

Опыт выявления спама в процессе бренд-мониторинга социальных сетей

© А. В. Суханов

sukhanov@rco.ru

© М. В. Калинина
ООО «ЭР СИ О»,
Москва

kalinina_m@rco.ru

© А. В. Антонов

alexa@rco.ru

Аннотация

В настоящей статье рассматривается опыт фильтрации спама и другого информационного шума в процессе мониторинга упоминаний брендов в социальных сетях с целью выявления отношения пользователей к брендам коммерческих компаний.

1 Введение

При создании автоматизированных информационных систем, предназначенных для тематического мониторинга сообщений социальных сетей (ВКонтакте, Твиттер, Живой журнал, Фейсбук, Одноклассники и т.д.), необходимо решать проблему фильтрации нерелевантных для данной задачи сообщений (спама, информационного шума). Для краткости будем называть такие нерелевантные сообщения спамом.

Разумеется, количество типов сообщений, которые будут отнесены к категории нерелевантных, зависит от цели мониторинга, т.к. вопрос о том, что считать спамом/шумом не имеет однозначного ответа.

Спам в общепринятом значении – рассылка коммерческой и иной рекламы или иных видов сообщений лицам, не выразившим желания их получать. Для электронной почты спам – несанкционированные нежелательные рассылки.

В рассматриваемой нами задаче отслеживания личного отношения пользователей к брендам коммерческих компаний будем считать спамом все то, что не содержит:

- мнения пользователя о бренде,
- опыта использования продуктов или услуг данного бренда,
- вопросов, связанных с деятельностью бренда.

При этом к категории спама, помимо обычных спам-сообщений (рекламные объявления, мошенничество, и т.п.), будем также относить перепечатки новостей и случайные (неинформативные) упоминания бренда.

2 Постановка задачи

Поток входных сообщений из социальных сетей отфильтрован по содержанию в сообщении наименования брендов и их синонимов.

Среди этих сообщений велика доля являющихся спамом в определенном выше смысле.

Задача состоит в том, чтобы отделить нерелевантные сообщения от значимых, и таким образом сузить поток сообщений, которые представляют интерес для потребителя (рис.1).

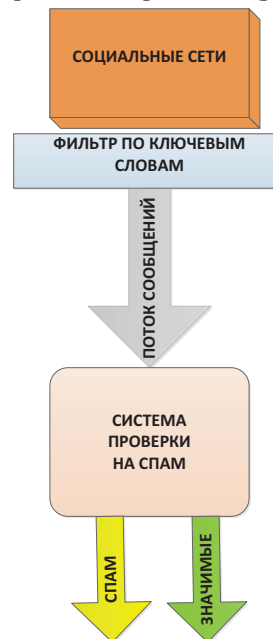


Рис 1. Разделение потока сообщений на значимые и спам

3 Классификация спама

При анализе сообщений, не содержащих оценки деятельности коммерческих компаний, нами были выделены следующие разновидности спама:

- Бренд как часть личного информационного сообщения:

- Звони мне на МТС;
- Встретимся у М-Видео;
- Требования: ответственность, коммуникабельность, грамотная речь, активность. Обязанности: обновлять программу СПС «Гарант».

- Перепечатки новостей – новостные сообщения пресс-релизы компаний с упоминанием искомой коммерческой компании:

- Ростелеком установил для видеотрансляции экзамена в южноуральских школах 1 139 программно-аппаратных комплексов;
- Верховный Суд РФ отменил нормативы на количество детей в игровых комнатах детских садов в зависимости от их площади //

Консультант Плюс.

- Мем – перепечатка популярных шуток, распространяемых пользователями социальных сетей:

- У лукоморья дуб зелёный. Есть Интернет на дубе том. Там виснет в «аське» кот учёный, Отбросив песни на потом. Там на неведанных дорожках Отлично ловит Мегафон.
- Я счастлива. Мегафон сказал, что будущее зависит от меня. Кока-кола обещала, что всё будет хорошо. Газпром утверждает, что мечты сбываются! Tefal думает обо мне! Maybelline говорит, что все в восторге от меня! А L'oreal подтвердил – Я ЭТОГО ДОСТОЙНА!

- Объявление – сообщение является коммерческим объявлением: предложение товаров и услуг, поиск сотрудников и т.д.:

- SIM карты. МТС;
- Мегафон. Цена 50р. В лс;
- В салон связи Билайн требуются сотрудники.

