чего используется классификатор, относящий пост к классу «спам» или «не спам»:

1) Грубая оценка поста по заданным правилам (например, низкая доля русских букв в тексте, завышенная доля спецсимволов, короткие/длинные тексты, наличие стоп-слов и их сочетаний и т.п.);

2) Выполняется обработка анализируемых текстов через ряд функций. В качестве исходных данных для таких функций выступают следующие значения: общее количество символов в тексте, количество символов кирилицы, латиницы и цифр, число хештегов, смайлов, спецсимволов. Фукнции ранжируют данные числовые значения и выдают вероятность того, что текст является спамом. Все оценки суммируются; на выходе у каждого текста формируется итоговая оценка;

3) Тексты с высокой оценкой на спам получают метку «спам»;

4) Производится очистка оставшихся текстов от спецсимволов, смайлов, хештегов, HTML-кода, ссылок и иных нетекстовых включений;

5) Нейронная сеть «sentence-transformers/stsb-xlm-r-multilingual» переводит очищенные тексты в вектора («эмбеддинги»), размерностью 768 чисел;

6) Полученные вектора классифицируются заранее обученной глубинной нейронной сетью на классы: «спам» и «не спам».

На *третьем этапе* после завершения очистки набора данных от спама происходит поиск таких сообщений, которые содержат вхождения ключевых слов. Список ключевых слов (маркеров) составляется заранее на этапе подготовки к исследованию. Список включает в себя основные словоформы и ключевые фразы, относящиеся к теме исследования. Составление списка ключевых слов (маркеров) требует участия специалиста в заданной предметной области. Таким образом, публикация из социальной сети относится к социально-направленной теме, если выполняются следующие условия: публикация не определена классификатором спама как «спам»; публикация содержит в себе один или более

ключевых слов (маркеров) из заданной предметной области.

На *четвертом этапе* проводится анализ тональности каждого сообщения из полученной отфильтрованной выборки с помощью модели на основе BERT. Классификация текста производится в соответствии с пятью классами: 1 – негативный; 2 – позитивный; 3 – нейтральный; 4 – речь; 5 – неизвестно.

На *пятом этапе* осуществляется определение общих тем и направлений публикаций. На основе текстов публикаций и обращений определяются кластеры, объединенные общей тематикой, и этим кластерам присваиваются общие названия, обозначающие суть проблемы или обращений. Кластеризация тем осуществляется следующим образом:

1) Между публикациями пользователей на основе эмбеддингов считается косинусная мера близости между векторами A и B:

$$cosine\_measure = 1 - \frac{A \cdot B}{\|A\|\|B\|} = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}};$$

2) Если косинусная мера менее 0.25, то публикации пользователей определяются как схожие по смыслу или значению и объединяются в один кластер;

3) После объединения в кластеры производится итеративный перебор пар кластеров, причем, если у меньшего кластера более половины публикаций принадлежит большему кластеру, то меньший кластер объединяется с большим кластером;

4) В качестве действующих кластеров выбираются такие, чей размер составляет не менее $l = \max(3, L^{0.25})$, где L – количество постов после проверок на спам;

5) Внутри каждого кластера подсчитывается количество биграмм (словосочетаний) с учетом синтаксиса в предложениях; N=3 наиболее частых биграмм становится названием кластера.

6) Для каждого кластера производится оценка уровня тональности $T=\max(count(T_n))$ n=1..5, где n соответствует номеру категории тональности текста.

Такой подход позволяет определить кластеры, их названия и тональности, с которыми далее осуществляют работу лица, принимающие решения. Таким образом, негативные кластеры считаются проблемными вопросами региона, а позитивные кластеры – наиболее значимыми достижениями региона, нейтральные – не содержат остросоциальные проблемы.

3. Программная реализация и результаты

Программная реализация предложенного подхода выполнена на языке Python в среде Цифровой платформы интегрального мониторинга [4, 5]. Исследование эффективности проведено на основе данных социальной сети VK за период в 3 мес. Сбор данных осуществлялся с помощью парсеров платформы через API. Так, за 3 мес. в анализируемом регионе выявлено около 700 тыс. постов, из них 480 тыс. постов классифицировано как спам. Из оставшихся 220 тыс. постов только 50 тыс. постов являются социально значимыми, из этих 50 тыс. постов относятся к позитивным 4 тыс. постов, а к негативным – 18 тыс. постов. В результате анализа позитивных постов выделено 9 кластеров, из них 7 отмечено аналитиком как целевые – содержащие актуальную информацию. В отрицательных постах выявлено 45 кластеров, из них 21 отмечены аналитиком как целевые. В таблице 1 представлена матрица несоответствий. Точность подхода составляет 0,42, полнота – 0,88, значение F-меры в исследовании 0,57.

Таблица I. Матрица несоответствий

Категория социально-значимого обращения		**Экспертная оценка**	
		Положительная оценка	***Отрицательная оценка***
Оценка метода	***Положительная оценка***	21 тыс.	29 тыс.
	Отрицательная оценка	3 тыс.	167 тыс.

Таким образом, не смотря на наличие неактуальных кластеров в результатах, всё же существенно снижается нагрузка на аналитика-оператора на анализ ключевых проблем региона, поскольку просмотр нескольких десятков тематик значительно менее трудоемок, чем поиск и просмотр исходных данных даже отдельно взятых публичных групп. В результате анализа определено, что нецелевые кластеры в негативных постах возникают в результате того, что пользователи социальных сетей обсуждают не реальные события или проблемы в регионе, а высказывают свое мнение о ситуации абстрактно или без конкретных фактов.

4. Заключение

Таким образом, в работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик. Подход программно реализован на основе Цифровой платформы интегрального мониторинга, применение которой позволяет снизить нагрузку на аналитика-оператора, выполняющего анализ ключевых проблем региона, за счет существенного сокращения количества просматриваемой информации.

Литература

[1] Аганбегян, А.Г. Анализ и прогнозирование социально-экономического развития регионов (методические заметки) / А.Г. Аганбегян // Среднерусский вестник общественных наук. – 2019. – Т. 14, № 4. – С. 15-28.

[2] Овчар, Н.А. Технологии исследования социального самочувствия горожан на основе анализа web-контента / Н.А. Овчар, А.С. Воробьев, Д.С. Парыгин, Н.П. Садовникова // Системный анализ в науке и образовании. – 2019. – Т. 1. – С. 83-92.

[3] Щекотин, Е.В. Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций / Е.В. Щекотин // Вестник ТомГУ. – 2021. – Т. 467. – С. 170-181.

[4] Сурнин, О.Л. Применение цифровой платформы интегрального мониторинга как средства бизнес-аналитики социально-экономического развития региона / О.Л. Сурнин, П.В. Ситников, А.В. Иващенко, О.К. Головнин [и др.] // Информ. технологии в управлении: сб. материалов. – СПб.: СПбГЭТУ ЛЭТИ, 2022. – С. 158-161.

[5] Ситников, П.В. Анализ социально-экономического развития региона на базе цифровой платформы интегрального мониторинга / П.В. Ситников, Е.А. Додонова, И.Н. Дубинина [и др.] // Информ. технологии и нанотехнол.: тр. конф. – Самара: Смр. ун-т, 2022. – С. 052032.