УДК 378.1

Определение образовательных интересов школьников на основе анализа пользовательских данных «вконтакте»[1]

У.С. Захарова, А.В. Фещенко

Национальный исследовательский Томский государственный университет, Томск, Россия
e-mail: fav@ido.tsu.ru

Аннотация: *социальные медиа являются важным элементом в коммуникационной политике современного университета, позволяют без посредников доставлять информацию до целевой аудитории, обеспечивают широкий территориальных охват при небольших финансовых затратах. Но существующие методы нацеливания рекламы в социальных сетях не позволяют университетам определять индивидуальные образовательные потребности и интересы потенциальных абитуриентов и предлагать им персональные рекомендации по выбору образовательных программ. По этой причине университеты во время рекрутинговых кампаний создают в социальных медиа универсальные сообщества с рекламой сразу всех образовательных программ. При таком подходе сложно разделить целевую аудиторию по интересам и сфокусировать её внимание на программах обучения, соответствующих этим интересам. Современные методы анализа пользовательских данных в социальных сетях позволяют университетам проводить рекрутинговую кампанию более эффективно. В нашей работе представляется опыт Томского государственного университета по применению методов анализа данных для выявления в социальных сетях абитуриентов с интересами к тому или иному профилю подготовки. В работе использованы методы контент анализа, статистики, анкетирования, data mining.*

Ключевые слова: *анализ данных, социальные сети, образовательные интересы, абитуриенты.*

При использовании университетом социальных сетей для рекрутинга возникает задача, связанная с отбором абитуриентов с сильным интересом к определенной предметной области и мотивацией к обучению. Стандартные

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 17-16-70004.

инструменты социальных сетей по сегментированию целевой аудитории используют в основном социальные, демографические и географические данные. Для выявления потребностей и интересов абитуриентов в сфере образования этих данных недостаточно, но они могут быть дополнены информацией о пользователе, содержащейся в его профиле: подписки к тематическим группам и страницам, публикации на стене, сеть связей и т.д. Подходы, позволяющие анализировать пользовательские данные и интерпретировать их для организации эффективного информационного воздействия, уже используются в политике и маркетинге. Основу этих подходов составляют методы лингвистического анализа и психодиагностики [Schwartz H., Kosinski M., Markovikj D., Mangal]. Но пока найденные решения не применяются университетами для выявления образовательных интересов и рекрутинга абитуриентов. Поэтому одной из задач нашего исследования является проверка гипотезы о возможности поиска потенциальных абитуриентов в социальных сетях для конкретных факультетов и направлений подготовки через выявление их интересов к соответствующим предметным областям.

Таким образом целью исследования является поиск методов выявления среди пользователей социальной сети старшеклассников с проявлениями интереса к той ли иной науке, сегментация аудитории по направлениям подготовки, ранжирование по степени проявления интереса, рекрутинг абитуриентов с наиболее выраженными интересами на соответствующие факультеты.

По нашему мнению, интерес старшеклассника к той или иной области знаний связан с вероятностью поступления на определенный факультет университета. В социальной сети интересы пользователя проявляются через тексты, опубликованные на странице его профиля и участие в сообществах, анализ которых, по нашему мнению, позволяет определить интерес к той или иной области знаний и классифицировать всех потенциальных абитуриентов на три группы интересов: гуманитарные, естественные, физико-математические науки, а затем дифференцировать в каждой группе пользователей по степени проявления (силе) интереса.

Для анализа текста мы прибегаем к методологии контент-анализа [Riffe, Lacy & Fico 2005]. Путем автоматизированного подсчета релевантных текстовых единиц (в нашем случае – отдельных тематически соотнесенных слов), мы планировали выявить заинтересованность отдельного пользователя – потенциального абитуриента – в конкретной области знания.

Дополнительно к методологии контент-анализа использовался метод статистики: дисперсионного анализа по Краскелу-Уоллису (Kruskal–Wallis one-way analysis of variance).

Основным инструментом для получения данных из социальной сети является Application programming interface (API). С помощью API возможно получить все публичные данные пользователя, в том числе поля профиля пользователя (имя и фамилия, город, страна, пол, образование, интересы,

любимые книги и т.д.), контент личной страницы пользователя (стены), а также список тематических групп.

Проверка гипотезы о возможности определения интересов пользователя социальной сети через анализ текстов его стены осуществлялась на студентах ТГУ. Приемная кампания в университет начнётся только в июне 2017 года и закончится в сентябре, поэтому проверить методы анализа на реальных абитуриентах пока не представляется возможным. На текущем этапе исследования в социальной сети «Вконтакте» были выбраны профили студентов ТГУ первого курса, и собраны тексты со стены, опубликованные до момента поступления в университет (до 01.08.2016). Из полученной выборки были исключены тексты объемом менее 10 Кб. Всего 232 текста, что составило 17% от первоначальной выборки.

С помощью контент анализа тематических сообществ составлены словари, определяющие принадлежность текста к одной из трех тем: гуманитарные, естественные и физико-математические тексты. Каждый из трех словарей состоит из 400 слов - маркеров. Мы сравнили тексты из профилей студентов с полученными словарями для проверки гипотезы о существовании связи между тематикой текстов на стене пользователя и выбором факультета при поступлении в университет (Таб. 1). Для 85% студентов гуманитарных факультетов доля лингвистических маркеров из гуманитарного словаря была больше, чем из естественнонаучного и физико-математического, 9% - меньше и для 6% доля гуманитарных текстов соразмерна либо физико-математическим, либо естественнонаучным.