- Омонимия бренда – в сообщении содержится слово омонимичное названию бренда, упоминания бренда как такового в нем нет:

- Дядя, зачем вы кричите, когда у вас мегафон есть?
- После объявления войны, на следующий день, 23 июня 1941г. на предприятиях, в учреждениях, учебных заведениях города Пугачева, в МТС, колхозах и совхозах района прошли митинги.

- Сбор денег – просьба о сборе денег на различные (обычно благотворительные) цели, бренд упоминается как способ сбора денег:

- Самые быстрые способы помочь: Яндекс Деньги 410011783225101 Мегафон 89247337271 БИЛАЙН 89681681308;
- Для материальных пожертвований вы можете воспользоваться удобным Вам реквизитом: Номер Мегафона +79312344317 Карта Сбербанка номер

67619600 0224223561 Киви-кошелек (QIWI Wallet): +79312344317 Яндекс-деньги 410011379200139

PayPal – пользователь sborlize@mail.ru

Webmoney: WMR-кошелёк R12533462...

- SEO-спам – сообщение не содержит осмысленный текст, используется для повышения посещаемости сайтов, популяризации хеш-тегов и т.п.:

- #mmc #mts #pretty #girl #новосибирск #novosibirsk #сибирь #красивая #девушка #beautiful #fashion #спорт #sport

4 Методы выявления спама

Для выявления спама в социальных сетях мы применили ряд технологий, которые можно разделить на следующие категории:

- Лингвистическая обработка;
- Сравнение с образцами спама;
- Сравнение с образцами новостей;
- Построение классификатора методом машинного обучения;
- Система правил на основе выявленных признаков.

4.1 Лингвистическая обработка

В первую очередь сообщение проходит лингвистическую обработку, в ходе которой ведется расчет характеристик текста, а также выявляются упоминания различных сущностей, и производится проверка срабатывания семантических шаблонов [2], описывающих характерные фрагменты текста, либо ситуации. В результате выделяются лингвистические признаки, которые в дальнейшем используются для проверки на спам-сообщения системой правил.

Для каждого типа спама были составлены шаблоны, описывающие условия присвоения признака сообщению. Шаблоны, построенные по специальным правилам на метаязыке, опираются на регулярные выражения и сочетания ключевых слов. При отождествлении шаблона с фразой или словом за сообщением закрепляется не только метка «спам», но и конкретный его тип.

К примеру, одна из разновидностей упоминания бренда – это упоминание названия сотового оператора в качестве уточнения при указании контактного номера телефона в объявлении: «приним заказов с 8.00 до 22.00 ежедневно по тел. 371-37-11(МТС)». Для выявления данного класса спама используются регулярные выражения для поиска в тексте номеров телефонов, рядом с которыми ищется упоминание бренда сотового оператора.

Новостные сообщения выявлялись по ключевым фразам, например: НАЗВАНИЕ БРЕНДА + сообщил/объявил/объявляет/планирует и т.д. (МТС объявляет о снижении цены на смартфоны).

Объявления также выявлялись по ключевым фразам, например: куплю/продам/продаю +

СУЩЕСТВИТЕЛЬНОЕ В ВИНИТЕЛЬНОМ ПАДЕЖЕ (*Продам детский развивающий коврик недорого, бу. За ценой в лс или по тел. мтс 5788714*).

Омонимия бренда определялась по контексту, в частности, сочетанием в одном предложении глаголов кричать/орать/говорить/вопить и т.д. с предложно-падежной группой “в мегафон” (*Потеряв телефон, стахановец-многостаночник орет в мегафон.*).

Выявление признака несвязного текста сделано на основе синтаксического анализа связности слов с учетом соотношения количества известных и неизвестных слов.

4.2 Сравнение с образцами спама

Пополняемая коллекция образцов спама позволяет построить обучающуюся систему фильтрации спама.

Существует несколько эффективных методов сравнения документов для выявления нечетких дубликатов [3]. В работе [1] описан метод обнаружения дублирующихся текстовых документов. За основу механизма выявления сообщений, сходных с образцами спама, мы взяли алгоритм шинглов. Был использован режим симметричной меры с порогом сходства (коэффициентом близости по Дайсу) равным 0,8 и длиной шингла равной 3 словам. В этом режиме, как показала практика, обеспечиваются наилучшие показатели полноты и точности отбора дублей и почти дубли сообщений.

Сообщения, сходные с одним из документов коллекции образцов спама, помечаются как спам.

