# Общественное мнение онлайн: сравнение структуры и тематики постов «обычных» и «популярных» блогеров Живого Журнала

Светлана Алексеева, Олеся Кольцова, Сергей Кольцов

НИУ ВШЭ, Санкт-Петербург, Россия
{salexeeva,ekoltsova,skoltsov}@hse.ru

**Аннотация.** Статья посвящена сравнению тематической структуры и основных статистических параметров постов «обычных» и «популярных» блогеров Живого Журнала. Исследование показало существенное тематическое сходство обеих выборок, была опровергнута гипотеза о большем интересе «топовых» блогеров к социально-политическим темам по сравнению с обычными блогерами. Различие между двумя группами заключается в меньшей активности и большей зашумленности данных среди «обычных» пользователей.

**Ключевые слова:** тематическое моделирование; LDA; Живой Журнал; общественное мнение.

## Введение

В настоящее время в сообществе интернет-профессионалов установилось представление, что блогосфера, наряду с другим пользовательским контентом, является важным источником общественного мнения интернет-активной части населения [1, 2]. В русскоязычном сегменте большая часть общественно значимых дискуссий сосредоточена на платформе Живого Журнала, поэтому именно этот ресурс выбран предметом исследования [3].

Перед исследователем само-сгенерированного общественного мнения в Живом Журнале стоит ряд методологических вопросов, которые необходимо решить перед проведением собственно социологического исследования. В частности, посты каких блогеров — «обычных» или «популярных» — выбирать для анализа? Какое количество текстов выбирать для анализа, если известно, что за неделю на страницах первых 2000 «топовых» аккаунтов Живого Журнала появляется огромное количество новых данных (в среднем 30000 новых постов и 480000 комментариев к ним)?

О. Кольцова и С. Кольцов в работе [2] показали успешность применения алгоритма LDA (латентного размещения Дирихле) для выявления тематической структуры больших совокупностей текстов Живого Журнала, что позволяет решить вторую проблему: автоматическим путем сформировать темы в исследуемой совокупности текстов и отобрать только те тексты, которые привязаны к

интересующим темам исследования (т. е. таким образом существенно сократив количество текстов для ручного анализа). Решению второй проблемы посвящено данное исследование.

## 1 Цели и задачи

Цель данной работы состоит в том, чтобы сравнить тематику и другие характеристики «популярных» и «обычных» блогеров; под первыми здесь понимаются блогеры, занимающие верхние позиции в рейтингах популярности.
Перед началом исследования мы сформировали две гипотезы:

1. топовые блогеры, ориентированные на публичность и лидерство в формировании общественного мнения, больше пишут для широкой аудитории и на темы, представляющие общественный интерес, тогда как «обычные» блогеры больше пишут о приватных и рекреативных вопросах для своих личных знакомых;
2. Обычные блогеры, не будучи профессионалами, в отличие от популярных блогеров, характеризуются меньшей активностью и смещением этой активности на выходные дни, тогда как популярные блогеры пишут, в основном, по будням.

## 2 Реализация

Данные собраны при помощи разработанного в Лаборатории интренет-исследований программного обеспечения BlogMiner, который позволяет закачивать и хранить посты и комментарии из Живого Журнала вместе с метаданными о аккаунте, времени и дате написания поста или комментария и ссылки на этой комментарий в Живом Журнале.
Выборка включила в себя все посты за месячный период (с 14 сентября по 14 октября 2013 года), созданные первыми 2000 блогерами по рейтингу «Социальный капитал» Живого Журнала[1] и 20000 случайных блогеров, представленных в данном рейтинге с 2001 по 150000 места; всего – 298967 постов и 2800154 комментариев. Предварительные исследования выявили, что количество постов резко падает после 150000 места, что обуславливает выбор данного ранга в качестве нижнего порога. Также ранее было показано, что 20000 нетоповых блогеров создают приблизительно столько же постов, как и первые 2000 блогеров, поэтому количество случайных блогеров было ограничено данным числом.
Автоматическое выделение тем, присутствующих в коллекции постов топовых и нетоповых блогеров, проводилось с помощью алгоритма латентного размещения Дирихле с сэмплированием Гиббса [4] с помощью разработанного в лаборатории программного обеспечения TopicMiner (http://linis.hse.ru/soft-linis). Необходимым параметром для алгоритма является задаваемое вручную количество

[1] http://www.livejournal.com/ratings/users/authority/?country=cyr

тем. После ряда тестов на основе непараметрического метода скачков [5] было определено, что оптимальным для нашей выборки является значения данного параметра равное 120 темам.
В результате тематического моделирования на основе данного алгоритма мы получили две матрицы: матрица, содержащая распределения слов по темам и матрица распределений документов по темам, при этом каждый столбец матрицы означает отдельную тему. Элементы матриц в каждой теме были отсортированы по убыванию. Таким образом были выделены 100 наиболее вероятностных документов по всем темам, которые были переданы двум кодировщикам для присвоения им ярлыков (интерпретации содержания тем)