Таблица 1. Соответствие тематики текстов со стены ВК студентов направлению подготовки.

Направление подготовки	Соответствие тематики текстов направлению подготовки		
	Соответствует	Не соответствует	Спорное
Естественнонаучное	64%	27%	9%
Физико-математическое	32%	58%	10%
Гуманитарное	85%	9%	6%

Таким образом, метод анализа текстов со стены пользователей «Вконтакте» для определения заинтересованности к тому или иному предметному профилю обладает рядом ограничений. Во-первых, объем текста на стене для объективного анализа, должен превышать 10 кб, что существенно ограничивает число анализируемых объектов со 100% до 17%. То есть этот метод

не применим к большинству профилей старшеклассников «Вконтакте». Во-вторых, метод анализа текстов с помощью специализированных словарей достаточно точен (85%) только для определения пользователей с гуманитарными интересами, для выявления пользователей с естественно-научными и физико-математическими интересами точность метода недостаточно высока.

Метод анализа контента стены при выявлении образовательных интересов абитуриентов планируется дополнить анализом тематических сообществ, в которых они состоят. Вступление в сообщество и подписка на страницу в социальных сетях может характеризовать интересы абитуриента. Если выбрать из спектра выявленных интересов абитуриента темы, имеющие отношение к образованию и познанию, то точность классификации абитуриентов по предметным областям может быть повышена.

В рамках исследования проведен анализ тематического содержания сообществ для 18000 абитуриентов только одного города, Томска. Из профилей абитуриентов выгружены и обобщены сообщества, в которых они участвуют. Из общего количества сообществ выбраны 959, только те, которые встречаются в профилях не менее 10 пользователей. Определение тематики сообщества проводилось вручную. В результате составлен классификатор сообществ и определена доля каждой рубрики в общем количестве сообществ.

Проверка классификатора на 992 студентах ТГУ показала, что 66% из них подписаны на группы и страницы, тематика которых может быть связана с той или иной предметной областью. Сравнение направления подготовки студентов с тематикой сообществ, на которые они подписаны представлена в таблице (Таб. 2).

Таблица 2. Соответствие тематики сообществ «Вконтакте» у студентов направлению подготовки.

Направления подготовки	Количество проанализированных студенческих профилей	Соответствие тематики сообществ направлению подготовки		
		Соответствует на 100%	Соответствует более чем на 30%	Соответствует менее 30%
Гуманитарные науки	324	88%	6%	6%
Физико-	199	17%	1%	82%

математические науки				
Естественные науки	139	4%	0%	96%

Точность выявления гуманитариев с помощью классификатора сообществ составила 94%. Невысокая точность определения интересов к физико-математическому и естественнонаучному контенту можно объяснить ограниченной выборкой сообществ для составления классификатора: из 959 проанализированных сообществ 231 соответствует гуманитарной тематике, 22 физико-математической и только 1 естественно-научный.

На текущем этапе исследования методы анализа текстов в профиле пользователей, а также групп и страниц, на которые они подписаны, позволяют идентифицировать с высокой точностью только гуманитариев. Применение этих методов к профилям потенциальных абитуриентов 2017 года, позволит определить относительную частоту упоминания лингвистических маркеров в текстах на стене и абсолютные значения по количеству тематических подписок на контент, связанный с интересом к гуманитарным наукам. Мы предполагаем, что эти данные позволят ранжировать всех пользователей по силе выраженности интереса и сузить целевую аудиторию во время работы по привлечению абитуриентов в социальных сетях. Оценить эффект данного подхода мы сможем после окончания приемной кампании в августе 2017 года. Ожидаемые результаты: расширение географии абитуриентов, увеличение конкурса на гуманитарные направления подготовки, увеличение доли первокурсников, узнавших об университете через социальные сети, уменьшение количества отчислений из университета в первый год обучения, повышение успеваемости в первый год обучения.

СПИСОК ЛИТЕРАТУРЫ

1.*Schwartz H.A.* et al. Personality, gender, and age in the language of social media: The open-vocabulary approach // PloS one. – 2013. – Т. 8. – №. 9. – С. e73791.

2.*Kosinski M.* et al. Manifestations of user personality in website choice and behavior on online social networks // Machine learning. – 2014. – Т. 95. – №. 3. – С. 357-380.

3.*Markovikj D.* et al. Mining facebook data for predictive personality modeling // Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA. – 2013.

4.*Rutter R.,* Roper S., Lettice F. Social media interaction, the university brand and recruitment performance // Journal of Business Research. – 2016. – Т. 69. – №. 8. – С. 3096-3104.

5.*Mangal N.*, Niyogi R., MilaniA. Analysis of Users' Interest Based on Tweets // Computational Science and Its Applications. – Springer, 2016. – V. 9790. – pp. 12-23.

6.*Fagerstrøm A.*, Ghinea G. Co-creation of value in higher education: using social network marketing in the recruitment of students // Journal of Higher Education Policy and Management. – 2013. – Т. 35. – №. 1. – С. 45-53.

7.*Riffe D.*, Lacy S.,Fico F. Analyzing media messages: Using quantitative content analysis in research. — Mahwah, NJ : Erlbaum, 2005.