4.3 Сравнение с новостями

Для решения задачи выявления новостного спама была применена система сбора новостей через настраиваемые каналы RSS.

При этом, учитывая то, что мониторинг социальных сетей ведется по теме брендов, которые определяются набором ключевых слов, отбираются не все новости, а подходящие по ключевым словам.

Дополнительно, чтобы не отбросить по ошибке сообщения в новостной спам, отбираются новости, которые содержат не менее 5 слов.

Проверка сообщений социальной сети на принадлежность к категории новостной спам производится также методом шинглов. Но используется несимметричная мера сходства – сообщение социальной сети должно быть сходным с одной из новостей, а обратное необязательно. Если проверяемое сообщение социальной сети (B) содержит большую долю словаря шинглов одного из новостных сообщений (A), то такое сообщение (B) относится к категории «перепечатка новостей».

Суть примененной несимметричной меры сходства можно выразить формулой:

$$M+ = \text{COMMON}/\text{SIZE_A}, \text{ где}$$

M+ – несимметричная мера сходства, при которой сообщение (B) должно содержать образец (A),

COMMON – количество совпадающих шинглов в словарях (A) и (B),

SIZE_A – размер словаря (A).

Порог сходства несимметричной меры был выбран 0,6 и длина шингла равной 2 словам.

4.4 Построение классификатора методом машинного обучения

Дополнительную информацию о том, что сообщение является или может быть спамом дает классификация сообщения методом Байеса.

Была использована система тематической классификации RCO, основанная на Байесовском классификаторе с распределением Пуассона [6].

Для класса спам-сообщений был построен лингвистический профиль на основе обучающей выборки, включающий в себя слова и словосочетания, появление которых однозначно свидетельствует о принадлежности сообщения к спаму. Отнесение классификатором сообщения к спаму является признаком, используемым системой правил для отнесения сообщения к категории «спам».

4.5 Система правил на основе выявленных признаков

Итоговым шагом в выявлении спама является применение системы правил. По ним формируется окончательное решение на основе признаков, выявленных на предшествующих стадиях обработки:

- Результатов лингвистического разбора;
- Данных о сходстве сообщения с ранее выявленным спамом или новостью (пресс-релизом);
- Результатов классификации сообщения;
- Сведений об источнике сообщения, его авторе, обращениях в тексте к определенным авторам, например, официальным.

Правила позволяют определить принадлежность сообщения к спаму, провести его классификацию, а также скорректировать решение – исключить некоторые сообщения из категории спама (например, если автор сообщения является VIP, т.е. знаменитостью или популярным блоггером).

Ниже представлены примеры правил, использованных для выявления спама.

Примеры правил

Сообщения относятся к спаму в случаях:

- Если автор сообщения спамер.
- Если сообщение распознано как дубликат спама.
- Если сообщению присвоен тематический профиль «Спам».

- Если автор сообщения или журнал относятся к официальным, количество упоминаний брендов в 50 раз меньше количества слов без учета знаков препинания, и сообщение не содержит обращения к автору-спамеру.

- Если сообщение не содержит упоминаний брендов.

- Если у сообщения в ходе лингвистической обработки выявлены упоминания бренда как контактного телефона, и их количество больше, чем других упоминаний брендов.

- Если сообщение не содержит обращений к авторам-спамерам и не содержит глаголов. При этом сообщение не содержит вопросительных предложений и оценочной лексики.

- Если сообщение не содержит обращений к авторам-спамерам и ему присвоен признак «количество предложений» со значением ≤ 3 .

- Если сообщение не содержит обращений к авторам-спамерам и ему присвоен признак «похоже на спам».

- Если сообщению присвоен лингвистический признак «объявление».

- Если сообщению присвоен лингвистический признак «попрошайничество».

- Если у сообщения выявлен признак бессвязного текста, в сообщении содержится ссылка и отсутствуют глаголы. Такое сообщение классифицируется как SEO-спам.

Сообщения исключаются из категории «спам» в случае:

- Если автор обработанного сообщения VIP.

5 Результаты

Применение комбинации описанных методов выявления спама в задаче бренд-мониторинга социальных сетей обеспечивает полноту выявления спама не ниже 80% и высокую точность (не ниже 95%). Более половины сообщений, отобранных для мониторинга по ключевым словам, помечаются системой как спам.

По опыту обработки сообщений социальных сетей по теме телекоммуникационных компаний за период 3 мая – 5 июня 2015 доля спама составляет от 53,2% до 65,9%, в среднем 59,6%.

