МЕТОД Т-ГРАММНОГО АНАЛИЗА ДЛЯ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ МАССИВОВ, ПОРОЖДЁННЫХ МАЛОГРАМОТНЫМИ НОСИТЕЛЯМИ СИНТЕТИЧЕСКИХ ЯЗЫКОВ

Н.В. Бобров
Московский государственный лингвистический университет
Москва

Анализ неструктурированных текстовых массивов социально-сетевого дискурса принадлежит к числу актуальных проблем последнего десятилетия. Актуальность данной проблемы обусловлена необходимостью мониторинга социальных процессов, прежде всего, с целью недопущения формирования преступных сообществ, которые в условиях современного уровня развития социальных сетей могут с катастрофической скоростью охватывать значительные массы людей независимо от их географического местонахождения.

Перенос коммуникативного взаимодействия активной части населения в электронно-медийную среду, с одной стороны, затруднил работу правоохранительных органов (традиционные схемы «отработки контактов» стали менее эффективны), с другой — предоставил им новые возможности: если раньше для начала работы требовалась «зацепка», то теперь широкоохватный мониторинг социальных сетей может сам по себе приносить необходимые результаты. В некотором роде это можно считать «презумпцией подозреваемости пользователей социальных сетей», но, поскольку данная «презумпция» не приводит к поражению пользователей соцсетей в каких-либо правах, то вряд ли возможно усматривать в этом нарушение основополагающих принципов права.

Между тем, чисто технически мониторинг социально-сетевого дискурса представляет собой несколько более сложную задачу, чем может показаться на первый взгляд [1], что связано, прежде всего, с неоднородностью контента социальных сетей. Мониторинг социально-сетевого дискурса, таким образом, подразделяется на четыре глобальных направления: анализ изображений, аудиоконтента, видеоконтента и собственно текстовых массивов. Первые три из них требуют серьёзных вычислительных мощностей и специальных алгоритмов выделения коммуникативно-прагматической информации. Но даже текстовый контент, в котором эта информация, можно сказать, «лежит на поверхности», представляет собой непростой объект для анализа. Прежде всего, это касается молодёжной социально-сетевой коммуникации на флективных языках (таких, как, например, русский), а также на некоторых младописьменных языках, реализуемой их носителями, не знающими (или не считающими нужным знать) своей нормативной письменности. Две острейшие лингвотехнологические проблемы — синтетический строй языков и малограмотность их носителей — здесь сплетаются в одну, не только острую, а ещё и тяжёлую.

В настоящее время задача мониторинга социально-сетевого дискурса обычно рассматривается с позиций контент-анализа и так называемого сентимент-анализа. Подходы, реализуемые в рамках данных парадигм, как правило, подразделяются на два класса: основанные на словарях и правилах и статистические [2, 3]. В основе первых лежит эвристическая установка, согласно которой основную семантическую и прагматическую нагрузку текста несёт лексика и в некоторой степени — синтаксис, а следовательно, заложив в анализатор описание того и другого, можно получить желаемый результат — меру соответствия поданного на вход системы текста заданным характеристикам. Вторые основываются на идее машинного обучения и реализуются чаще всего с использованием технологии нейронных сетей. Оба класса подходов имеют свои преимущества и недостатки (первый является в большей степени детерминистическим, но это же делает его более чувствительным к изменению характера входных данных, второй — обладает большей гибкостью, но при этом результат сильно зависит от обучающей выборки и не может быть однозначно интерпретирован). При этом оба класса подходов оказываются чувствительны к уже упомянутым выше отклонениям от орфографической нормы, хотя и по разным причинам: в первом случае — потому, что анализ текста основан на поиске непосредственных лексических соответствий (при этом предполагается, что люди в основной своей массе пишут грамотно, поэтому ошибками «можно пренебречь»), во втором — потому, что, начиная с некоторой пороговой концентрации, ошибки начинают «размывать» полученную статистическую модель.