## 3 Результаты

Проанализировав две группы пользователей Живого Журнала с точки зрения тематической структуры и других социологических показателей мы получили:

- Данные нетоповых блогеров сильно зашумлены: 25% от всех постов (42300) в случайной выборке были написаны одним аккаунтом спамерского происхождения. Данный феномен удалось обнаружить при помощи построения графика распределения количества постов на пользователя, а также распределения количества постов по дням недели (все эти тексты были выложены в Живой Журнал в период с 9 по 14 октября 2013 года).
- Активность нетоповых блогеров гораздо ниже, чем у топовых: большинство нетоповых блогеров, которые вообще имеют посты за исследуемый период, имеют по одному посту, в то время как у топовых блогеров этот показатель равен 40-60 на аккаунт. Кроме того, почти три четверти постов «обычных» пользователей не получили ни одного комментария, у топовых блогеров не получили комментариев менее трети постов. Кроме того, в постах топовых блогеров нередко встречаются дискуссии не менее чем из 10 комментариев, у нетоповых блогеров таких дискуссий крайне мало.
- Обычные блогеры склонны больше писать в будние дни, чем в выходные, причем примерно в той же мере, в которой и популярные блогеры; таким образом, вторая часть гипотезы 2 не подтвердилась.
- Для сравнения тематического состава топовых и нетоповых блогеров нами были просуммированы вероятности отнесения постов топовых и нетоповых блогеров к тем или иным темам. Затем была выделена доля каждой темы в общем весе тем у топовых и нетоповых блогеров по отдельности и было установлено, что в обеих выборках распределение тем практически идентично. Таким образом, мы не можем подтвердить нашу гипотезу о том, что топовые блогеры больше пишут на социально-политические темы, а нетоповых блогеров больше волнуют темы отдыха и личных взаимоотношений. При проведении тематического моделирования из изучаемой выборки не были удалены спамерские аккаунты, и наибольшее различие в тематике обеспечивается именно ими.

- В целом, тематическая структура постов топовых и нетоповых блогеров сходна с результатами предыдущих исследований авторов [5], и отличается в основном событийными темами.

## 4 Заключение

В результате проведенного нами исследования мы можем утверждать, что «обычные» и «популярные» блогеры, в равной степени интересуются как социально-политическими вопросами, так и личной и рекреационной сферами. В отсутствии различия в тематической структуре можно было бы советовать социологам использовать тексты не только популярных блогеров, но и обычных пользователей Живого Журнала для выявлении общественного мнения в блогосфере. Однако, меньшая активность и б*о*льшая зашумленность данных не позволяет этого сделать. Можно также сделать вывод о целесообразности расширения совокупности текстов для изучения онлайнового общественного мнения путем присоединения к общей выборки комментариев к постам популярных блогеров (которые в основном создают нетоповые блогеры).

**Благодарности.** В данной научной работе использованы результаты проекта «Социально-политические процессы в Интернете», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

## Список источников

1. González-Bailón S., Banchs R.E., Kaltenbrunner A. Emotions, Public Opinion, and U.S. Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions // Human Communication Research. - Newark, DE, 2012. Vol. 38. № 2. P. 121–143
2. Koltsova O., Koltcov S. Mapping the public agenda with topic modeling: The case of the Russian livejournal // Policy & Internet. – UK: Wiley-Blackwell, 2013. Vol. 5. № 2. P. 207–227
3. Etling B. et al. Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization. - Rochester, NY: Social Science Research Network, 2010.
4. Griffiths T.L., Steyvers M. (2004) Finding scientific topics // Proceedings of the National Academy of Sciences, 101. P. 5228–5235.
5. Sugar, C.A. and James, G. M. Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach, Journal of the American Statistical Association, 2003; 98(463): 750-763.

# Vox Populi Online: The Comparison of Posts' Structure and Topics Among the "Regular" and "Popular" Bloggers on LiveJournal

Svetlana Alekseeva, Olesya Koltsova, Sergei Koltsov

Higher School of Economics, Saint Petersburg, Russia
{salexeeva,ekoltsova,skoltsov}@hse.ru

**Abstract.** The paper is devoted to comparison of topical structure and basic statistical parameters among the "regular" and "popular" bloggers on LiveJournal. The study has shown a significant topical similarity between both of the user groups. The hypothesis that "popular" bloggers are more interested in social and political topics rather than "regular" ones has been rejected. The discovered difference between the groups is in "regular" users' lesser activity and increased data noise among them.

**Keywords:** topic modeling, LDA, LiveJournal, public opinion.