В таблице №1 приведено количество и доля помеченных системой спам-сообщений по источникам за указанный период.

Таблица 1. Количество выявленного спама по источникам

Источник	Всего	Помечено системой	
ВКонтакте	335494	196181	58,5%
Твиттер	248726	147330	59,2%

Одноклассники	8430	7017	83,2%
Фейсбук	7105	3905	55,0%
Живой журнал	10104	7781	77,0%
Youtube	3932	3657	93,0%
GooglePlus	2269	1736	76,5%
Google.Play	2596	1723	66,4%
Instagram	10759	7467	69,4%
Ответы@mail.ru	2532	695	27,4%
Я.ру	1	0	-
4pda.ru	8058	3780	46,9%
Хабрахабр	137	82	59,9%
Прочие	16184	9732	60,1%
Итого	656332	391090	59,6%

В таблице №2 приведены количества и доля помеченных системой спам-сообщений по датам.

Таблица 2. Количество выявленного спама по датам

Дата	Всего	Помечено системой	
05.06.2015	83721	52831	63,1%
04.06.2015	43784	24434	55,8%
03.06.2015	34647	20939	60,4%
02.06.2015	7597	4473	58,9%
01.06.2015	7560	4895	64,7%
31.05.2015	4965	3127	63,0%
30.05.2015	9419	5627	59,7%
29.05.2015	42675	23847	55,9%
28.05.2015	14187	8887	62,6%
27.05.2015	13043	7959	61,0%
26.05.2015	5289	2812	53,2%
22.05.2015	11196	6924	61,8%
21.05.2015	17647	11635	65,9%
20.05.2015	23900	14587	61,0%
19.05.2015	20055	11251	56,1%
18.05.2015	21908	13071	59,7%
17.05.2015	19640	10906	55,5%

16.05.2015	18611	11112	59,7%
15.05.2015	53061	30214	56,9%
14.05.2015	60480	35552	58,8%
13.05.2015	32129	19165	59,7%
12.05.2015	11488	7101	61,8%
08.05.2015	5620	3023	53,8%
07.05.2015	23079	13118	56,8%
06.05.2015	14882	8563	57,5%
05.05.2015	20585	13134	63,8%
04.05.2015	17584	11070	63,0%
03.05.2015	16861	10338	61,3%
Итого	656332	391090	59,6%

Для повышения полноты выявления спама необходимо исключить сообщения, порожденные генераторами текстов, чему может способствовать применение подходов, описанных в работах [4, 5].

6 Заключение

Большую долю информации для принятия решения об отнесении сообщения социальной сети к спаму дает лингвистическая обработка. Ее дополняют методы сравнения с образцами спама, новостями и классификация, а также учет атрибутов сообщения, его автора и источника. Обобщение же всех имеющихся характеристик и признаков сообщения системой правил позволяет не только повысить достоверность отнесения сообщения к спаму в целом, но и к конкретному его типу.

Литература

- [1] Андреев А.М., Березкин Д.В., Козлов И.А., Симаков К.В. Метод обнаружения дубликатов в потоке текстовых документов. // Труды 16й Всероссийской научной конференции "Электронные библиотеки: перспективные

- методы и технологии, электронные коллекции" – RCDL'2014, Дубна: 2014.
- [2] Ермаков А.Е., Плешко В.В. Семантическая интерпретация в системах компьютерного анализа текста. // Информационные технологии. – Москва, 2009, № 6, с. 2-7.
- [3] Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль, Россия, 2007. – Том 1, С. 166-174.
- [4] Павлов А.С., Добров Б.В. Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Труды 11й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL'2009, Петрозаводск: 2009.
- [5] Павлов А.С. Методы обнаружения массово порождаемых неестественных текстов на основе анализа разнообразия тематической структуры текстов // Труды 13й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL'2011, Воронеж: 2011.
- [6] Плешко В.В., Поляков П.Ю., Ермаков А.Е. RCO на РОМИП 2009 // Труды РОМИП 2009. (Петрозаводск, 2009г.). – Санкт-Петербург: НУ ЦСИ, 2009 – с. 122-134.

Experience of Spam Detection within the Social Network Brand Monitoring Task

A. V. Sukhanov, M.V. Kalinina, A.V. Antonov

The article focuses on the experience of spam and other information noise filtering within the task of brand monitoring of social networks in order to identify users' opinion about commercial brands.