Предлагаемый подход к анализу социально-сетевого дискурса был впервые сформулирован нами в ходе работ по проекту 14-18-01059 (научный руководитель – д-р филол. наук, проф. Р.К. Потапова), поддержанному Российским научным фондом, в 2015 году. Он находится, в некотором смысле, посередине между словарно-ориентированными и статистическими подходами и позволяет обойти обе обозначенные выше проблемы. Суть его состоит в следующем. Если мы имеем лексему, которая в текстах встречается, например, в сотне различных вариантов (парадигма склонения, помноженная на сопоставимую с ней по объёму парадигму искажения), то с некоторой большой вероятностью мы всё равно имеем компоненты этой лексемы, которые остаются инвариантными. Если мы разобьём лексему (скажем, 7-буквенную) на

перекрывающиеся t-буквенные цепочки, то мы получим некоторое множество t-грамм, являющееся её «портретом». Например (для t=4):

яффшоке -> яффш, ффшо, фшок, шоке

Другие варианты этой же лексемы дадут:

яфшоке -> яфшо, фшок, шоке

яффшоооооке -> яффш, ффшо, фшоо, шооо, оооо, ооок, ооке

[я в] шоке -> шоке

Видно, что все полученные цепочки перекрываются. В связи с этим было выдвинуто и проверено предположение о том, что разбиение словоформ на t-граммы можно использовать для агрегации многообразных дублетных форм в некоторые «гнёзда», более широкие, чем традиционные словообразовательные гнёзда, причём это можно делать без привлечения словарных баз данных: проблема морфологического и грамматического анализа здесь «решается» сама, так как (и это надо признать со вздохом) грамматически обусловленная вариативность словоформ социально-сетевого дискурса порой уступает другим видам вариативности (см., опять же, приведённый пример). Это заодно означает и лёгкую переносимость технологии на другие языки. Эксперименты, проведённые в 2015-2016 годах, показали, что оптимальное число t для русскоязычных текстовых массивов составляет 3–4. Оптимальность в данном случае определяется как соотношение релевантных и посторонних включений в результирующих гнёздах.

Исследования показали, что t-граммы сами по себе подчиняются закону Ципфа, причём их распределение в основном (хотя и не полностью) коррелирует с распределением соответствующих им лексем. Таким образом, можно проводить предварительный частотный анализ сначала t-грамм, а уже потом лексем, попадающих в отфильтрованные t-граммным частотным анализом автоматически сформированные гнёзда. При этом ещё на стадии t-граммного анализа можно проводить предварительный коллокативный анализ с целью выявления вероятных лексических коллокаций, которые могут служить маркерами для обнаружения текстов, представляющих потенциальный оперативный интерес.

Как показали предварительные эксперименты, метод t-граммного анализа можно применять и для определения меры сходства текстов и их кластеризации, что может быть полезно при проведении предобработки больших текстовых массивов, содержащих разнородный материал.

Необходимо отметить, что несомненным преимуществом предлагаемого подхода является его малая ресурсоёмкость, что делает его пригодным для поточной обработки больших текстовых массивов «на лету». Кроме того, что, как уже было сказано, алгоритм t-граммного анализа не требует использования словарных баз данных и сложных вспомогательных алгоритмов лемматизации и морфологического анализа, он фактически позволяет рассматривать текст как последовательность 32- или 64-битных целых чисел (в зависимости от числа t и используемой кодировки), что обеспечивает максимальное быстродействие и производительность t-граммных текстовых процессоров при минимальных затратах.

В настоящее время разработано два исследовательских приложения, использующих описанный подход, ведутся работы по созданию программного продукта, пригодного для непосредственного использования в интересах Российской Федерации.

Работа выполнена при поддержке Российского научного фонда, грант № 14-18-01059.

ЛИТЕРАТУРА

1. Potapova R., Potapov V. Polybasic Attribution of Social Network Discourse / Ronzhin A., Potapova R., Németh G. (eds) Speech and Computer. SPECOM 2016. Lecture Notes in Computer Science, vol 9811. Springer, Cham, 2016.
2. Bing Liu. Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing / под ред. N. Indurkhya и F. J. Damerau. 2010.
3. Анна Пазельская, Алексей Соловьев. Метод определения эмоций в текстах на русском языке // The international conference on computational linguistics and intellectual technologies “Dialogue 2011”: конференция. Москва, 2011. С. 510 - 522.